From Skill to Being: The Embodied Education Turn and Falsification-Based Assessment in English Education in the Age of Artificial Intelligence

Shi Wei

2026-06-24

ChinaRxiv

Abstract

In response to the existential anxiety of English education in the age of artificial intelligence, most existing mainstream solutions have failed to move beyond the rut of instrumentalism. This paper attempts to propose a "falsification benchmark" for the evaluation of English education: rather than seeking to verify whether education has taken place, it safeguards the bottom line of education by falsifying those instructional designs that structurally exclude the possibility of students "becoming different persons through English." Drawing on embodied cognition theory and second language acquisition research, this paper attempts to define three exclusive dimensions of embodied education in foreign language learning: the cross-cultural reconstruction of the L2 body schema, the generation of cross-cultural heterogeneous intersubjectivity, and the dual expansion of the boundaries of the bilingual world. The study shows that the inexhaustibility of experience is one of the core boundaries that distinguish humans from algorithms. On this basis, the paper further differentiates between "evaluation logic" and "hosting logic," and proposes hosting tools such as critical incident analysis and sound walls to provide a witnessing space for traces of transformation that cannot be evaluated. The falsification benchmark is not intended to "prove" the success of education, nor is it meant to replace quantitative evaluation; rather, it serves as a logical defense of the bottom line of education. This paper seeks to provide an ontological reference for the value reconstruction of English education in the age of artificial intelligence, as well as theoretical resources for constructing an autonomous knowledge system for foreign language education in China.

Full Text

From Skill to Being: The Embodied Education Turn and Falsification-Based Assessment in English Education in the Age of Artificial Intelligence

Shi Wei

(School of Foreign Languages, Jingdezhen University, Jingdezhen 333400)

Abstract: In response to the existential anxiety surrounding English education in the age of artificial intelligence, existing mainstream approaches have largely failed to move beyond the constraints of instrumentalism. This paper attempts to propose a "falsification-based baseline" for English education assessment—not by seeking to prove whether education has occurred, but by falsifying instructional designs that structurally exclude the possibility that students may "become different persons through English,"thereby safeguarding the educational bottom line. Drawing on embodied cognition theory and second language

acquisition research, this paper attempts to define three exclusive dimensions of embodied education in foreign languages: the cross-cultural reconstruction of second-language body schemas, the generation of intersubjectivity among culturally heterogeneous subjects, and the dual expansion of bilingual world-boundaries. The study shows that the inexhaustibility of experience is one of the core boundaries that distinguish human beings from algorithms. On this basis, the paper further differentiates between the "logic of assessment" and the "logic of bearing witness," and proposes bearing-witness tools such as critical incident analysis and voice walls to provide evidential space for traces of transformation that cannot be assessed. The falsification-based baseline is not intended to "prove" educational success, nor does it aim to replace quantitative evaluation; rather, it serves as a logical line of defense for the educational bottom line. This paper attempts to provide an ontological reference for the reconstruction of values in English education in the age of artificial intelligence, and to offer theoretical resources for building an autonomous knowledge system for foreign language education in China.

Keywords: falsification-based baseline; embodied education; English education; artificial intelligence; second language acquisition; foreign language education assessment

Classification No.: H319.3

From Skill to Being:

The Embodied Education Turn and Falsification-Based Assessment in

English Education in the Age of Artificial Intelligence

Wei SHI (E-mail: 20031@jdzu.edu.cn)

(School of Foreign Languages, Jingdezhen University, Jingdezhen 333400, China)

Abstract: Addressing the existential anxiety in English education in the age of artificial intelligence, this paper argues that existing mainstream responses have largely failed to break free from instrumentalism. It proposes a "falsification-based baseline" for English education assessment—one that falsifies instructional designs structurally precluding students from "becoming different persons through English," thereby safeguarding the educational bottom line. Drawing on embodied cognition and second language acquisition research, the paper attempts to define three exclusive dimensions of foreign language embodied education and distinguishes the "logic of assessment" from the "logic of bearing witness," proposing tools such as critical incident analysis and voice

walls for witnessing unassessable traces of transformation. The baseline does not aim to prove educational success, nor does it intend to replace quantitative evaluation; it serves as a logical defense of the educational bottom line. This paper attempts to provide an ontological reference for value reconstruction in the age of AI, and offers theoretical resources for building an autonomous knowledge system in Chinese foreign language education.

Keywords: falsification-based baseline; embodied education; English education; artificial intelligence; second language acquisition; foreign language education assessment

1. Introduction: Misplaced Anxieties and Research Gaps in English Education in the Age of Artificial Intelligence

1.1 Misreadings of Crisis Manifestations and Mainstream Interpretations

Since 2024, the global language-education community has been undergoing continuous upheaval. The Middlebury Institute of International Studies at Monterey in the United States ceased operations, and many universities in China adjusted or suspended enrollment in master' s programs in English. These events have been widely interpreted as the direct consequences of the impact of generative artificial intelligence technologies. Behind this interpretation lies an unexamined presupposition: that the core value of English education resides in its instrumental outputs. It is precisely this presupposition that has made "machines can do it better" a fatal blow to English programs.

In response to the crisis, academic discussion has mainly focused on two paths. The first is "addition," namely cultivating interdisciplinary talent such as "English + law" and "English + economics and trade," in an attempt to maintain the irreplaceability of the discipline through the layering of skills. The second is "multiplication," namely using artificial intelligence to empower teaching and emphasizing human–machine collaboration. At the policy level, responses have also been made: five departments, including the Ministry of Education (2026), issued the "*Artificial Intelligence + Education*" *Action Plan*, stressing the deep integration of AI and education. Yet its framework still focuses on "application orientation" and "intelligent literacy," without touching on the ontological questioning of English education. Although these two types of solutions differ in their approaches, they share the same logical starting point: the essence of English is a skill system that can be decomposed and measured. If this presupposition is not reflected upon, the anxiety will be difficult to truly dispel—when generative artificial intelligence can complete all kinds of standardizable language tasks at zero marginal cost, "more complex combinations of skills" are merely a suspended sentence that delays the verdict.

1.2 Formulation of the Core Research Question

The core question this article seeks to address is: If generative artificial intelligence can efficiently complete standardizable language tasks such as translation, writing, and information extraction, what exactly is the irreplaceable core value of English education? On what basis should we distinguish genuine "person cultivation" from "skills training" that can be replaced by algorithms?

This article explores, at the ontological level, the distinctive meaning of embodied person cultivation in foreign-language education, and attempts to construct falsification-oriented evaluation criteria so as to provide a theoretical reference for value reconstruction.

1.3 Review of Domestic and International Research Trajectories and Their Limitations

Around the above issues, existing research has formed three main trajectories.

The first trajectory focuses on the relationship between artificial intelligence and foreign-language education. Cai Jigang (2025) demonstrates through experiments that using ChatGPT for academic-paper writing can significantly promote the improvement of university students' critical thinking ability; however, it also reveals that technological impact is only an external factor, while the fundamental issue lies in the ambiguous positioning within the discipline itself. Wen Qiufang (2024; 2025a; 2025b) systematically discusses the research-topic framework, teaching-method reforms, and curriculum-model innovations of foreign-language education in the age of artificial intelligence, advocating a shift from "language-skills training" to "the cultivation of intercultural competence and critical thinking." Warschauer and Xu (2024) further point out that the impact of generative AI on language education is not a matter of technological iteration, but a redefinition of the very concept of "language ability." The common contribution of these studies lies in their acute recognition of the existence of the crisis; however, their proposed responses still remain at the level of curriculum adjustment and ability recombination, without sufficiently engaging the ontological question of "what English education ultimately is."

The second trajectory introduces embodied cognition theory into foreign-language education research. Wang Yin (2022; 2023) systematically expounds the theoretical framework of embodied-cognitive linguistics, emphasizing that language understanding is rooted in the interaction between bodily experience and cognitive processing. Ni Chuan

Bin (2023), in a meta-analysis of 56 empirical studies on embodied effects in foreign languages, showed that factors such as embodiment, teaching methods, age of acquisition, foreign-language proficiency, and linguistic distance can all influence embodied effects in foreign languages. The study by Jiang Meng and Zhao Sisi (2014) further revealed the embodied mechanism of affordance extraction in discourse comprehension. However, existing studies have transformed "em-

bodiment" from a philosophical concept into a teaching strategy and promoted the shift of classroom activities from "disembodiment" to "bodily participation." Their limitation, however, lies in narrowing "embodiment" to "bodily activity supporting memory" or "multimodal teaching" ; they have not fully completed the theoretical leap from "embodied teaching" to "embodied education," still less have they anchored the foreign-language-specific embodied mechanisms that distinguish foreign-language learning from mother-tongue education.

The third line of inquiry concerns reform of the evaluation system in foreign-language education. Existing evaluation takes quantifiable skill outputs as its core indicators and standardized tests as its principal tools. The fundamental predicament of this system is that it can measure only those skills that can be replaced by algorithms, while it cannot touch the essence of educating persons. Scriven's (1967) classic distinction between formative evaluation and summative evaluation revealed the multiple functions of evaluation; Broadfoot and Black (2004), reviewing the development of educational evaluation over the previous decade, pointed out that the paradigm shift toward formative evaluation was still constrained by the inertia of traditional evaluation culture; Black and Wiliam (1998) argued that formative evaluation should be "assessment for learning," rather than merely "assessment of learning" ; Torrance and Pryor (1998), in a systematic classroom-level investigation of the implementation mechanisms of formative evaluation, found that teachers' everyday assessment practices often slid unknowingly toward a summative logic. Torrance (2007) more profoundly warned of the risk of alienation in evaluation—when assessment procedures completely dominate the learning experience, "standards compliance" replaces "genuine learning." As Biesta (2010) criticized, when education is simplified into measurable "learning outcomes," education itself is hollowed out of meaning. This predicament also stands in sharp contrast to the requirement in the *Overall Plan for Deepening the Reform of Educational Evaluation in the New Era* (2020) to "eliminate utilitarian evaluation tendencies that focus only on scores and only on advancement to higher levels of schooling." The existing evaluation system for foreign-language education has not yet formed an implementable framework suited to the essence of educating persons.

In sum, existing research has three major areas requiring further deepening: first, insufficient ontological reflection—few studies directly inquire into the irreplaceable existential foundation of English education in the age of artificial intelligence; second, the generalization of embodied research—the concept of embodiment has been downgraded to a teaching strategy and has not been anchored in the distinctive attributes of foreign-language learning; third, a blank space in evaluation criteria suited to the essence of educating persons—the existing evaluation system has difficulty effectively distinguishing genuine from pseudo education.

1.4 Academic Contributions and Structural Arrangement of This Article

In response to the above shortcomings, this article offers two contributions. At the theoretical level, it attempts to define the three exclusive dimensions of embodied education in foreign languages (Chapter 2). At the methodological level, it attempts to construct falsification-oriented evaluation criteria, using falsification to exclude teaching designs that only seemingly educate persons and thereby uphold the bottom line of education (Chapter 3). Chapter 4 demonstrates the application of the criteria, and Chapter 5 summarizes the conclusions and implications.

The structure of the full article is as follows: Chapter 2 constructs the ontological connotations; Chapter 3 proposes falsification-oriented criteria; Chapter 4 demonstrates their application; and Chapter 5 summarizes the conclusions and implications.

2. Ontological Construction of Embodied Education: The Irreducible Core of Foreign-Language Education

2.1 The Theoretical Origins of "Embodiment" and Its Ontological Transcendence

In cognitive science, the concept of "embodiment" points to a core insight: language comprehension is rooted in sensorimotor experience. The neuroscientific research of Gallese and Lakoff (2005) shows that understanding verbs such as "grasp" will activate large···

the brain' s motor cortex; abstract concepts, too, often acquire a bodily basis through spatial metaphors. The phenomenological tradition provides a philosophical foundation for this insight (Merleau-Ponty, 1945/2012). We do not use the brain to "compute" the meaning of language; rather, we use the whole body to "feel" and "perform" language. For example, when we learn to apologize in English, we do not merely memorize the word "sorry" ; we learn a new bodily posture—the degree of tension in the vocal cords, the point at which the gaze falls, and the rhythm of breathing all differ from those involved in apologizing in the mother tongue. This is the core of "embodiment" : language is an action of the body, not a symbol in the mind.

On this basis, this paper advances a stronger ontological thesis: English is not only an external instrument for human beings, but also a core medium that participates in the construction of the subject. Genuine English education does not enable students to "possess" a skill; rather, it enables students to "become different people through English." As Heidegger (1927/1962) reveals, when a skilled craftsman uses a hammer, the hammer has already been integrated into his bodily action and is no longer perceived as an object; the highest state of English education is likewise this: not "I speak English," but "I become a certain kind of person through English."

The above ontological thesis is grounded in a phenomenological fact: each self is, in existential terms, an irreplaceably unique being. This is not an "individual difference" in the statistical sense; rather, it means that genuine educational transformation can only be an internal event of this unique being, not the achievement of external indicators. Therefore, whether a curriculum safeguards the possibility of such an internal event cannot easily be judged through standardized measurement of "outputs" ; instead, it must be examined by asking whether the curriculum structure has systematically excluded that possibility—this is precisely the logical root of the falsification criterion. To go a step further, algorithmic "personalization" is a spatial combination of variables, whereas human embodied experience is temporal: when the same subject faces the same situation at different times, the ontological depth of that subject is already different, because every encounter irreversibly changes the subject itself. This irreversibility within temporality constitutes one of the ontological boundaries that algorithms find difficult to reach.

2.2 The Three Exclusive Dimensions of Embodied Person-Formation in Foreign-Language Education

The following attempts to define the exclusive connotations of embodied person-formation in foreign-language education from three dimensions: corporeality, relationality, and generativity. Each dimension distinguishes the essential differences between mother-tongue education and foreign-language education, and anchors them in the distinctive contexts of second-language acquisition.

(1) Corporeality: Cross-Cultural Reconstruction of the Second-Language Body Schema

Mother-tongue acquisition is a process in which the body schema and the language system are generated synchronously. As children learn their mother tongue, they also acquire a matching set of articulatory muscle patterns, pragmatic bodily postures, and breathing rhythms for emotional expression. This body schema tends to stabilize around adolescence and becomes the "mother-tongue body."

Foreign-language learning is entirely different. On the basis of an already stabilized mother-tongue body schema, it completes a dual process of "de-mother-tonguing" and "re-embodiment." Learners need to break free from the muscular memory of mother-tongue pronunciation, overcome the bodily inertia brought about by mother-tongue word order, and at the same time establish a new English body schema. The degree of vocal-cord tension, gaze placement, and bodily posture involved in apologizing in English differ systematically from those involved in apologizing in the mother tongue. This is not "performance," but bodily learning of another mode of existence. When learners sense the changes in bodily posture brought about by this difference, language is more likely to begin to become "embodied."

(2) Relationality: The Generation of Cross-Cultural Heterogeneous Intersubjectivity

In the mother tongue, intersubjectivity occurs within the scope of shared cultural scripts. Speakers and listeners jointly possess cultural tacit understandings about "how to express opposition," "how to apologize," and "how to make requests." Such tacit understanding makes communication smooth, but it also limits the extent to which subjectivity can be expanded.

Intersubjectivity in a foreign language, by contrast, faces conflicts and negotiations among cultural scripts. When a Chinese student says in English for the first time to a foreign peer, "I disagree with your point of view," the student is not merely learning and using a sentence pattern, but is also confirming the boundaries of the self within a heterogeneous language. This process is accompanied by a genuine sense of tension—fear of offending the other person, fear of expressing oneself improperly, and fear of being misunderstood. The other person' s response—nodding, questioning, or silence—will reflexively alter the student' s self-understanding. This is precisely the core of intersubjectivity: the self is constituted in and through the responses of others. Lantolf' s (2000) sociocultural theory further reveals how foreign-language learners achieve cognitive leaps within the "zone of proximal development" through interaction with real others. Levinas (1961/1969) points out that to face the face of the Other is to be summoned to respond, and that the responsibility of this response constitutes subjectivity. Generative artificial intelligence can simulate dialogue and generate grammatically correct responses, but it has no "self" that can be changed; therefore, it is difficult for it truly to "bear" the ethical weight of response.

(III) Generativity: The Double Expansion of the Boundaries of a Bilingual World Wittgenstein (1953/2001) points out that "the limits of language are the limits of the world" —the mother tongue provides the first symbolic system for partitioning the world, while a new language means acquiring a second power to reorganize experience. Norton' s (2013) research on language-learner identity also shows that foreign-language learners' "investment" in the target-language community is, in essence, a pursuit of an "imagined" or "actual" membership—this is precisely the empirical footnote to "becoming a different person through English."

This generativity is characterized by inexhaustibility. Through the resonance of a particular vowel in an English poem, a student may suddenly understand the shared emotional ground across cultures; through reading an English philosophical text, the student may reconstruct a judgment about a moral issue; through telling their own story in English, the student may discover that they can become someone "more direct, more humorous, and more willing to express emotion." Such experiences are difficult to predict, difficult to reduce to scores on an "intercultural competence scale," and even more difficult to generate algorithmically. It should be emphasized that the "inexhaustibility" here is not an epistemological statement of something "not yet exhausted," but an open

sequence in an ontological sense. What algorithms process is probabilistic calculation within a closed set; no matter how "personalized" their outputs may be, they are recombinations of variables along spatial dimensions, and every output is a historyless present. Human generative experience, by contrast, is temporal: when Student B, while reading a poem, suddenly understands her mother' s silence because of the resonance of a certain vowel, if this instant were required to occur again, its ontological thickness would already be different—because she is no longer the person who had not yet understood. It is precisely this irreversibility within temporality that makes generative experience difficult to rehearse, difficult to replicate, and difficult to reduce to any score on any scale.

The three dimensions above constitute the distinctive core of embodied foreign-language education. Bodily reconstruction, relational generation, and open-ended expansion are all indispensable; the systematic absence of any one dimension means that the curriculum structurally excludes the possibility of existential moments, and thus constitutes the core object of the falsification benchmark review in Chapter 3.

2.3 Key Clarification: The Boundary Between Embodied Education and Generative Artificial Intelligence

The opposite of embodied education is not "technology," but "disembodiment"—that is, treating language as an abstract exchange of information detachable from bodily perception and action. The traditional grammar-translation method narrows English into a system of symbolic rules, and students' brains are treated as "information processors"; this is typical of disembodied education. By contrast, role-playing and service learning regard English as "bodily action," which is the embryonic form of embodied education.

Generative artificial intelligence can serve as an auxiliary environment for embodied education. For example, it can simulate high-risk intercultural dialogue scenarios and provide learners with space for practice; it can play the role of dialogue partners from different cultural backgrounds, helping

learner become familiar with heterogeneous intersubjective interaction. For example, some universities in China have explored a teaching model of "AI rehearsal + real interpersonal interaction": generative AI is first used to simulate high-risk scenarios such as cross-cultural job interviews and international academic defenses, allowing students to complete embodied practice before entering real intersubjective interaction contexts. This both gives full play to the auxiliary value of technology and safeguards the core boundary of education as person formation.

However, generative artificial intelligence has a boundary that it can hardly transcend: it is difficult for it to replace real intersubjective encounters that carry ethical weight. When learners know that the other party in a dialogue is a machine, the risk of being "misunderstood" is lowered, the weight of "assum-

ing responsibility" is hollowed out, and the triggering conditions for existential moments are correspondingly weakened. Even if future artificial general intelligence (AGI) can simulate human conversation, it will still be unable to replace real intersubjective encounters—because "assuming responsibility for response" constitutes the core of subjectivity, whereas a machine has no "self" that can be transformed.

Therefore, the position of this article is not to reject technology, but to draw a bottom-line boundary for its application in English education: generative artificial intelligence may serve as a "practice field" leading toward embodied education, but it cannot become the "endpoint" of education. Genuine embodied education must ultimately return to encounters among real people and between heterogeneous cultures.

2.4 Experiential anchor: empirical support for embodied education

Ni Chuanbin' s (2023) meta-analysis of 56 studies on embodied second-language acquisition shows that factors such as embodiment, teaching methods, age of acquisition, foreign-language proficiency, and language distance can all influence the embodied effects of foreign-language learning: the stronger the embodiment of language components and teaching methods, the more evident the embodied effects, and the better the outcomes of foreign-language learning. The study by Jiang Meng and Zhao Sisi (2014) further revealed the embodied mechanism of affordance extraction in language comprehension. Macedonia' s (2019) meta-analysis also shows that learning foreign-language vocabulary through the execution of actions is more effective than traditional static memory, and can significantly narrow gender gaps in memory performance.

These studies provide evidence for the mechanism of embodied cognition. It should be noted, however, that the embodied education proposed in this article completes, on this basis, a theoretical leap from cognitive mechanism to an ontological claim, and cannot be reduced to a list of teaching strategies for "using bodily activity to aid memory." It is not a checklist of teaching strategies, but a fundamental revaluation of the nature of English education.

3. Paradigm shift: constructing falsifiability benchmarks for the evaluation of English education

3.1 Core defects of existing evaluation systems and the methodological shift

Existing evaluation systems center on standardized tests and attempt to "verify" the success of teaching, but their fundamental dilemma lies in the fact that they can measure only skills that can be replaced by algorithms (see 1.3). The reason education as person formation is difficult to verify through standardization is that its core is the generation of subjectivity—whether a student, through English, "becomes a different person" is a unique event in an individual life his-

tory, characterized by singularity, unpredictability, and non-replicability. Any attempt to convert it into a quantitative scale is a simplification of the essence of education as person formation (Biesta, 2010; Torrance, 2007).

In response to this dilemma, this article draws on Popper' s principle of "falsification" (Popper, 1959, 1963) and proposes a methodological shift: the essence of education as person formation is difficult to standardize and "verify," but ineffective teaching can be logically "falsified." We cannot force students to produce existential moments, nor can we measure their "magnitude," but we can identify those instructional designs that structurally and thoroughly preclude the possibility of existential moments. This shift will transform the logic of evaluation

The shift from "proving success" to "identifying failure" both preserves the critical edge and fundamentally prevents a slide toward new standardization.

3.2 Core Definition and Features of the Falsification-Based Benchmark

On the basis of the methodological turn described above, this paper proposes a falsification-based benchmark for the evaluation of English education:

If the overall design of an English course or program, in principle, systematically excludes the possibility of existential moments in which students complete a reconstruction of subjectivity through English, then that design should, in terms of its educative essence, be judged unqualified.

The core concept of this benchmark—existential moments—requires a dual delimitation. At the theoretical level, it refers to moments in which students, in the process of using English, experience self-loosening, reorganization, and enhancement. At the pedagogical level, as stated in the relational dimension in Section 2.2: when a student uses English for the first time to express a genuine dissenting opinion to a peer, accompanied by nervousness while speaking, relief after being listened to, and later the feeling that "because of this sentence in English, I became a braver person," this moment differs from "having produced a grammatically correct sentence." The difference lies in the fact that the former involves a transformation of subjectivity, whereas the latter involves only the operation of a skill.

The falsification-based benchmark has two core features. First, it is not used to "prove" the success of teaching. Existential moments cannot be compelled, predicted, or standardized; if they are set as rigid assessment indicators, their authenticity is easily damaged. Second, it is used to "falsify" the failure of teaching. By examining whether a course design structurally excludes the possibility of existential moments, unqualified instructional designs can be identified.

3.3 Hierarchical Differentiation and Boundaries of Application of the Falsification-Based Benchmark

The falsification-based benchmark applies at three levels, which must be distinguished to avoid misuse. The table below presents the scope of application, core logic of verification, transparency requirements, and rules of use for each level.

Table 1 Application Levels and Rules of Use of the Falsification-Based Benchmark

Applicable Level	Scope of Application	Core Logic of Verification	Transparency Requirements	Rules of Use
Course/Program Level	Complete degree programs, semester-long courses	Verify whether the course systematically excludes embodied reconstruction, intersubjective interaction, and generative spaces	A structural exclusion statement must be attached; the evaluated party is allowed to submit a counter-appeal; the full review checklist is made public	Overall judgment: if systematic exclusion is present, the course is unqualified
Teaching-Activity Level	A single class session, a single teaching activity	Verify whether the activity is a "closed design" and whether it forms an educative closed loop with other activities	The grounds for judgment must be traceable to specific design elements; a simplified exclusion statement is attached	Do not judge a single class session as unqualified; only identify risk-bearing activities

Applicable Level	Scope of Application	Core Logic of Verification	Transparency Requirements	Rules of Use
Baseline-Logic Level	All levels	Verify whether the design maintains the baseline of "leaving open the possibility for existential moments"	Judgment must refer to a publicly available anonymized case database; the benchmark text may be consulted and questioned	This is a baseline for qualification, not a standard of excellence

The above hierarchical distinctions are intended to prevent misuse of the benchmark. The benchmark applies only to course design itself; for specific prohibitions, see Section 3.5.

3.4 Testing Rules and Application Scenarios for the Falsifiability Benchmark

Based on the hierarchical distinctions discussed above, the falsifiability benchmark is tested according to the following logic: for a given course or project, one examines in sequence whether it leaves open the possibility of existential moments along the three dimensions of embodiedness, relationality, and generativity. If any one of these three dimensions is systematically excluded, the design should be judged unqualified.

The table below presents the testing results for seven typical teaching behaviors, in order to demonstrate the operability of the benchmark.

Table 2 Falsifiability Test Results for Typical Teaching Behaviors

Typical teaching behavior	Falsifiability test result	Core basis for judgment
Pure grammar-translation teaching	Falsified (unqualified)	It is excessively narrowed into the teaching of symbolic rules, lacks authentic intersubjective action scenarios, and systematically excludes the possibility of existential moments.
Standardized oral training that only corrects pronunciation	Falsified (unqualified)	Language is reduced to acoustic signals; there is only vocal output, with no embodied action or intersubjective interaction.
Simulated negotiation with no real consequences	Highly likely to be falsified	Risk is excessively reduced; students know that "this is only practice," and the triggering conditions for existential moments are systematically excluded.
Standardized speaking class containing a 5-minute authentic sharing segment (meeting the three requirements: genuine listening, risk without punishment, and genuine response)	A risk-bearing segment exists and improvement is needed; the course as a whole is not judged unqualified	Although the course as a whole is mainly standardized training, there is a structural opening, and the possibility of existential moments is not systematically excluded.
Project-based course with authentic interaction, but with the reflection segment quantified for grading	Relationality is qualified; generativity is at risk	Authentic intersubjective interaction leaves room for relationality, but the reflection segment is incorporated into the grading system, so the spontaneity of generativity may be weakened.

Typical teaching behavior	Falsifiability test result	Core basis for judgment
Cross-cultural role-play involving real risk	Not falsified (qualified)	It leaves complete possibilities for embodied reconstruction, intersubjective interaction, and generative space.
Authentic community translation service with a reflection segment	Not falsified (qualified)	It has authentic intersubjective consequences and, together with the reflection segment, provides space for existential generation.

The essence of falsifiability judgment is structural review, not item-by-item checking. To determine whether a course systematically excludes the possibility of existential moments, one should not look merely at whether it "has a certain activity," but rather at whether its overall design preserves structural possibilities for embodied reconstruction, intersubjective interaction, and generative openness.

When the three requirements are simultaneously satisfied, that segment constitutes a structural opening rather than a procedural display. Accordingly, the logic for judging borderline cases is as follows: if the overall course design does not systematically exclude the possibility of existential moments across the three dimensions, and only local closed segments exist, then it should be judged as "having risk-bearing segments and needing improvement," rather than as "unqualified." Audit feedback should point toward "how to enlarge this opening in the next round of teaching."

Horizontal projects can likewise be incorporated into this testing framework: if a project merely uses students as labor, takes completion of deliverables as its sole objective, does not include a reflection segment, and does not generate authentic intersubjective interaction, then its design in principle excludes the possibility of existential moments and should be falsified. Conversely, if the project sets tasks that require students to assume real responsibility for cross-cultural communication and is accompanied by reflective discussion segments, then it leaves space for existential transformation and is not falsified.

3.5 Transparency Principle: Making the Adjudication Procedure Visible

The critical force of the falsification benchmark should not return to educational practice in the form of yet another black box. The transparency principle requires that the adjudication procedure be fully exposed within a visible space where it can be examined, discussed, and challenged. Specifically, it includes four mutually supporting operational levels:

First, the disclosure of a structured review checklist. Evaluators must conduct their work on the basis of a public checklist grounded in the three dimensions. The checklist must be accompanied by explanatory notes to prevent its items from being alienated into rigid requirements.

Second, the mandatory writing of a statement of structural exclusion. When making a judgment of "unqualified,"evaluators must provide a written statement specifying on which dimension, because of which design features, what kind of existential possibility has been excluded. This statement is delivered together with the evaluation result and becomes an object open to public discussion.

Third, the gradual accumulation of an anonymized case database. Where conditions permit, anonymized cases that have undergone review should be made public as a common frame of reference, helping evaluators calibrate their judgment scales and reducing judgment drift caused by differences in personal experience.

Fourth, the establishment of a reverse appeal mechanism. The evaluated party has the right to submit a rebuttal within a specified period, identifying design elements that the evaluator failed to notice and that leave openings for existential moments. The evaluator must respond in writing, and if the rebuttal is valid, the conclusion must be revised.

Together, these four layers of design—public checklists, mandatory statements, case-based reference, and reversible appeals—ensure that, while the falsification benchmark holds the educational bottom line, it is itself also placed in a position where it can be criticized.

3.6 The Critical Force of the Benchmark and the Avoidance of Misuse

The falsification benchmark is a bottom-line supplement to existing evaluation systems, not a replacement for them. It does not deny the measurement value of quantitative evaluation for skills training; it is used only to hold the bottom line that "education must not be reduced to skills training that can be replaced by algorithms," thereby forming a complementary relationship with existing evaluation systems.

To prevent the benchmark from being misused as a "new standardization," three core prohibitions corresponding to its boundaries of applicability must be made explicit. First, its scope of application at the course/project level must not be exceeded by turning it into a performance indicator requiring teachers

to "produce existential moments" —compulsory sharing of transformative moments would damage the authenticity of transformation. Second, its scope of application at the level of bottom-line logic must not be exceeded by using it for student evaluation or horizontal comparison—we are judging whether the course design has excluded possibility, not whether a particular student has "successfully transformed." Third, its positioning as a "qualification bottom line" must not be exceeded by treating it as a basis for certifying an "excellent course" —it is intended to answer "whether it is qualified," not "whether it is excellent." Only by holding these three boundaries can the falsification benchmark function as a "line of defense," rather than becoming a new shackle.

Supporting carrier tools, such as critical incident analysis and sound walls, provide a space for witnessing traces of transformation in the process of educating people.

4. Application Demonstration and Value Extension of the Falsification-Based Benchmark

4.1 Application Demonstration of the Falsification-Based Benchmark: Taking Two College English Courses as Examples

It should be noted that the core contribution of this paper lies in theoretical construction and methodological innovation. The cases in this section are used only to demonstrate the application process and operability of the benchmark, rather than to provide a large-sample empirical test. The limitations of the study will be further discussed in terms of the means and boundaries of validating this benchmark.

To demonstrate the operability of the falsification-based benchmark, this section takes two typical college English courses as examples and presents the complete testing process. The descriptions of the two courses are both based on typical features of actual course syllabi, but typological treatment has been applied to highlight the contrast.

Case A: College English test-preparation course

This course is a College English Band 4 preparation course for first-year non-English majors. Its teaching content includes vocabulary memorization, grammar cloze exercises, reading-comprehension techniques, translation templates, and writing formulas. The typical form of teaching activity is: the teacher explains problem-solving techniques, students complete simulated exercises, and the teacher releases answers and provides analysis. Assessment consists of staged mock tests and a final Band 4 mock examination. The course syllabus does not involve authentic cross-cultural interaction, intersubjective risk-taking, or discussion of open-ended questions.

According to the testing logic in Table 2 of Section 3.4, this course systematically excludes the possibility of existential moments in all three dimensions

(pure symbolic training, no authentic interaction, and no open space). It is therefore falsified and may be regarded as essentially unqualified in terms of its educational formation function.

Case B: Intercultural Communication English Course

This course is a core professional course for second-year English majors. Students work in pairs, use English to investigate the needs of new immigrants, design a counseling plan, carry out the counseling in practice, and report and reflect on the experience to the whole class at the end of the semester. The course includes simulations of intercultural conflict scenarios: students play parties from different cultural backgrounds and conduct English debates in situations requiring moral judgment. Assessment includes a project report, reflective journal, and peer feedback; no standardized closed-book examination is used.

This course does not systematically exclude the possibility of existential moments in any of the three dimensions (authentic embodied action, genuine intersubjective risk, and open space retained in reflective journals). It is therefore not falsified and may be regarded as essentially qualified in terms of its educational formation function.

This demonstration shows that the falsification-based benchmark can quickly identify the educational bottom line of a course solely through analysis of the course syllabus and descriptions of teaching activities. It does not require classroom observation or the collection of student data; rather, it can complete the judgment at the logical level. This gives it relatively strong operability in curriculum review, teaching-quality evaluation, and planning for disciplinary development.

4.2 Inexhaustibility: The Core Boundary and Carrier Tools of Educational Formation

A potential risk of the falsification-based benchmark is that it may be misread as meaning that "as long as a course has not been falsified, it is a good course." It must be made clear that passing the falsification test is only a baseline condition for educational formation, not a sufficient condition. A course may leave open the possibility of existential moments, but whether students actually experience such existential moments depends on the teacher

···specific guidance, the creation of a classroom atmosphere, students' individual states of readiness, and many other factors. These factors are inexhaustible—they are difficult to design in advance, difficult to produce in a standardized manner, and difficult to predict algorithmically.

The inexhaustibility of experience is precisely one of the core boundaries that distinguishes human beings from algorithms. Algorithms process probabilistic calculations within closed sets; their outputs are reproducible and optimizable. What human education safeguards, however, is one-time, unpredictable "events"

that belong to particular subjects (see Section 2.2, Generative Dimension). Such moments are difficult to reduce to scores on any quantitative scale.

However, refusing algorithmic evaluation does not mean refusing all forms of attention. The key lies in distinguishing two entirely different logics: the logic of evaluation and the logic of bearing traces.

The logic of evaluation is judgment. The questions it answers are: "Is this qualified?" and "What level has this reached?" A pass/fail benchmark belongs to this logic, but it is a special exception—it judges only "unqualified," not "excellent." Quantitative testing also belongs to this logic; it answers questions such as "How large is the vocabulary?" and "How high is the grammatical accuracy?" The endpoint of the evaluative logic is judgment, and its products are grades, scores, or binary conclusions of qualified/unqualified.

The logic of bearing traces is witness. The questions it answers are: "What happened here?" and "What traces were left behind?" It does not seek judgment, does not produce scores, and is not used for comparison. Its products are not data, but traces—things that can be looked back on, listened to, and remembered, but that need not be measured.

The two cannot be reduced to one another, nor can they substitute for each other. If they are conflated—for example, if one attempts to use trace-bearing tools (such as reflection journals or sound walls) to "prove" educational outcomes, or incorporates them into teacher performance appraisal and course-evaluation indicators—then the vitality of these trace-bearing tools will be quietly damaged. Students may gradually learn to produce "conversion narratives" that meet expectations; teachers may tend to select those traces that "look profound" ; and the space for bearing traces risks being compressed into a more refined measurement tool. If the logic of bearing traces is confused with the logic of evaluation, its function of witnessing will be fundamentally impaired. Therefore, trace-bearing tools must operate independently of all scoring, ranking, and certification systems. Their ethical bottom line is this: any attempt to use trace-bearing evidence to "verify" teaching success departs from the original intention of the logic of bearing traces.

Based on this distinction, this paper proposes two supporting trace-bearing tools for witnessing the process of educating people, rather than judging educational outcomes.

(I) Key Event Analysis

At the end of the semester or the conclusion of a course, students are invited to look back and describe "a moment when I felt a change in myself in the process of using English" (if any). Note that this is not a required assignment. Students may choose "there was no such moment," which is itself a meaningful trace. Teachers read these narratives not in order to assign grades, but in order to understand: whether the course, in its structure, made such moments possible; which designs facilitated transformation; and which designs hindered

transformation. The stance of reading is listening, not grading. The disposition of the narratives is determined by the students—they may be read only by the teacher, may be provided anonymously for research use, or may be shared with the class after consent is obtained.

(II) Sound Wall

The voice wall is an open, voluntary, anonymous public holding space, which can be implemented with the aid of digital platforms. Its operating principles are as follows:

- **Voluntary participation:** Students decide for themselves whether to leave a trace; leaving no trace at all is also entirely normal;
- **Freedom of attribution:** They may choose to use their real names, pseudonyms, or complete anonymity;
- **Freedom of form:** A passage of text, a line of poetry, a word, or an emoji may all become traces;
- **Polyphonic response:** Others may add their own voices on top of existing ones—resonance, further questioning, and silent attention are all forms of response;
- **The teacher' s role:** The teacher is not a grader, but a listener and a weaver. At the end of the semester, the teacher may select a group of voices and present them to the whole class without revealing identities: "These are some of the echoes that appeared in our classroom this semester."

The voice wall is not an assessment tool. It does not attempt to "capture" moments of existential presence; it merely provides a dwelling place for those moments that are willing to be shared. Whether anyone leaves a voice, what kind of voice is left, and how many voices there are—none of these constitutes an "evaluation" of the course. They are only traces of education, preserved honestly and safely, waiting to be revisited or forgotten.

The reason for deliberately introducing a carrying tool is that, logically, the falsification benchmark requires only the course syllabus and descriptions of teaching activities in order to make a judgment; it does not need to enter students' inner worlds. Yet educators cannot merely hold the bottom line; they also need to provide a visible plane for those things that cannot be judged. The falsification benchmark is intended to prevent us from doing harm—it safeguards the lower limit beneath which education must not fall; the carrying tool, by contrast, allows those accidental and unpredictable good things not to vanish silently within the blind spots of the assessment system—it opens up the possibilities of education that cannot be predetermined.

4.3 Ultimate Value: The Educational Ethics of Resisting Algorithmic Governmentality

The argument of this paper ultimately points to a deeper proposition: the inexhaustibility guarded by embodied education is, in essence, a resistance to the domination of the human by "computability" under algorithmic governmentality. Williamson points out that algorithmic governmentality is reshaping the underlying logic of education—breaking the learning process down into predictable and optimizable data flows, replacing judgment with metrics, and replacing teachers with algorithms (Williamson, 2021). Under this trend, education faces the risk of degenerating into a "workshop for training human skills," with its primary reason for existence becoming the supply of disciplinable labor for algorithmic society.

The core legitimacy of English education lies precisely in its safeguarding of what is difficult to compute. When a language is not taught as a tool, but experienced as a mode of human existence, it opens up a domain that algorithms can scarcely reach: a process in which a learner uses a heterogeneous language to redivide the world, reinterpret the self, and reestablish relations with others. This process is difficult to accelerate, difficult to standardize, and difficult to outsource to any machine.

Therefore, the core claim of this paper is not to return to some romantic imagination of a pre-technological age, but, at a time when algorithms are increasingly invading the core of education, to draw an inviolable bottom line: any instructional design that in principle eliminates the possibility for students to become different people through language, no matter how efficient its outputs may be, can hardly be regarded as genuine education.

This bottom line is not only the starting point for English education' s self-defense in the age of artificial intelligence, but also one of the core grounds for the existence of the humanities in the intelligent age.

5. Conclusions and Implications

5.1 Core Research Conclusions

The study in this paper can be summarized in two core conclusions.

First, the existential anxiety surrounding English education in the age of artificial intelligence stems, to a large extent, from instrumentalism' s forgetting of the foundation of educating persons. This paper attempts to take embodied person-formation as the ontological foundation of English education, anchoring it in three exclusive dimensions: see 2.2 for details. In doing so, it responds to the question of the irreplaceable core value of English education.

Second, in response to the difficulty that the essence of educating persons is hard to assess through standardized evaluation, this paper attempts to construct falsification-oriented evaluation criteria. By using falsification to exclude

instructional designs that foreclose the possibility of educating persons, it safeguards the educational bottom line: see 3.2 for details. Supporting carrier tools provide an evidentiary space for traces of transformation: see 4.2.

5.2 Practical Implications

Based on the above conclusions, this paper proposes three practical implications.

First, English curriculum design should return to the ontological foundation of embodied person-formation. The review of course syllabi should not focus only on "which knowledge points are covered" or "which skills training is arranged"; it should also ask: Does this course leave room for students to undergo embodied reconstruction, intersubjective interaction, and generative spaces in English? Does it create real and risky scenarios of intersubjective encounter? In university English curriculum reform, it is possible to try reducing the proportion of purely test-oriented training and increasing concrete person-forming components based on real projects, such as service learning, intercultural debate, and community translation. The design of horizontal projects should go beyond the logic of "revenue generation" and regard students as subjects of education rather than as human capital for completing tasks.

Second, educational evaluation reform should be deepened by breaking through the quantitative framework of instrumentalism. In institutional evaluation, program accreditation, and course review, a falsification-oriented bottom-line review mechanism may be introduced to identify instructional designs that systematically exclude the possibility of educating persons. At the same time, carrier tools such as critical incident analysis and voice walls should be incorporated as independent evidentiary spaces parallel to quantitative evaluation systems, rather than as supplements to the evaluation toolbox.

Third, boundaries should be drawn for the application of generative artificial intelligence in English education. The boundary of technological application has already been set out in 2.3. Curriculum design should remain alert to the solutionist tendency to use AI to replace real intersubjective encounters, and should avoid reducing education to an information-processing process of human–machine interaction.

5.3 Research Limitations and Future Prospects

The core contribution of this paper lies in theoretical construction and methodological innovation, but it also has clear limitations. First, the transparency procedures for the falsification-oriented criteria and the construction of a case database still need to be gradually accumulated and calibrated in real teaching-evaluation contexts. Through a typological comparison of two courses, this paper has demonstrated the application logic of the criteria; however, the specific items of the review checklist, the writing norms for exclusion statements, and the classification framework of the case database still need to be tested and revised in pilot applications across multiple universities and different educational stages.

As a logical review tool, the "verification" of falsification-oriented criteria is not a matter of large-sample statistical testing, but of procedural transparency and case accumulation; its reliability depends on what the community forms through long-term practice

a consensus of judgment, rather than experimental data. Second, the principles of spatial design and the ethics of implementation for carrier tools (such as critical-incident analysis and voice walls) remain to be further explored. How can a safe, voluntary carrier space be designed—one not penetrated by the logic of assessment? When teachers listen to students'transformative narratives, what ethical boundaries should they observe (such as privacy protection and not using narratives for evaluation)? The core of these questions is not "how to analyze narrative materials," but rather "how to safeguard a space not colonized by assessment logic." In this respect, current research offers only preliminary ideas and still requires more systematic theoretical discussion and practical validation. Third, the modes in which foreign languages enact their threefold educational dimension may differ across educational stages—basic education, higher education, and adult education—and this article has not undertaken a detailed discussion of such differences.

Future research may be expanded in the following directions. First, based on a negative benchmark, tools for curriculum design and curriculum-review checklists oriented toward embodied education can be developed and piloted in foreign-language departments across multiple universities. In this process, anonymized case libraries can be accumulated in parallel to test the operability of the transparency procedure and the consistency of judgments. Second, the design principles and ethical frameworks of carrier tools (such as voice walls and learning-legacy platforms) should be explored, examining how, without coding or scoring narrative materials, traces of transformation can be safely witnessed and shared. The key in this direction is not to develop new analytical techniques, but to create an institutional place of repose for those things that cannot be evaluated. Third, the implementation conditions and constraints of foreign languages as embodied education at different educational stages should be compared, so as to provide evidence for graded curriculum design. Fourth, the dialogue between the theory of embodied education and theories of second-language acquisition should be further deepened, exploring the theoretical interfaces between the three dimensions of embodiment, relationality, and generativity and existing core concepts in second-language acquisition, such as identity, investment, and imagined communities, thereby providing theoretical resources for constructing an autonomous knowledge system for foreign-language education in China.

References

[1] Cai Jigang. The application of artificial intelligence in universities: opportunities for improving critical-thinking ability [J]. *Foreign Language Teaching Theory and Practice*, 2025(1): 61–66.

[2] Jiang Meng, Zhao Sisi. A study on the extraction of affordances in the process of discourse comprehension [J]. *Foreign Language Teaching and Research*, 2014, 46(4): 584–595.

[3] Ministry of Education and four other departments. Notice on issuing the "Artificial Intelligence + Education" Action Plan [EB/OL]. (2026-04-02) [2026-06-02]. http://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html.

[4] Ni Chuobin. The embodied effect of foreign languages and its pathways and mechanisms [J]. *Foreign Language Teaching and Research*, 2023, 55(6): 876–886.

[5] Wang Yin. A new perspective on English teaching: the theory and practice of embodied linguistics [J]. *Shandong Foreign Language Teaching*, 2022, 43(3): 9–21.

[6] Wang Yin. The theoretical value and historical significance of embodied linguistics [J]. *Journal of Tianjin Foreign Studies University*, 2023(2): 1–9.

[7] Wen Qiufang. Will foreign-language education in the era of artificial intelligence produce a subversive revolution? [J]. *Modern Foreign Languages*, 2024, 47(5): 597–607.

[8] Wen Qiufang. Frameworks for topic selection and methodological suggestions for research on foreign-language education in the AI era [J]. *Foreign Language World*, 2025a(2): 2–10.

[9] Wen Qiufang. AI-empowered POA reform in foreign-language teaching: an innovative exploration oriented toward real-time interaction [J]. *Foreign Language Teaching Theory and Practice*, 2025b(4): 40–50.

[10] CPC Central Committee, State Council. Overall plan for deepening the reform of educational evaluation in the new era [EB/OL]. (2020-10-13) [2026-06-02]. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202010/t20201013_494381.html.

[11] Biesta G. *Good education in an age of measurement: ethics, politics, democracy* [M]. Boulder: Paradigm Publishers, 2010.

[12] Black P, William D. Inside the black box: raising standards through classroom assessment[J]. *Phi Delta Kappan*, 1998, 80(2): 139-148.

[13] Broadfoot P, Black P. Redefining assessment? the first ten years of assessment in education[J]. *Assessment in Education: Principles, Policy & Practice*, 2004, 11(1): 7-26.

[14] Gallese V, Lakoff G. The brain' s concepts: the role of the sensory-motor system in conceptual knowledge[J]. *Cognitive Neuropsychology*, 2005, 22(3-4): 455-479.

[15] Heidegger M. *Being and time*[M]. Macquarrie J, Robinson E, trans. New York: Harper & Row, 1962.

[16] Lantolf J P. *Sociocultural theory and second language learning*[M]. Oxford: Oxford University Press, 2000.

[17] Levinas E. *Totality and infinity*[M]. Lingis A, trans. Pittsburgh: Duquesne University Press, 1969.

[18] Macedonia M. Embodied learning: why we should take it seriously for language teaching[J]. *Frontiers in Psychology*, 2019, 10: 1-7.

[19] Merleau-Ponty M. *Phenomenology of perception*[M]. Landes D A, trans. London: Routledge, 2012.

[20] Norton B. *Identity and language learning: extending the conversation*[M]. 2nd ed. Bristol: Multilingual Matters, 2013.

[21] Popper K R. *The logic of scientific discovery*[M]. New York: Basic Books, 1959.

[22] Popper K R. *Conjectures and refutations: the growth of scientific knowledge*[M]. London: Routledge & Kegan Paul, 1963.

[23] Scriven M. The methodology of evaluation[M]//Tyler R W, Gagné R M, Scriven M. *Perspectives of curriculum evaluation*. Chicago: Rand McNally, 1967: 39-83.

[24] Torrance H, Pryor J. *Investigating formative assessment: teaching, learning and assessment in the classroom*[M]. Buckingham: Open University Press, 1998.

[25] Torrance H. Assessment as learning? how the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning[J]. *Assessment in Education: Principles, Policy & Practice*, 2007, 14(3): 281-294.

[26] Warschauer M, Xu Y. Generative AI for language learning: entering a new era[J]. *Language Learning & Technology*, 2024, 28(2): 1-4.

[27] Williamson B. *The digital governance of education: data, algorithms, and the rise of platform schools*[M]. London: Bloomsbury Academic, 2021.

[28] Wittgenstein L. *Philosophical investigations*[M]. Anscombe G E M, trans. Oxford: Blackwell, 2001.

Corresponding author: Shi Wei (E-mail: 20031@jdzu.edu.cn)

Author Contributions Statement

Shi Wei: proposed the research ideas and designed the research plan; was responsible for the literature review, construction of the theoretical framework, and methodological innovation; completed the drafting, revision, and final-version revision of the paper.

Note: Figure translations are in progress. See original paper for figures.