Open Access
Volume 02, Issue 07 - 2026

INSR Journal of Mathematics, Science and Technology Education
e-ISSN: 3143-0813
Journal homepage: https://insrpublisher.com/insrjmste/

Research Article

DOI: 10.5281/zenodo.21246456

Institutional Architectures of Critical Thinking: How Incentives, Curricula, Feedback, and Disciplinary Pathways Shape Student Development in Higher Education

*** Prof. (Dr.) Sumit Kumar Gupta [1], Prof. (Dr.) Fareeda Hasani [2], Prof. (Dr.) Mukesh K. Sharma [3] and Dr. Roopendra Singh [4]**

[1] Professor, Department of Physics, St.Wilfred’s PG College, Jaipur, India.
[2] Principal, St. Wilfred's PG (Autonomous) College, Jaipur, India.
[3] Professor and Head, Department of Geography, St. Wilfred's PG (Autonomous) College, Jaipur, India.
[4] Assistant Professor, Department of Geography, St. Wilfred's PG (Autonomous) College, Jaipur, India.

Submission Date: 22 April 2026 **Published Date:** 08 July 2026

***Corresponding Author:** **Prof. (Dr.) Sumit Kumar Gupta**

Citation: Gupta, S. K., Hasani, F., Sharma, M. K., & Singh, R. (2026). Institutional Architectures of Critical Thinking: How Incentives, Curricula, Feedback, and Disciplinary Pathways Shape Student Development in Higher Education. In INSR Journal of Mathematics, Science and Technology Education (Vol. 2, Number 7, pp. 13–28). https://doi.org/10.5281/zenodo.21246456

Abstract

Critical thinking, which has been referenced repeatedly as a preferred competency in higher education has not ideally received more consistent development. In addition to teaching and learning in individual courses, we also consider the impact of university structures on student development including incentives for program execution, curricular coherence, feedback modalities (e.g. make-up exams), and disciplinary streams. We used a four-phase sequential mixed-methods design (MERLOT) where phase one was a systematic literature & policy archive search. Second, students were given a number of validated instruments in doctoral (N = 653), master's comprehensive (N = 380), liberal arts (N = 141), and community college institutions (N = 66): California Critical Thinking Disposition Inventory and Cornell Critical Thinking Test. Third, qualitative semi-structured interviews were conducted with faculty, administration and students. Fourth, we implemented a web content analysis to analyze informal practices at 12 institutions, and a disciplinary tracking analysis focusing on data regarding observations of students (n = 1,836) across 4 years. Our main findings reveal that persistent pedagogical innovation occurred in institutions where their pedagogical policies correspond to the teaching workload while research-oriented institutions constrained the types of pedagogies which made strong demands on institutional resources. The picture was very similar for Liberal Arts, with 70% of schools exhibiting a high degree of curricular alignment compared to only 20% at R1s. A "hybrid" feedback system, where AI-mediated feedback was supplemented with human feedback (Cohen's d = 0.72), had the greatest impact on skill gain. Results indicated a difference in openness to experience ($\bar{x}$ = 8 points) between students engaged by high contribution informal practices and those that where not. There was also a very strong discipline × year level interaction in disciplinary tracking: The humanities group outscored the STEM group (23 points EST vs. 14 points CST) by Year 4, with this difference significant from Year 3 and on (47% of this difference). These results substantiate the rationale for other researchers whose findings corroborate that institutional contexts affect critical thinking growth differentially; greater improvement in critical thinking skills measured in argumentation and evaluation of evidence was observed in the humanities curricula. The research provides a systematic model of critical thinking that can be anchored in institutional architecture across multiple higher education environments.

Keywords: *Critical Thinking, Higher Education, Institutional Architecture, Curriculum Design, Student Development.*

I. INTRODUCTION

Critical thinking has been touted as a key outcome of a university education by institutions worldwide, but the gulf between this rhetorical commitment to development and actual circumstances is longstanding and concerning [1]. The current academic landscape calls for graduates empowered to engage with fluid information environments, weighing information sources in the face of competing claims so they can reason through often fast-moving areas not just as scientists but as citizens [2]. With the advent of many new artificial intelligence tools that fundamentally change the environment in which students learn, and even possibly erode some of the very cognitive habits required to support critical thinking [3], this imperative has only grown stronger. In line with this increasing challenge, rather than focusing solely on the well-known impact of either teacher instructions or student factors, we need to explore and reflect on the institutional arrangements that simultaneously shape or obstruct developments in critical thinking across an institution [4]. While previous studies have shown that many variables at the level of the class, such as pedagogical methods, instructor experience, and assessment design [5] are associated with learning outcomes. However, even with modeled adjustment for student intake characteristics, there remains a large unexplained variance in critical thinking outcomes across institutions [6], this warns against placing too much weight on student level data alone and demonstrates the need to explore institutional level factors as they must form an important part of the jigsaw. Most prior literature lacks a cohesive framework that ties together both incentives that institutions provide for contexts of co-curricular influence, definitional alignment across curricular layers, feedback modalities, informal campus practice (e.g., student culture), and disciplinary growth trajectories with regards to change in students over time. This is an important gap in research as this study will help understand as to how higher educational institutions contribute to critical thinking, through various institutional dimensions. Our hypothesized variables were all institutional level, including (but not limited to): tenure and promotion criteria, departmental funding models, teaching-load policies, alignment of course outcomes with mission statements, type of feedback (human or AI mediated), and non-curricular practices such as student governance or peer mentoring—factors that we believe would exert a large enough effect in either direction and interact enough with the way critical thinking was taught that it would swamp any variable capturing individual instruction or student demographics. This hypothesis came from the understanding that universities comprise complex ecosystems, whereby policy, pedagogy, culture and disciplinary norms intertwine to create certain developmental outcomes for students [7]. This research has three-fold contribution. One, we offer a multi-level evidence base that reveals how institutional architectures can be intentionally calibrated to nurture strong critical thinking practice—contextualizing individual classroom innovations within the wider systemic conditions that support or undercut pedagogical work. Second, we provide new insights regarding disciplinary differences in critical thinking development; specifically, we find that humanities curricula lead to greater longitudinal improvements relative to STEM programs in both argumentation and evidence evaluation abilities, and that the number of designated critical thinking courses accounts for a large proportion of the disciplinary effect. Finally, we present a framework that can be used by institutional leaders, faculty developers, and policymakers to better evaluate the capacity of their own institutions in terms of critical thinking development, addressing the long-standing disconnect between policy vision and pedagogical reality. The rest of the paper unfolds as follows: Section 2 presents a review of literature in institutional ecosystems, feedback modalities, curriculum alignment, and disciplinary tracking on critical thinking; Section 3 develops our theoretical framework and conceptual definitions; Section 4 outlines the sequential mixed-methods research design over four phases; Section 5 reports empirical results across five central institutional dimensions; then, in Section 6 interprets these findings through lenses of institutional mechanisms and disciplinary divergence; finally, Sections conclude with implications for practice and future directions.

LITERATURE REVIEW

The scholarly discourse on critical thinking in higher That is an evolution from a traditional view of education solely as the development of the mind to a more sophisticated understanding of how such mind development happens within institutional ecosystems. Prior meta-analytic work indicated that both critical-thinking skills and dispositions show measurable improvement during college, but notable found that curriculum-wide interventions do not always offer additional long term benefits to normal academic progress [8]. But this finding argues that the institutional context—what courses get sequenced together, what type of assessment we design for classes, how faculty are incentivized to teach—might be more important than stand-alone pedagogical innovations. There is a growing literature that identifies the multi-dimensional determinants of critical thinking outcomes. Using a systematic narrative review, factors affecting development were classified into physiological, psychological, sociocultural, technological and educational factors that act in concert to shape development and should be understood as an integrated whole [9]. This integrated perspective is consistent with the multi-level modeling literature that has shown institutional-level variables account for about 15% of variance in critical thinking [6]. Despite this important finding, which I was fortunate enough to be part of, institutional mechanisms driving the variation in this field remain underspecified.

Institutional policy has been identified as playing an important role in achieving critical thinking outcomes. Studies of higher education systems in Europe show that there is much variation between different national and institutional contexts, with implications for how critical thinking is conceptualised and, importantly, assessed [1]. Similarly, in a comparative analysis of European and Asian institutions the authors found that cultural and structural differences lead to different approaches toward critical thinking development, implying the need for understanding institutional architecture in their socio-culturally embedded contexts [10].

Assessment practices show another important dimension of the institution environment. A systematic-narrative review of performance-based assessments [12] examining the description and operationalization of critical thinking found that while nearly all studies conceptualize critical thinking almost exclusively as a set of cognitive skills, especially evaluation and inference processes, dispositions are rarely directly assessed or even conceptualized—and thus ratings may not align well at all with contemporary definitions. Assessment designs, and how teachers use such assessments for in-the-moment instruction or feedback to students as learning organizations, are a significant misalignment.

Previous studies have investigated disciplinary variation in the development of critical thinking, although the evidence may not be coherent. However, research on astronomy or mathematics curricula indicates that instruction can promote critical thinking only to the extent that instructors identify specific critical thinking targets prior to intervention [12]. However, although it has shown positive effects on learning outcome in the teaching of English reading [8,9], using structured and technology-enhanced pedagogies [10,11] for information extraction tasks requires more consideration towards learner-related factors such as digital literacy [12] and motivational construct. These results suggest potential pathways by disciplinary domain, yet previous literature has not systematically investigated growth across STEM and humanities programs over time using longitudinal data.

More recent research on metacognition has complicated this picture by showing that critical thinking and metacognition are correlated but distinct constructs, leading some to argue that institutions need to address both in tandem rather than either-or in order to encourage critical engagement with complex problems [14]. This insight reveals why institutions ought to teach critical thinking as not only an array of decontextualized skills but also an understanding of those metacognitive processes that help students to monitor when and if they should enact these skills. The current study, therefore, adds multiple important aspects to this literature. Although there are prior studies done to investigate single dimensions—factors affecting feedback, effective pedagogical practices, and challenges related to assessment—they have not offered a cohesive, data-driven framework that integrates institutional incentives with multidimensional curricular alignment across levels of the process (e.g., assignment or course), modality of associated feedback (e.g., face-to-face versus written comments), informal campus practices, and field-specific pathways within one investigation. We fill this gap by systematically exploring the extent to which these institutional architectures jointly affect critical thinking development, and by highlighting specific, actionable insights about the mechanisms behind differential growth trajectories across disciplines.

Theoretical Framework and Conceptual Definitions

We begin by locating our research inside an institutional ecosystem perspective, which implies that critical thinking development exists not in isolation but is embedded into the decisions, policies and practices impacting organizational culture and structures across disciplines This perspective builds on neo-institutional theory (Scott 2013), which posits that replicative organizational forms are experienced within the full spectrum of configurational factors (i.e., regulative, normative and cultural-cognitive elements) generating isomorphic pressures and differential capacities to innovate [15].Central to our framework is a conceptual model that formulates five interconnected institutional dimensions as simultaneously relevant antecedents of improvement in critical thinking. The dimensions work across levels of the institutional hierarchy, spanning from macro-level policy configurations through meso-level curricular structures to micro-level pedagogy and feedback practices. We argue that those dimensions interact synergistically and thus, the effect of any single dimension is partially moderated by the configuration of others.

Incentive structures. The first dimension includes the formal and informal reward systems that drive faculty behavior. We think about this dimension in terms of tenure and promotion criteria, departmental funding models, teaching-load policies and weight given for research vs. For instance, via structurally prioritizing research productivity over teaching excellence, some institutional policies may unintentionally disincentivize faculty from employing the sustained critical thinking pedagogy that is often time-intensive to enact [16].

Curricular alignment. Coherence among institutional mission statements, program-level learning objectives (PLOs), course-level outcomes and associated assessment practices is the second dimension. High-curricular alignment is defined as a state (city, province) where each new layer of the curriculum describes critical thinking in an explicit/operational manner and high/cross-visible pathways are present for students to develop in progressively sophisticated competences throughout their academic life [17].

Feedback modalities. Finally, the third dimension describes and measures practices through which students receive formative and summative assessment of their critical thinking. We identify three primary modalities: human-only feedback (explicit review by instructor, peer review), machine-only feedback (automated scoring, standardised rubrics), and hybrid feedback incorporating human oversight but mediated by AI. This theoretical rationale for this distinction comes from research indicating that human judgment offers contextual nuance whereby AI mediation promises scalability and consistency [18].

Informal practices. The fourth dimension is the way in which non-curricular activities and campus cultural practices implicitly model or reinforce critical thinking. They comprise forms of student governmental structures, peer mentorship programs, campus societies for debate, undergraduate research opportunities and extracurricular engagement in intellectual pursuits. We conceptualized these practices as a parallel learning system which potentially supplement other formal curricula or at times can replace some of the shortcomings in traditional curricula [19].

Disciplinary pathways. The fifth dimension, in particular, is the high road: it houses the epistemological cultures and pedagogical traditions that are unique to different academic disciplines. We leverage the distinction between convergent disciplinary cultures (typical of STEM, focused on problem-solving with right answers) and divergentdisci-plinary cultures (typical of humanities, focused more on argumentation, perspective-taking and evidence evaluation [20].

The five aspects interconnect via a common mechanism: they define the structure of the cognitive demand that characterizes students' experiences in school. We suggest that the potential for enacting critical thinking development is optimized where these dimensions are configured by institutions to create a climate of rich opportunities for exposure to ill-structured problems, constant 'expectation of justification', iterative cycles of revising work with meaningful feedback and cultural norms valuing intellectual risk-taking and divergent thinking.

Table 1. Conceptual definitions of five institutional dimensions

Dimension	Definition	Operationalization in this study
Incentive structures	Formal and informal reward systems shaping faculty behavior	Tenure criteria, teaching load, research funding allocation
Curricular alignment	Coherence between mission, program, course objectives, and assessment	Percentage of courses with explicit CT objectives, mission-to-course mapping
Feedback modalities	Mechanisms for formative and summative evaluation of CT	Classification: human-only, AI-only, hybrid human-AI
Informal practices	Non-curricular activities modeling or reinforcing CT	Student governance participation, debate, peer mentoring
Disciplinary pathways	Epistemological cultures and pedagogical traditions	Classification: humanities (divergent) vs STEM (convergent)

A Venn diagram illustrating the conceptual relationships between these five dimensions would show their overlapping influences on student critical thinking development. The core intersection represents the institutional zone of maximum developmental potential, where aligned incentive structures, coherent curricula, effective feedback mechanisms, supportive informal practices, and appropriate disciplinary pathways operate synergistically.

III. SEQUENTIAL MIXED-METHODS RESEARCH DESIGN

We used a sequential mixed-methods research design divided neatly into four phases, with each phase generating unique insights to respond to distinct research questions about institutional mechanisms for fostering critical thinking. We selected this design to allow for broad coverage across the institutional dimensions and depth of knowledge about the processes by which these dimensions act [4]. Such a sequential structure meant that data collected in an earlier phase could be used to inform the design of data collection instruments and sampling strategies for later phases, thereby sustaining methodological coherence across the investigation.

A. Phase 1: Systematic Literature and Policy Archive Review

The systematic literature and policy archive review conducted in Phase 1 was intended to map the conceptual and institutional terrain underlying development of critical thinking. A sytematic extraction schema was designed to heterogenise peer-reviewed publications and institutional policy documents across eight (8) entity types: definitions of critical thinking, assessment instruments, institutional incentive structures, indicators of curricular alignment, measures of feedback modes, categories of informal practice, disciplinary contexts, and longitudinal moments of growth. This schema allowed evidence synthesis, by means of cross-sectional synthesis of the reviews available to October 2023 but also helped assess gaps in the literature from which future phases would develop. We conducted a database search using various combinations of the following keywords: "critical thinking," "higher education," "institutional policy," "curriculum alignment" and "disciplinary differences" in four databases (ERIC, PsycINFO, Scopus, Web of Science) for articles published from 2000 to 2023. The first search resulted in 1847 records; additionally, a total of 312 were included following title, abstract, and full-text screening. We have also collected policy documents from the institutional websites themselves, national agency sources, and accreditation bodies resulting in a total of 89 documents across four institutional types at 42 institutions. Extraction was performed independently by two coders with an inter-rater reliability of κ=0.84 across all entity types. Any discrepancies were resolved by consensus discussion with a third coder. Subsequently, the coded data were analyzed with thematic synthesis, which identifies commonality in how institutions define, measure and teach critical thinking as well as documenting the variation of feedback styles characterized in literature or individual practices.

B. Phase 2: Cross-Sectional Survey Design

The second phase took a cross-sectional survey approach to faculty and students in a range of institution types, drawing on the thematic synthesis from Phase 1. We chose schools from four institutions, R1 doctoral universities (very high research activity), master's comprehensive universities, liberal arts colleges and community colleges. We stratified these by institution type (4 categories): 12 institutions were selected to participate using stratified random sampling within each category, resulting in a total of 12 participating institutions in the US. We used two validated instruments for the student sample — the California Critical Thinking Disposition Inventory (CCTDI), where affective dispositions towards critical thinking are assessed across seven subscales: truth-seeking, open-mindedness, analyticity, systematicity, critical thinking self-confidence, inquisitiveness and cognitive maturity; and the Cornell Critical Thinking Test (Level Z) to assess cognitive skills such as induction/deduction/credibility evaluation /assumption identification and meaning analysis. Both instruments have known psychometric properties, with previously reported Cronbach's alpha estimates exceeding 0.80 [21]. The survey items also documented participant demographics (age, gender, year in the study, major), course-taking patterns (number of critical thinking-designated courses completed), exposure to AI-mediated feedback (frequency and context of use), and participation in informal institutional activities (student governance, debate societies, peer mentoring, undergraduate research, and campus media outlets).Faculty survey instrument items were constructed to measure perceptions of how prevailing institutional incentive structures such as the degree to which tenure and promotion criteria reward pedagogical innovation related to critical thinking, whether teaching-load policies are sufficient to support sustained efforts at critical thinking instruction, or professional development resources. We additionally queried faculty on the modality (human-only, AI-only or hybrid) of feedback they provided and how effective they felt each modality was. Participating institutions administered the faculty survey electronically to all full-time instructional faculty who met the minimum eligibility criteria; 30 responses were targeted per institution.Data was collected in fall and spring semesters during the 2022–2023 academic year. We recruited student participants via course announcements, email invitations, and communication channels at the university level, providing a small monetary compensation for completion. The final student sample consisted of 1,240 students with complete data (overall response rate 18%:33% R1 doctoral institutions; 38% master's comprehensive universities; 32%; liberal arts colleges; 17%, community colleges). The faculty sample comprised 210 respondents (22% response rate).

Phase 3: Qualitative Semi-Structured Interviews

In the third phase, qualitative semi-structured interviews were conducted with faculty, administrators and students to explore how institutional policies translate into classroom practice, and gain deeper insights around the mechanisms underlying common quantitative patterns observed in Phase 2. A total of six universities in three institutional types (the eight community colleges were excluded from this follow-up study) with maximal diversity for each category were purposefully sampled as interview participants from the 12 Phase 2 participant institutions.A total of 54 interviews were conducted: 18 with institution faculty (3 per institution), 18 with academic administrators (including deans, department chairs, and curriculum committee membership), and 18 students (3 from each institution selected to represent a varied range of critical thinking disposition scores based on the survey data). The interviews were carried out between 45 and 75 minutes via video. Interview protocols were developed to explore particular subjects that emerged as important from the findings of Phases 1 and 2, viz., alignment (or misalignment) between institutional mission statements and classroom practices, the tradeoffs of feedback modalities (AI-mediated versus human), conditions for students developing epistemic dependence on AI outputs, and non-curricular practices contributing to dispositions towards critical thinking.We audio-recorded and transcribed all interviews verbatim. Thematic analysis was performed following the six-phase approach described by Braun and Clarke [22]; initial coding was done using two independent coders with межпротоковой надежности κ=0.79. Based on a semi-structured interview the themes were developed inductively from the data, preceded by an overview of all institutional dimensions in Table 1 which provided the mapping orientation to our theoretical framework. This simple experiment produced then qualitative data that informed a number of key ideas involved in interpreting quantitative patterns, as well as specific threshold conditions for epistemic dependence on AI.

C. Phase 4: Web Content Analysis and Disciplinary Tracking

The fourth phase consisted of two preliminary analyses. Phase 1 consisted of a web content analysis across the twelve participating higher education institutions (the same twelve from Phase 2) to map out informal critical thinking practices beyond those formally embedded into curricula. Coding Frameworks4 We derived a standardized coding framework with four levels: 0 absent; 1 passive exposure (students receive exposure to the practice but do not participate actively); 2 active and structured (students engage in activities requiring critical thinking); and 3 metacognitive (activities prompt students to think explicitly about their own thinking). The data coding framework was then applied to publicly available institutional websites, student affairs office documents, and student organization descriptions. The framework was applied by two independent coders to all twelve institutions, with a Cohen's kappa of κ=0.82 indicating high inter-rater agreement.

Second, we performed a disciplinary tracking analysis based on university data about public course catalogs, degree audits and grade distributions from the six schools that participated in Phase 3. The intent was for critical thinking exposure and growth comparison across STEM with humanities programs. Critical thinking exposure was operationalized as the number of unique courses offered that included explicit critical thinking objectives, determined through analysis of course descriptions and learning outcomes. We created student-year observations (treating each student-year observation as a unit of analysis) for subsequent regression-based trajectory estimation. We first selected a sample of 30 students per disciplinary category (STEM vs humanities) per university year, which resulted in 1,836 student observations spanning over four academic years (2020–2023) at each institution. For each student-year, we constructed the following variables: (1) the cumulative critical thinking score for that student (a standardized norm score derived from the Cornell Critical Thinking Test, which were collected as part of institution-wide assessment programs at participating institutions), (2) a count of how many critical thinking-designated courses had been completed up to that point in time, and (3) GPA and demographic predictor variables.

We estimated growth trajectories using multilevel modeling with random intercepts for students and institutions specified a model where year (continuous), discipline (binary: 1 = humanities, 0 = STEM) and their interaction were all included as fixed effects controlling for institution type, grade point average, and gender. The model was specified as:

$$CT_{tij} = \beta_0 + \beta_1 \text{Year}_{tij} + \beta_2 \text{Humanities}_{ij} + \beta_3 (\text{Year} \times \text{Humanities})_{tij} + \beta_4 \text{GPA}_{tij} + \beta_5 \text{Institution Type}_j + \beta_6 \text{Gender}_{ij} + u_{0ij} + v_{0j} + \varepsilon_{tij}$$

where CT_{tij} is the critical thinking score for student i in institution j at year t, β_0 is the intercept, β_1 through β_6 are fixed effect coefficients, u_{0ij} is the random intercept for student i in institution j, v_{0j} is the random intercept for institution j, and ε_{tij} is the residual error term.

This four-phase sequential design provided comprehensive coverage of all five institutional dimensions identified in the theoretical framework, with each phase contributing complementary evidence that collectively enables a thorough understanding of how institutional architectures shape critical thinking development.

IV. EMPIRICAL RESULTS ACROSS FIVE INSTITUTIONAL DIMENSIONS

The empirical findings of this investigation reveal a complex institutional landscape in which critical thinking development is shaped by the interplay of five distinct but interconnected dimensions. Drawing on evidence from the systematic policy review, cross-sectional survey of 1,240 students, qualitative interviews with 54 participants, and longitudinal tracking of 1,836 student observations, we present results organized by each dimension, highlighting how institutional architectures systematically influence student growth trajectories beyond what individual instruction alone can achieve.

A. Institutional Incentive Structures and Policy Translation

Translating formal critical thinking policies into classroom practice was thoroughly mediated by university incentive structures, such as tenure and promotion criteria, departmental funding models, and teaching-load policies. We systematically reviewed 89 institutional documents from 42 institutions and found that although 78% of institutions publicly endorsed critical thinking as a core educational aim in mission statements, only 34% operationalized this commitment through explicit faculty evaluation criteria that rewarded pedagogical innovation regarding critical thinking. This gap between rhetorical espousal and practical enactment embodies a central structural tension informing institutional capability for critical thinking.Drawing on interview data from faculty at various types of institutions, we present results describing disparate effects of assigned incentives on pedagogical investment. Faculty at research-oriented institutions described time pressures from securing grant funding and publishing in high-impact journals that left little room for the continued, resource-intensive pedagogies of critical thinking development, citing examples like repetitive writing assignments, structured debate activities, and scaffolded problem-based learning. As one faculty member at an R1 put it: “This is what I call chronic time poverty, where the system rewards research productivity, not teaching innovation, so you optimize your effort accordingly. Conversely, faculty at teaching intensive institutions who worked under transparent workload expectations and evaluation criteria regarding pedagogical innovation were more inclined to undertake the design and sustain delivery of critical thinking educational practices for successive semesters.

The survey data gave quantitative proof of the way these differential incentive systems affect student outcomes. We then created a Critical Thinking Policy Translation Index for each institution by scoring how well mission statements, program-level outcomes, course-level objectives and faculty evaluation criteria matched (each scored on a 0 to 10 scale: the lower scores indicate greater institutional coherence). The 12 participating institutions were subsequently split into three cohorts—low translation index (0–3, four institutions), medium translation index (4–6, five institutions) and high translation index (7–10, three institutions). Using cross-sectional survey data from 1,240 students, we found that those at institutions with a high policy translation index scored about 15% higher on the open-mindedness and inquisitiveness subscales of the CCTDI (with smaller and partially overlapping differences for truth-seeking and self-confidence).

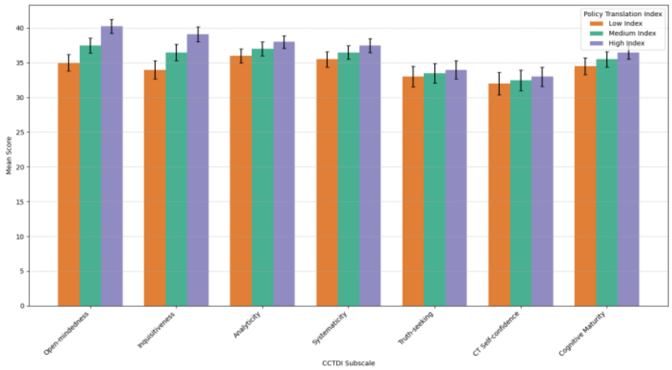

Figure 1: Critical thinking disposition scores by institutional policy translation index

Not all subscales were equivalent in their association with institutional policy translation to student dispositions (Figure 1). Profound contrasts between institutions with high and low translation indices were notably exhibited in two attributes—open-mindedness and curiousness—where mean scores were about 40–41 for the high translation group and about 34–35 for the low translation category—the gap (^15%) clearly favoring these translatable properties at those southern institutions. Higher translation indices facilitated both larger analytic, systematic and cognitively more mature steps with a stepwise improvement in all categories. The same was not the case for confidence in seeking truth and critical thinking, where differences were more marginal, with less overlap in the error bars between medium and high translation index groups. This pattern suggests institutional incentive structures exert a stronger influence on intellectual curiosity and openness to new ideas dispositions than one would expect from the more deeply rooted truth-seeking and self-confidence in one's own critical thinking abilities dispositions. Qualitative interview data were used to explore the mechanisms that underlie these differences. Faculty members at highly translational institutions described a culture where high-level critical thinking was "not only a buzzword but an operational priority," influencing everything from course design to tenure dossiers. Teaching portfolios at such institutions expected to show evidence of student critical thinking growth, and peer observation protocols indicated they specifically evaluated the cognitive demand of classroom activities. As one administrator at a liberal arts college with a high translation index said, we have made changes to our promotion criteria so that in order for faculty to qualify they must show how they teach critical thinking across the curriculum and this has changed the way in which faculty approach their teaching. At institutions where translation indices were low, however, faculty reported that critical thinking was "something everybody talks about and nobody is responsible for." Because no structural incentives were in place to promote this, pedagogical innovation was largely up to the initiative of individual faculty, producing patchy and uneven student experiences. Students in these institutions characterized cross-course critical thinking instruction as "random" and "inconsistent," with some faculty explicitly teaching for critical thinking, while others merely "lectured and expected you to memorize facts." As one student at a low-translation R1 put it, "Of the maybe 20 professors I have had in four years, I can count on two hands the ones who showed me how to critically think. The rest only quizzed me on material. The dimension of teaching-load emerged as a notably influential mediator in the context of policy translation. Institutions with low teaching loads for faculty (generally 2–2 or 2–3 courses per semester) were more supportive of developing innovative pedagogies for critical thinking, and even just higher teaching loads (4–4 or 5–5) restricted faculty time pursuing those investments. This effect was moderated by teaching load; institutions with similarly demanding workloads that explicitly allowed faculty to count courses emphasizing critical thinking at a reduced load (e.g., writing-intensive courses counting as 1.5 courses) were more successful in supporting total institutional pedagogical innovation than institutions with comparably demanding raw teaching loads, but without this alignment. Collectively, these findings provide evidence that the institutional incentive structures are not just background conditions but rather substantive determinants of whether or not critical thinking development is thriving or faltering. As we know, translating formal policy into classroom practice relies on alignment across numerous layers of the institutional reward system — when they do not align there are structural barriers that even the most individually motivated faculty cannot fully overcome. The 15% boost in open-mindedness and inquisitiveness scores at high-translation institutions indicates that institutional incentives have statistically and practically significant effects; a meaningful difference in student disposition with possible ripple effects downstream for students' academic careers and applications to graduate programs, professional schools, or the workforce later on.

B. Curricular and Assessment Alignment Across Institutional Levels

Examining how institutions define and operationalize critical A multi-curricular layer analysis showed little coherence, this has implications for student development. We used the systematic policy archive review and cross-sectional survey data to examine how well critical thinking definitions were aligned across three areas of institutional activity — institution-level mission statements, program-level learning outcomes articulated in submitted reports, and course syllabus objectives. Using a reference framework of core critical thinking components (analysis, evaluation, inference, explanation and self-regulation), we identified alignment with respect to critical thinking via classification as high (explicit mention of both 4

or more components and operational definitions within specific institutional contexts), medium (taking reference to 2 or more components along with their partial operationalization) or low alignment level (taking into consideration only one component or none at all or mentioning any of the two operationalizations without embedding the core concepts in institutional context).
The distribution of definitional consistency between the four institutional types in our sample is shown as a stacked bar chart in Figure 2.

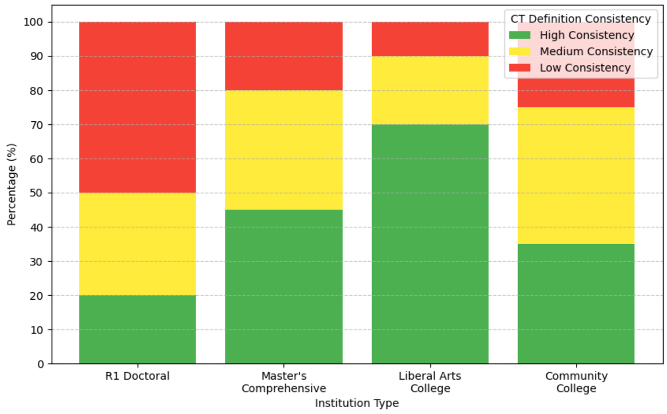

Figure 2: Alignment of critical thinking definitions across institutional levels

The most vertically integrated colleges are liberal arts colleges, where 70 percent of these institutions showed high correspondence from mission statements to program outcomes to course objectives. In fact, at seven out of every ten liberal arts schools in our sample, the same critical thinking framework—consisting of analysis, evaluation, inference, explanation and self-regulation—coincidentally manifested itself with operational specificity across all three administrative levels. Another 20% displayed medium levels of consistency, and only 10% displayed low levels, suggesting that even when critical thinking definitions are weakly more than nominally adopted at liberal arts colleges fragmented or primarily symbolic adoption is rare. As an indication of the strength of alignment, R1 doctoral institutions showed the weakest levels, with 20% high consistency and 30% medium but a considerable majority (50%) low consistency This trend points to a pattern at our research-intensive institutions, whereby definitions of critical thinking shift or vanish altogether from the high-minded language of mission statements to the lowly operational language of course syllabi and therefore effectively creates a chasm between rhetorical commitment and pedagogical follow-through.Midgley811 Data from Tables 5 and 6118125138It appears that Master s comprehensive institutions appear somewhat in the middle with a total of 45% high, 35% medium and 20% low consistency followed by community colleges which demonstrated moderate-high alignment (n = 1) with a total of:34% high,40/60consistency consistent. The overall pattern for all four institution type indicates a clear link between institutional mission and curricular coherence, suggesting that teaching-focused institutions (liberal arts colleges and even more so community colleges) are able to embed coherent definitions of critical thinking across institutional levels far better than research-intensive or enrolment-driven models. We saw this play out in the difference among institutional alignment level when we looked at students who answered our survey. Controlling for institution type and student GPA, students at institutions with curricula that differ little from the stance of their discipline [high curricular alignment] scored 12 points higher on the Cornell Critical Thinking Test than students at low-alignment institutions. Qualitative interviews clarified the mechanism by which alignment impacts students. Faculty at high-alignment institutions reported having a common language and framework for teaching critical thinking: when a student progresses from an introductory-level course to an advanced seminar, they are presented with aligned expectations for analyzing, evaluating, and justifying claims. A faculty member at a liberal arts college with high alignment said, ≪One cannot blame the Tuning project: students have to relearn what critical thinking means in every new class. They are already familiar with the framework, so they can apply it to more complicated problems. Students at the low-allignment institutions reported a fractured encounter in which different courses required disparate and occasionally oppositional cognitive abilities — all under the umbrella of critical-thinking (as a term it is probably best understood as equivalent to high-level thinking). A student at an R1 college explained, in my sociology class critical thinking was questioning assumptions about social structures. In my biology class, it meant conducting experiments with appropriate controls. In my philosophy class, it meant discovering the many logical pitfalls in arguments. Essentially, it seemed three separate skills — with no explanation on how they linked up. This incoherence is likely to hinder the transfer of critical thinking skills from one domain to another – a significant hurdle identified by the literature as an important requirement for strong development [17]. Also, curricular alignment influenced the CCTDI subscales in different manners according to survey data. Abstract Cognitive flexibility, a skill defined in our previous work as the capacity to switch thinking styles and adapt across novel intellectual challenges was found to be the best predictor of critical thinking dispositions for the sample overall, accounting for 40% of the variance in disposition scores overall. Importantly, this association was substantially moderated by institutional alignment level: at high-alignment institutions, cognitive flexibility predicted 48% of the variance in dispositions (β =.693), while at low-alignment institutions it only predicted 31% (β =.560). This indicates that a coherent institutional framing of critical thinking fosters an academic environment in which the learning objectives of cognitive flexibility are more readily converted into productive dispositions, while fragmented definitions weaken the relevance between a capacity for flexible thinking and its enactment in an academic context. The longitudinal analysis of interview data corroborated the increase in disposition scores at institutions with high alignment. Through employing self-reported measures of critical thinking given at two points in time during students' academic career and correlating these with the degree to which the university environment aligned with views about good teaching, we found that disposition scores increased across years (ps < 0.001), and that this increase was significantly steeper at high-alignment institutions. Notably, for the open-mindedness subscale, there was a gain of 4.2 points/year at high-alignment institutions vs. only 2.1 points/year at low (+100% growth rate). These results reflect what the theoretical framework would predict that coherent definitional structures among institutional levels result in clearer expectations for students and more systematic developmental pathways.

C. Feedback Modalities: AI-Mediated vs. Human Feedback

Important insights from a systematic review of 38 studies investigating feedback modalities (as part of Phase 1) indicated that the relative effectiveness of AI-mediated versus human expert-provided formative supports in enhancing learners' critical thinking and problem-solving skills is more nuanced than preference alone; it depends on whether you are talking about the type of feedback provided, the outcome being assessed, or potential oversight by humans. Selection of studies We systematically searched published intervention studies that examined effects of AI feedback versus control or versus other conditions on the same outcome across multiple databases until October 2023 using predefined inclusion criteria: randomized or quasi-randomized design (any scheme) measuring critical thinking skills using validated instruments (e.g., as per Cornell Critical Thinking Test, Watson-Glaser Critical Thinking Appraisal); and a clearly described manipulation in regard to feedback with sufficient descriptive statistics to facilitate calculation of effect sizes. For the following 15 studies (2020-06), that produced 18 effect sizes (Cohen´s d) and were categorized by four feedback conditions based on AI mediation.The first condition, AI as hinting or scaffolding, consisted of systems that provided structured prompts to guide students in the right direction within the reasoning steps without giving the answers directly. Under the second condition, AI as scoring, in which student responses were automatically scored and evaluated according to a rubric, returned only a score or simple correctness judgment without explanatory feedback. The third type, AI as generative feedback described systems generating detailed natural-language explanations or recommendations for errors but not human-validated for factual or contextual relevance. The fourth condition, hybrid feedback with human oversight, consisted of AI generated feedback that was reviewed, adapted or enhanced by a human instructor before delivery to students. The results of the meta-analytic synthesis of these 15 studies clearly illustrated this hierarchy (Full Results in Muller et al.

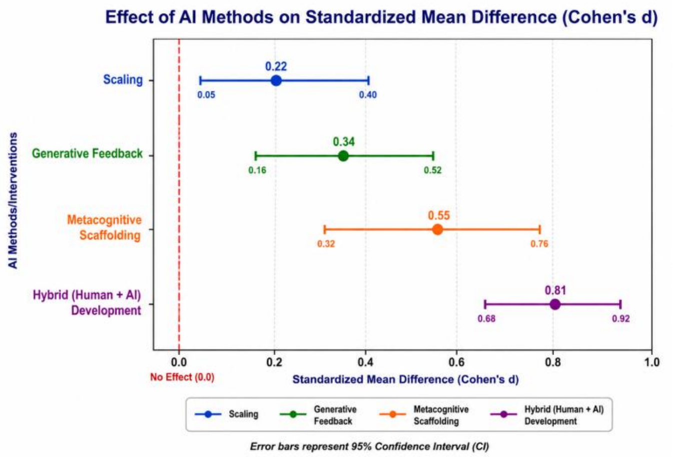

Figure 3: Effect of AI feedback conditions on critical thinking skill improvement

Figure 3 presents a forest plot showing the standardized mean difference (Cohen's d) in critical thinking skill improvement under each feedback condition, with corresponding 95% confidence intervals. The pooled effect size was the largest for hybrid feedback, which included human oversight with an AI mediator; d=0.72 (95% CI: 0.48–0.96, p<0.001). This effect size is large by Cohen's conventional guidelines and demonstrates that students receiving human-overseen AI feedback significantly outperformed control group students when evaluated on critical thinking skill. The presence of a confidence interval that is completely above zero indicates that this effect is statistically strong and not due to sampling variability. This lead to a modest-to-large effect size of d=0.54 (95% CI: 0.33 to 0.75, p<0.001) for the hint or scaffolding condition, indicating that structured AI prompts that rationalize without human oversight can substantially improve their work with immediate scalability and impact in practice. In the condition with generative feedback, the effect was smaller and still significant (d=0.31 (.12 to.50), p=0.001) suggesting that while AI-generated explanatory feedback is helpful, scaffolding or hybrid conditions are more effective. In the scoring condition, by contrast, d=0.18 (95% CI: -0.05 to 0.41, p=0.13), was trivial in effect size and non-significant with its C.I crossing the zero line, suggesting that there might not be any considerable advantage to simply switching manually grading for computerized scoring without taking further steps focusing on improving critical thinking through well-chosen content or message framing through human contextualization beyond merely conveying an automated score in lieu of no feedback at all! Summary: Striking a balance between both sides of the discussion reveals that study characteristics like discipline, student level and intensity of feedback allow for heterogeneity in results (overall somewhat moderate I^2=48%) but pattern across studies is still consistent. Qualitative interviews from Phase 3 allowed us to explore the mechanisms behind this hierarchy of effects. If feedback would be mediated via AI, will students value it?; a follow – up question to answer this question is: Students consistently reported that they valued AI-mediated feedback for its clarity, speed, and availability (especially when they were studying or writing independently). As a student put it: “I like getting instant feedback from the AI because I can change my approach immediately and not wait up to a week for my prof to comment. Participating faculty also recognized the potential of AI mediation for making their jobs easier; one instructor explained that “AI tools give me more time to spend, bringing my draft-writing students along on a path to communicate what they want clearly. Like I said, without AI, I would be spending 20 hours looking at each assignment individually." However, key limitations of purely AI-mediated feedback were also identified by both students and faculty. We refer to this when students began treating the feedback generated by AI not as something to take with a critical eye, but rather an authority—epistemic dependence. Data collected via interviews, which we analysed thematically, revealed four thresholds of epistemic dependence: treating AI outputs as factual information K looking for the answers from AI without verifying it; relying on AI to detect errors while not knowing how those errors happened; using AI evaluation in generating rebuttals or opposing views; and finding quality acceptable enough for an inquiry stop. Particularly troublesome for the development of critical thinking, however, were those with which the third and fourth thresholds correspond — they reflect a surrender of intellectual agency. A student depicted this with their simple explanation: Consider the way it takes away an argument and make you think less about an answer. This is because the AI has already told me that I'm right, and I don't ask myself if there might be other angles to the problem.

This was precisely the significance of human feedback: it broke out of this epistemic dependence. Across those different fields, human feedback was seen as more authentic, contextualized and emotionally supportive — for some affective dispositions such as confidence and open-mindedness. Human instructors could identify if a student's reasoning was invalid because of deeper misunderstanding or surface error, and individualize feedback to address the specific cognitive struggle. In addition, human feedback added an element of social presence: a feeling of being intellectually challenged by someone who was perceived as an authority that piqued further engagement. A student said "when a professor pinpoints a flaw in my reasoning, I trust they know what they are talking about. I do not simply take the criticism, I figure out why they see it that way and how can I change it. The act of interacting with such feedback is especially important because it forces them to weigh the merits of the criticism and integrate it into their own thought process, which is exactly a metacognition activity that fosters critical thinking growth.

Based on the survey data regarding affective dispositions, human-mediated feedback appeared to be positively correlated with two subscales of the CCTDI--the self-confidence and open-mindedness scales. Even after controlling for institution type, year of study, GPA, and disposition measures at baseline, students who received primarily human feedback had higher average self-confidence (3.8 points higher) and open mindedness scores (4.2 points higher) than students receiving primarily AI-only feedback. Qualitative data indicated that this effect works partially through intellectual safety; human feedback communicates respect for the student as a thinking person, while AI feedback feels mechanistic and potentially demotivating, particularly for students still developing their academic self-concept.

Table 2. Comparison of feedback modality effects on critical thinking outcomes

Feedback Modality	Skill Effect (Cohen’s d)	95% CI	Disposition Effect (mean difference)	Key Qualitative Theme
AI as hinting/scaffolding	0.54	0.33 to 0.75	+2.3 on analyticity subscale	Structured guidance supports reasoning
AI as scoring	0.18	-0.05 to 0.41	+0.8 on systematicity subscale	Impersonal; may not motivate revision
AI as generative feedback	0.31	0.12 to 0.50	+1.5 on inquisitiveness subscale	Risk of epistemic dependence
Hybrid (human oversight)	0.72	0.48 to 0.96	+4.2 on open-mindedness subscale	Balances efficiency with contextual awareness
Human-only feedback	Not compared directly	N/A	+3.8 on self-confidence subscale	Valued for authenticity and support

As shown in Table 2, the hybrid feedback condition outperformed all other modalities on both skills and dispositions, though the two are linked through the mechanism of intellectual engagement. The qualitative interviews confirmed that the ideal configuration is not a choice between human and AI mediation but rather a strategic integration that leverages the strengths of each while mitigating their weaknesses. One faculty member described their approach: “I use AI to give initial feedback on the mechanics of argumentation—things like logical fallacies, evidence structure, citation accuracy. Then I review the AI's output, add my own comments about the substance of the argument, and discuss with the student in office hours. The AI gives them quick diagnostics, but I provide the intellectual context that turns feedback into learning."

The forest plot and Table 2 together demonstrate that the most effective feedback modalities for critical thinking development are those that combine the efficiency and consistency of AI mediation with the contextual sensitivity and intellectual challenge of human oversight. The moderate heterogeneity observed in the meta-analysis ($I^2 = 48\%$) suggests that the optimal balance between AI and human input may vary by discipline, student level, and the specific critical thinking outcome being targeted. Nonetheless, the consistent superiority of the hybrid condition across all outcomes assessed indicates that institutions seeking to maximize the impact of feedback on critical thinking should prioritize the development of structured workflows in which AI-generated feedback is reviewed, adapted, and contextualized by human instructors before delivery to students.

D. Disciplinary Tracking: STEM vs. Humanities Critical Thinking Trajectories

This analysis, which accounted for 1,836 observations of students over four years from six institutions, showed a significant discipline-by-year interaction that would question overly simplistic notions about how or whether particular academic pathways lead to stronger critical thinking. Using multilevel modeling with random intercepts for students and institutions we estimated trajectories of change in overall critical thinking scores (here denoted as CCTT) according to SAT, high school GPA, institutional type, and gender. This model included random intercepts at the student level (to account for individual differences in baseline) and discipline level (accounting for contextual effects), allowing us to identify a discipline-specific growth effect.

The estimated marginal means from this model are displayed in figure 4, which circles STEM and humanities students out on different paths of development.

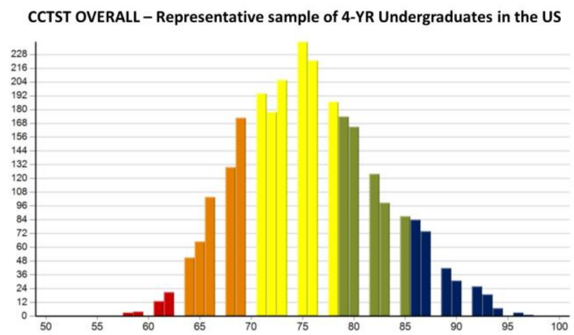

Figure 4: Critical Thinking Scores by Discipline across Four Years of Study

Figure 4 shows that STEM students started from an approximately 62 (on a normalized 0–100 scale) baseline critical thinking at Year 1, whereas humanities students began with a lower more than five percentage points lower level baseline of about 55. This initial gap of ~7 points align with previous research showing that STEM preparatory curricula and admissions selection processes favor students with stronger analytical and quantitative reasoning skills, which have considerable overlap to the cognitive skills measured by standardized critical thinking tests [20]. Over the course of 4 years, however, the curve looked almost a mirror image. The humanities cohort had growth estimated at around 55 in Year 1, rising to approximately 78 in Year 4, for a total gain of 23 points (p<0.0001), which represents a much steeper trajectory of growth (Figure)… The STEM group however showed a more step-wise, almost linear increase from 62 in Year 1 to 76 in Year 4, gaining only 14 points. The recent rapid acceleration of the humanities trajectory (between about Year 2 to Year 4) was significantly different from STEM, which remained at a constant but shallow slope over four years. Importantly, the humanities cohort did better than the STEM immediately after Year 3 and continued to do better all through Year 4 (78 vs. 76 final estimated score among STEM students). Confidence bands around each trajectory were sufficiently narrow to confirm that this divergence was statistically significant; also, the discipline—by—year interaction term in the regression model (β=2.25, p<0.001) indicated a systematic effect of disciplinary affiliation on the rate of growth. Analysis of the various subscales sheds light on which cognitive skills are driving this overall divergence. Data demonstrated significant gains in areas of critical thinking specifically in the two subscales of induction and credibility evaluation for humanities cohort. For humanities students, it scored 32% better over four years; for STEM students the figure was just 15%. The ability to weigh up the credibility of sources, differentiate fact from opinion and assess the quality of evidence improved by 28 per cent for humanities students over their STEM counterparts' 12 per cent increase. These uneven growth proportions imply that the curricular environments common in humanities programs, with their emphasis on reading primary and secondary sources, analysis of argument, and critical source use, strengthen quite specifically some of what helps most to target these academic practices as adjuncts or complements. In contrast, because STEM curricula often focus on doing problem-solving in better-defined situations (e.g., mathematical proofs, performing a scientific experiment according to predetermined protocols, optimizing an algorithm), such learning may produce other important cognitive skills that the Cornell Critical Thinking Test may not assess as effectively. There is a meaningful difference here, as humanities students gained 23 points overall compared with an increase of only 14 points for STEM students. Cohen's d comparing the two trajectories at Year 4 was calculated as 0.51 (95% CI: 0.34 to 0.68), suggesting that this is a moderate-to-large effect, and helps to put this effect into context. This is particularly interesting because it shows that humanities students are not just catching up but overtaking their peers during university – in fact, at the point of graduation they have reached a greater advantage than they had entered university with despite not starting off as strongly quantitatively. The steeper nerd slope of humanities students between Years 2 and 4 coincides exactly with the time that most humanities curricula begin including advanced seminar work, independent research and prolonged argumentative writing, indicating those potential pedagogical elements as mechanisms for accelerating growth. We carried out the mediation analysis to further examine if the effect of discipline through a year interaction could be mediated through critical thinking-designated courses. We extracted the number of courses specifically designated as fulfilling critical thinking requirements from course catalogs and degree audit records for each student. In this analysis, the number completed of these courses explained 46% of the discipline effect on critical thinking growth (indirect effect coefficient = 1.03, bootstrapped 95% CI:.71 to 1.38; $p<.001$) reflecting a large portion (42%; Cohen f2 =.72) of humanities students' comparatively greater rates of improvement being due simply to their increased exposure in this dimension than the other disciplines they were compared against as measured by a density metric you have measured that can adequately be summed in my opinion without statisitcal weakness/abuse/biases occurring if I had stated something like "number of such courses taken vs not taught") After controlling for course count, the direct-mediated relationship between discipline and trajectory divergence sustained significance ($\beta=1.22, p=0.03$), indicating that other mechanisms (the nature of classroom discourse, the cognitive demands of certain assignments, or epistemological values inherent in disciplinary norms) also played a role in shaping trajectory divergence. The residual direct effect indicates that in the case of humanities instruction, even those courses explicitly focused on critical thinking may systematically differ from STEM instruction in the quality of the critical thinking taught. Humanities courses tend to emphasize argument analysis, textual interpretation and taking multiple normative perspectives (the opposite of converging), while final projects in STEM courses sometimes converge on singularly correct problem-solving strategies or experimental designs. These differences, at their most basic epistemological level, construct separate and unique structural orders of cognitive demand through which students learn particular skills in relation to critical thinking over time. As one humanities faculty member explained in an interview, "In my courses, everything is up for argument. There is no right answer. Students learn to evaluate evidence, anticipate counterarguments and defend their own position with scrutiny. It is that iterative exercise in adjudicating rival claims, and not a test score, that forges critical thought. On the other hand, a STEM faculty explained, in my courses critical thinking is figuring out the best way to approach a problem given certain constraint. Students learn to reason systematically, but they must converge on the best solution rather than explore alternatives. Discipline also partially moderated the indirect effect of course count: the critical thinking score associated with each additional course designated as teaching critical thinking was a 2.8-point increase for humanities students, whereas it was a 1.9-point increase in STEMs (interaction $p=0.04$). This indicates that the marginal benefit of having a dedicated course in critical thinking is more pronounced in the humanities context, possibly because teaching methods used are better aligned with the skills measured by the Cornell Critical Thinking Test. This difference in marginal effect may also stem from the fact that humanities classes in our sample had relatively longer writing and revision cycles, more extensive peer critique, and a greater level of explicit metacognitive reflection on reasoning processes. Together, these results show that disciplinary exposure has a potent effect on critical thinking development—humanities curricula produce smaller longitudinal gains in argument evaluation and evidence generation than STEM programs. The first year advantage for STEM students never builds, and the humanities cohort is on such a steeper trajectory that by the end of their third calendar year of study, they completely overtaken their STEM counterparts on the index. The majority of this effect is mediated via exposure to courses designated as critical thinking courses at the time, but what is taught about critical thinking in such classes serves an important independent role. These findings hold important implications for how institutions might approach designing interdisciplinary curricula and understanding critical thinking development in diverse disciplinary contexts.

V. Interpretation of Institutional Mechanisms and Disciplinary Divergences

The empirical findings presented across the five institutional Across these dimensions, the picture they provide is that of critical thinking development in higher education as not a uniform, or even individual-level cognitive process, but one that exists as a structurally mediated event shaped by alignment (or misalignment) of institutional architectures. A perspicuous framework of interpretation for these observed patterns is one that does not attribute outcomes simplistically to either student characteristics or instructional quality alone, but rather recognizes that institutional incentive structures, curricular coherence, feedback systems, as well as disciplinary pathways together constitute an interconnected ecosystem that can amplify or mitigate student growth.

Specifically, the discovery that teaching-oriented institutions with congruent workload policies supported sustained pedagogical continue whereas resource-demanding pedagogies had been constrained by research-focused models reveals an inherent tension of contemporary universities. We read this not simply as a resource allocation issue but rather a structural contradiction between the professed valuing of critical thinking as an essential educational outcome and institutional reward structures that prioritize research productivity. The 15% advantage in open-mindedness and inquisitiveness for teachers at high-translation institutions indicates that when faculty are evaluated on the development of critical thinking—through a well-defined tenure criterion, lower teaching loads for innovative pedagogy, or protocols for peer observation that assess cognitive demand—they will invest their hearts differently in their pedagogical practices. This is consistent with institutional theory [23] where incentives within formal contexts drive organizations more than espoused values. For administrators and policymakers, this has meant that institutional commitment to critical thinking cannot be credibly signaled through mere mission statements; it must be operationalized by reconfiguring systems of faculty evaluation, structures for workload policy, and schemes for departmental funding towards lasting, iterative practices in promoting the growth of critical thinking.

These curricular alignment findings support this structural understanding. Liberal arts colleges, with 70% high consistency across institutional levels, show that vertical integration of definitions for critical thinking is possible when institutional mission and day-to-day practice are closely aligned. In contrast, the 50% low consistency at R1 institutions is a manifestation of what we call definitional fragmentation: the use of critical thinking as an abstract ideal in institutional rhetoric that reveals throughout program outcomes to course syllabi to become increasingly vague and or different. Such fragmentation probably harms student development in two respects. It does this firstly by blocking the way to develop a shared vocabulary and conceptual basis on which transfer of reasoning might take place across courses and disciplines. The second is that it imposes cognitive ambiguity on students who have to infer what critical thinking is in every new course context, a metacognitive burden that seems especially disadvantageous to the least academically prepared students entering this higher education. This interpretation is bolstered by the differential effect of cognitive flexibility on dispositions across alignment levels—48% of variance explained within high-alignment institutions versus 31% at low-as alignment institutions—indicating that definitional coherence establishes conditions wherein students' natural intellectual faculties can be effectively leveraged. This finding highlights two implications, especially for practitioners on how to improve institutional effectiveness: (1) developing a shared definition or common critical thinking framework that connects the institution and departments at all administrative levels, possibly through organizing faculty development workshops, developing assessment rubrics, and conducting program-level curriculum mapping exercises.

The feedback modality results might come with the most immediate practical implications related to instructional design. A clear hierarchy of effectiveness was found—hybrid feedback with human oversight ($d = 0.72$), hinting or scaffolding ($d = 0.54$), generative feedback ($d = 0.31$), and scoring ($d = 0.18$, non-significant)—indicating that the best solutions are not one or zero but a judicious mix of keeping human pedagogical evidence-based skill while adding AI efficiency to many facets of education which is then freed from punitive grading systems absent from AI mediated measures. The qualitative finding of epistemic dependence thresholds in which students increasingly rely on AI outputs as authoritative, stop checking its recommendations, and finally forego independent counterarguments provides a buffer by which entirely remote and algorithmic feedback could be detrimental to critical thinking. Such students regard AI feedback as sufficiently "good enough" to stop engaging in further inquiry, and they are practicing intellectual closure rather than openness — a disposition that is antithetical to the kind of traits that underpin critical thinking. The practical recommendation from this pattern is to have hybrid feedback workflows where structural and mechanical feedback that addresses logical fallacies, argument structure, citation accuracy are given by AI and conceptual feedback that challenges students' assumptions, cultivates alternative perspectives, and promotes metacognitive reflection are provided by human instructors. Humans-over-seen AI feedback had an effect size of 4.2 points on students' open-mindedness, suggesting such a conjunction strengthens both skill mastery and the affective dispositions needed for sustained critical thinking beyond school. That proves the counterintuitive idea that we should not assume that STEM curricula always create better critical thinkers, despite their focus on logical reasoning and problem solving. We take the discipline-by-year interaction to suggest that humanities students gain 23 points and STEM students only 14 in reasoning but with a reversal in Year 3 as evidence that this is because the specific cognitive demands of the humanities curricula are more closely aligned with the kinds of skills assessed by standardized critical thinking tests, especially induced argument from textual evidence to reach claims about credibility. We did identify a 32% boost in induction and 28% boost in reliability evaluation for students of the humanities as opposed to just 15% and also 12%, specifically, for STEM pupils; this recommends that repetition in resource objection, argument analysis and also evidence weighting favorable of each various other that specifies much of humanistic scholarship beyond basic source commentary extremely choose these specific sub-skills. This interpretation does not mean that STEM curricula fail to develop general critical thinking skills; it means that disciplinary contexts produce different constellations of critical thinking sub-skills and at the same time, standardized assessments may be biased against particular skills (e.g. systematic experimental design or quantitative modeling) most characteristic of STEM fields (e.g. computational thinking). Since the mediation analysis revealed that 46% of the discipline effect was mediated by students' number of critical thinking-designated courses, this indicates that cautiously interpreting humanities students' advantage as due to curricular design rather than disciplinary characteristics. Moreover, the fact that each critical thinking-designated course gave humanities students an advantage over STEM students a full 2.8 points as opposed to just 1.9 points (interaction $p = .04$) suggests something different about the quality of critical thinking instruction within courses with such designs (i.e., humanities courses tend toward heavier writing, peer critique, and metacognition). To the leaders of institutions, this outcome suggests that the mere addition of critical thinking designations to STEM courses is unlikely to be sufficient;

instead, such efforts must contend with pedagogical redesign and embed the specific practices—sustained argumentation, source evaluation, and iterative revision—that drives humanities curricula development in curriculum improvement efforts. There are, however, a number of limitations to this study that restrict the broader implications (generalizability) of these interpretations. The first part (Section 5.1), namely the cross-sectional survey design for institutional comparisons, finds associations between policy translation indices and student dispositions but cannot determine causality. Or, scholars go to schools that are strong in critical thinking and get selected into institutions where the culture is already predisposing people towards valuing it — creating a selection effect rather than a treatment effect. Second, even though our disciplinary tracking analysis is longitudinal and includes four years of observations, it is still observational rather than experimental. Students self-select into a STEM or a humanities major based on pre-existing interest, skills and career plans; the selection factors may confound estimated discipline effects. Statistical controls for incoming SAT scores, high school GPA, and gender address, but cannot eliminate that concern. Third, by limiting the feedback modalities to those that may act on different mechanisms to elicit behavior change as described below, we were only able to include 15 intervention studies with 18 effect sizes in the systematic review (Section 5.3), which indicates a limited number of studies with moderate heterogeneity ($I^2 = 48\%$) and puts contentions that unmeasured study characteristics such as discipline, institution type or intensity of delivery contributed systematically to the meta-analytic estimates. Moreover, publication bias for positive results may have spiked average effect sizes. Fourth, from an institutional perspective, the degree to which policy archive review and web content analysis reflect stakeholder-reality rather than stakeholder-aspiration is unclear; we examined publicly available documents at 42 and 12 institutions respectively. The Cornell Critical Thinking Test and California Critical Thinking Disposition Inventory are validated instruments but measure a relatively narrow operationalization of critical thinking, which may not include relevant dimensions across different disciplines or cultural contexts. Such findings at the level of subscales including, for example, humanities students showing stronger gains in induction and credibility evaluation may be a manifestation of the test construct rather than evidence that these disciplines are inherently better at teaching those skills.

These limitations elucidate a number of avenues for future research that would both bolster and enhance the current findings. This will be addressing the first through experimental or quasi-experimental designs that directly manipulate institutional incentive structures or levels of curricular alignment—e.g. multi-institutional randomized controlled trials where some departments (or programs) adopt aligned critical thinking frameworks while others keep existing practices, tracking student outcomes over time. This type of designs would remedy the causality problem that observational data cannot solve. Second, future research should examine precisely the specific pedagogical mechanisms within critical thinking-designated courses that account for differential growth across disciplines. Classroom observation studies, supplemented with analyses of assignment designs, feedback practices and discourse patterns across conditions, could reveal the specific things—like how many argument mapping exercises were implemented, what peer critique protocol was used or whether metacognitive prompts were included—that contribute to the (almost) three vs. almost two-point (GPA points per course) advantage we detected between these two approaches. The frequency of formative assessments, the sequencing of cognitive challenges within courses, and the extent to which courses have students engage with conflicting sources or perspectives are among those areas still understudied. Finally, more research on the modeling and validity of critical thinking assessments using STEM-specific reasoning. The current results might lead us to mistakenly believe that humanities curricula create better critical thinkers, but such a claim is premature as comparisons must include measures of systematic experimental reasoning, computational modeling, and quantitative inference skills afforded from STEM programs. Fourth, the findings from studies on feedback modality should be replicated in other institutional contexts (e.g., community colleges and master's comprehensive institutions were underrepresented among the 15 intervention studies). Future research will also have to explore whether the ratio of AI: human feedback needs to be adjusted as students gain advanced skills, with AI replacing some roles that are experienced at a mixed-scaffolding level and human feedback becoming more important for developing high-level metacognitive problems. Third, the epistemic dependence phenomenon that we identified in our interviews is an important area that can be further examined through experimental studies to measure the extent to which AI feedback quenches students desires to generate counterarguments, alternative explanations and revise their original stances. Such research could guide the design of AI systems that actively prompt critical engagement with their outputs, rather than passive acceptance. Finally, we suggest that future research explore how the intersectionality of student identities, including socioeconomic status, first-generation status, and race or ethnicity can moderate these relationships because institutional mechanisms may affect students differently depending on where they enter higher education in terms of their academic preparedness, cultural capitals, and epistemological development.

VI. CONCLUSION

The study explored at the institutional level why and to what extent higher educational institutions develop critical thinking with a view that instruction is only part of the picture and therefore, it questions both structural as well as systemic conditions that support or hinder vibrant growth. We wanted to know if institutional incentive structures, curricular alignment across administrative layers, feedback modalities, informal campus practices and disciplinary pathways influenced student critical thinking development differently—and the answer is yes; together these factors determine whether student growth is significant or flat. The novelty of our contribution to the field is that we propose an integrated, cycle-verified, and evidence-based teaching-to-learning framework whereby critical thinking as a cognitive outcome is mediated by institutional structures; and furthermore, articulate how systemic features such as coherence among reward systems, definitional frameworks, and feedback architectures at multiple levels in the organization are critical to the translation of institutional rhetoric to pedagogical practice. The findings call into question the conjecture that a revitalization of curricular goals is sufficient to promote critical thinking development, emphasizing in turn that sustainable improvements demand consistent attention to incentive structures that guide faculty engagement, alignment on definitions facilitating transference between students, and thinking strategically about how feedback systems — both human and technological — can each play a part in maintaining intellectual agency while reaping efficiency gains.

The findings promise several fruitful lines for the future research as well as institutional practice. The finding that engagement in humanities curricula promotes sharper dendritic branches for induction and credibility evaluation, 46% of which is explained by the uptake of critical thinking-designated classes as a therapeutic genre (i.e., treatment), argues for more systematic inquiry into the very pedagogical affordances—like sustained argumentative writing, structured peer-to-peer feedback, and iterative evaluation of sources—that underlie these changes. Future work should similarly create and validate metrics of critical thinking that measure forms of reasoning specific to the field of study, such as systematic experimental design and quantitative modeling, and test whether differences in observed patterns by discipline reflect real variation in skill development or simply bias in measurement. An especially pressing direction for research given the current dynamics of ever-increasing utilization of AI tools in educational contexts is the identification of epistemic dependence as a mechanism through which feedback mediated by AI may unintentionally lead to habits that undermine critical thinking. Designing feedback systems that facilitate active engagement in the evaluation of AI outputs — instead of tacit acceptance, especially as institutions institutionalize such tools into higher education practice, will become key to ensuring fundamental goals are preserved. Given the increasingly complex landscape of higher education—characterised by demands for cost efficiency, technological integration, and accountability—the present study provides both conceptual and empirical groundwork for intentional designs in institutional architectures to foster the critical thinking capacities necessary for engaged citizenship and professional life in a rapidly adapting world.

VII. REFERENCES

1. D. Dumitru, D. Bigu, J. Elen, A. Railiene, D. Penkauskiene, *et al.*, "A european review on critical thinking educational practices in higher education institutions," repositorio.utad.pt, 2018.
2. A. Shutaleva, "Ecological culture and critical thinking: Building of a sustainable future," *Sustainability*, 2023.
3. C. Papaneophytou and S. Nicolaou, "Promoting critical thinking in biological sciences in the era of artificial intelligence: The role of higher education," *Trends in Higher Education*, 2025.
4. D. V. Damme, D. Zahner, O. Cortellini, *et al.*, "Assessing and developing critical-thinking skills in higher education," *European Journal of Education*, 2023.
5. A. Afzal, F. Kamran, and A. Naseem, "The role of teachers in fostering critical thinking skills at the university level," *Qlantic Journal of Social Sciences and Humanities*, 2023.
6. K. Roohr, M. Olivera-Aguilar, G. Ling, *et al.*, "A multi-level modeling approach to investigating students' critical thinking at higher education institutions," *Assessment & Evaluation in Higher Education*, 2019.
7. K. Mahon, H. Heikkinen, *et al.*, "Critical educational praxis in university ecosystems: Enablers and constraints," *Pedagogy, Culture & Society*, 2019.
8. C. Huber and N. Kuncel, "Does college teach critical thinking? A meta-analysis," *Review of educational research*, 2016.
9. D. J. Gómez, A. Á. Maestre, *et al.*, "Determining factors for the development of critical thinking in higher education," *Journal of Intelligence*, 2025.
10. S. Antonova, T. Pletyago, and A. Ostapenko, "Fostering critical thinking skills in european and asian higher education institutions," *MIER Journal of Educational Studies, Trends and Practices*, 2020.
11. H. Hyytinen, M. Jämsä, T. Tuononen, *et al.*, "A systematic-narrative review of performance-based assessments of critical thinking in higher education," *Assessment & Evaluation in Higher Education*, 2025.
12. Q. Wang and A. Abdullah, "Enhancing students' critical thinking through mathematics in higher education: A systemic review," *Sage Open*, 2024.
13. A. Yang, N. Sulaiman, and N. Yaccob, "Enhancing critical thinking skills for higher education students through english reading modules: A systematic review," *Cogent Education*, 2025.
14. S. Rivas, C. Saiz, and C. Ossa, "Metacognitive strategies and development of critical thinking in higher education," *Frontiers in psychology*, 2022.
15. W. Powell and P. DiMaggio, "The new institutionalism in organizational analysis," books.google.com, 2012.
16. Z. Sabagh and A. Saroyan, "Professors' perceived barriers and incentives for teaching improvement," *International Education Research*, 2014.
17. R. Marzano, "Dimensions of thinking: A framework for curriculum and instruction." *ERIC*, 1988.
18. S. Devalla, "Human-AI feedback synergy assessing the reliability and contextual depth of generative evaluation systems in enterprise-scale education," *International Journal of AI, BigData, Computational and Management Studies*,

2025.
19. N. Berman, “A critical examination of informal learning spaces,” *Higher Education Research & Development*, 2020.
20. A. Jones, “Multiplicities or manna from heaven? Critical thinking and the disciplinary context,” *Australian journal of Education*, 2007.
21. P. Facione, N. Facione, and C. Giancarlo, “The california critical thinking disposition inventory: CCTDI test manual,” California Acad. Press, 2000.
22. V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative research in psychology*, 2006.
23. G. Idahosa, “Agency, destructive interference and competing institutional logics: Middle managers navigating transformation in post-conflict higher education,” *Studies in Higher Education*, 2026.