Can AI Be Trusted to Plan Lessons? Auditing the Alignment of Generative-AI Primary English Lesson Plans with China's Curriculum Core Competencies

Ziheng Zhang *

Northern Arizona Institute, Shanghai Normal University Tianhua College, Shanghai 201815, China

* Corresponding Author Email: zz372@nau.edu

Abstract. Generative artificial intelligence (GenAI) is entering primary teachers' lesson-planning routines, yet to our knowledge no study has audited its output against the competency framework of China's English Curriculum Standards for Compulsory Education (2022 Edition). Three chatbot models (Claude, DeepSeek, Gemini) each generated lesson plans for four official primary teaching cases from Appendix V of the standards, under a naive and a curriculum-anchored prompt, twice per condition (48 plans; 1,767 activities). Two independent automated coders of different model lineages applied an instrument derived verbatim from the standards (activity-tier kappa = 0.77), and each plan was compared with its official exemplar using Porter's alignment index. Contrary to earlier audits, the plans were not tier-deficient: transferring-and-creating activities averaged 22.9% of each plan's coded activities, matching the official cases (23.3%), though naive prompts fell modestly short (17.2%). Alignment gaps concentrated in cultural awareness: under naive prompts only 25% of plans operationalised Chinese-foreign cultural comparison, rising to 96% under curriculum-anchored prompts (cultural-awareness score, $p < .001$). Document-level competency scores otherwise saturated, and repeated generations varied noticeably. Findings support prompt-anchoring plus teacher review rather than wholesale trust or rejection.

Keywords: Generative artificial intelligence, lesson planning, curriculum alignment, core competencies, primary English education, content analysis.

1. Introduction

China's *English Curriculum Standards for Compulsory Education (2022 Edition)* reorganised the goals of school English around four core competencies – language ability, cultural awareness, thinking capacity, and learning ability – and prescribed the English learning activities approach, in which classroom activities progress from learning and understanding, through applying and practising, to transferring and creating [1]. Implementing this reform falls first on the lesson plan – the document where a teacher translates curriculum requirements into objectives, activities, and assessment – and teachers' preparedness for that translation has been a persistent concern of curriculum reform in China [2]. At the same time, generative artificial intelligence (GenAI) chatbots have rapidly entered teachers' planning routines: reviews and studies of ChatGPT in language education and teacher education identify lesson planning and material generation among the most common teacher-facing applications, while noting that empirical evidence on the teacher side remains far thinner than on the learner side [3, 4].

Whether such AI-drafted plans can be trusted is beginning to be audited empirically, but almost exclusively against Western yardsticks. The largest audit to date coded 310 plans generated for Massachusetts civics standards and found that only 10% of 2,230 activities promoted higher-order thinking [5]. In foreign-language education, zero-shot ChatGPT plans were of generally high quality, yet showed extreme output variability and a bias toward practices long refuted by research [6]. Related audits against the Universal Design for Learning (UDL) framework [7], general content analyses [8, 9], perception and classroom studies [10, 11, 12], and culturally responsive design tools [13] likewise employ Western standards or generic pedagogical frameworks. No study we could locate evaluates AI-generated lesson plans against the competency framework of China's 2022 curriculum standards – the yardstick governing the classrooms concerned; Chinese-language

scholarship, for its part, has so far focused on how GenAI can empower lesson design rather than on auditing its output against curriculum standards [14]. Nor do existing audits use the alignment-measurement tradition developed for curriculum research [15, 16], and few report inter-coder reliability statistics.

This study therefore audits GenAI lesson planning against the 2022 standards themselves. Three widely accessible chatbot models each generated plans for four primary-level teaching cases published in Appendix V of the standards, under a naive prompt and a curriculum-anchored prompt, twice per condition (48 plans in total). Every plan was coded, activity by activity, with an instrument derived verbatim from the standards, and compared with the official exemplar for the same unit using Porter's alignment index [15]. Three questions are asked. RQ1: How well do AI-generated primary English lesson plans align with the four core competencies and the three tiers of the activities approach? RQ2: How do model choice and prompt condition affect alignment? RQ3: Where do AI plans fall short relative to the official exemplars? The study contributes the first criterion-referenced audit of GenAI lesson plans under China's competency framework, a reusable standards-derived coding instrument with dual-coder reliability evidence, and practical guidance on what curriculum-anchored prompting does – and does not – repair.

2. Background

2.1. The 2022 curriculum standards as an audit criterion

The 2022 standards define the four core competencies as the "educational value focal point" of the English curriculum and specify level goals for each: Level 1 for Grades 3–4 and Level 2 for Grades 5–6 in the primary phase [1]. For classroom implementation, the standards prescribe the English learning activities approach, requiring activity sequences that move from learning and understanding (perceiving and noticing; retrieving and sorting out information; generalising and consolidating), through applying and practising (describing and interpreting; analysing and judging; internalising and applying), to transferring and creating (reasoning and arguing; critiquing and evaluating; imagining and creating) [1]. Two features make the standards unusually well suited as an audit criterion. First, the goal tables decompose each competency into three subdomains with observable, level-specific descriptors, so a coding scheme can be derived from the document rather than invented by researchers – an approach previously applied to textbook analysis under the senior-secondary standards [17]. Second, Appendix V of the standards publishes complete official teaching cases for Grades 3–6, providing government-issued exemplars against which generated plans can be benchmarked on identical topics.

2.2. Auditing AI-generated lesson plans

Existing audits converge on three findings. First, AI plans are structurally fluent but cognitively shallow: 90% of activities in the Massachusetts civics corpus sat at the remember, understand, or apply levels, and prompting for "highly interactive" plans produced only minor change, with activities at the create level actually declining [5]. Second, outputs are unstable: identical prompts produced extreme variability across generations, and added prompt specificity did not guarantee higher quality [6]. Third, plans inherit historical biases: models trained on classic pedagogical texts reproduced practices such as scripted-dialogue recitation that communicative language teaching has long abandoned [6], fabricate resources such as non-existent books [9], and their weaknesses are hard to detect: experienced teachers could not reliably tell AI and human plans apart [10]. Whether these findings transfer to a competency-based curriculum in a non-Western system remains untested: framework-referenced audits exist for UDL [7] and culturally responsive teaching [13], but not for a national competency framework, and the sole classroom experiment found no significant between-group advantage for ChatGPT-planned lessons [11].

2.3. Measuring alignment

Curriculum research offers a mature toolkit for quantifying agreement between documents and standards: Porter's alignment index computes 1 minus half the summed absolute discrepancies between two content-proportion matrices, a 0–1 statistic that Porter cautions has no absolute threshold and is best used comparatively [15], while Webb's criteria add categorical concurrence and depth-of-knowledge consistency [16]. Comparative evaluation of alignment models further shows that whatever scheme is chosen, its inter-rater reliability must be demonstrated [18]. This tradition has not previously been connected to the auditing of AI-generated teaching materials; the present study makes that connection by using the standards' own activity tiers as the content dimension of a Porter-style comparison.

3. Method

3.1. Design

The study is a 3 (model) × 2 (prompt condition) × 4 (teaching case) design with two independent generations per cell, yielding 48 AI-generated plans, each benchmarked against the official Appendix V case for the same unit (N = 52 documents). The audit is criterion-referenced: every coding category is taken verbatim from the 2022 standards.

3.2. Materials: units and official anchors

Four official teaching cases were drawn from Appendix V of the standards, spanning both primary levels and three plan formats: U1 "Meet My Family" (Grades 3–4, four-lesson unit), U2 "Do Family Chores" (Grades 5–6, four-lesson unit), U3 "Do You Do Chores at Home?" (Grades 5–6, single dialogue lesson), and U4 "Lantern Festival Riddle Games" (Grades 5–6, three-lesson integrated practice activity). Using the standards' own cases fixes topic, grade band, and scope.

3.3. Plan generation

Three chatbot models of different vendors were used: Claude (claude-opus-5, Anthropic), DeepSeek (deepseek-v4-pro, DeepSeek), and Gemini (gemini-3.7-flash at high reasoning effort, Google). All plans were generated on 1 September 2026 through the vendors' command-line interfaces in single-turn mode with default sampling settings (the interfaces do not expose temperature control); model identity for every run was verified from interface-written log artifacts rather than model self-report. Each cell was generated twice because identical prompts can yield highly variable outputs [6]. Two prompt conditions were used. The naive condition (P1) mirrors the broad requests teachers typically write [5]: "Write a teaching plan for a primary school English class in China on the topic '[topic]' ([hint]) for [grade] students. The plan should cover [scope]. Write the plan in English." The curriculum-anchored condition (P2) added the four core competencies by name, the level goals, the three tiers of the activities approach with all nine sub-activity descriptors, Chinese-and-foreign cultural integration, and learning-strategy and assessment requirements. A constant delivery instruction (save the complete plan to a specified file) was appended in all conditions. The 48 plans ranged from 2,020 to 9,586 words (median 5,512.5).

3.4. Coding instrument

The instrument operationalises two axes, both derived verbatim from the standards. Axis 1 (document level) records, for each competency, whether each of its three subdomains – e.g., for thinking capacity: observing and differentiating, summarising and inferring, critiquing and creating – is substantively operationalised by at least one activity at the unit's level, and whether the plan's objectives merely mention the competency. A mechanical rule converts these indicators to a 0–3 score: 0 = absent; 1 = mentioned in objectives but operationalised by no activity (surface alignment); 2 = one subdomain operationalised; 3 = two or more. Axis 2 (activity level) assigns every student-facing

activity exactly one tier of the activities approach (learning-understanding, applying-practising, transferring-creating, or outside the tiers), following an iron rule adopted from prior audit work [5]: activities are coded by what students are actually required to do, never by the plan's own labels, and multi-task blocks receive the highest tier actually reached. Boundary rulings (e.g., role-playing a given dialogue is applying-practising; creating an original dialogue is transferring-creating) were fixed in the codebook before coding. Table 1 summarises the instrument.

Table 1. The coding instrument (all categories quoted from the 2022 standards).

Axis	Categories	Source in the standards
Activity tier (per activity)	learning-understanding (LU); applying-practising (AP); transferring-creating (TC); outside tiers (NA)	Teaching suggestions, nine sub-activity descriptors
Competency subdomains (per document)	3 subdomains × 4 competencies (e.g., CA1 comparing-judging; CA2 adjusting-communicating; CA3 appreciating-internalising)	Level-goal tables, Levels 1–2
Score rule (mechanical)	0 absent; 1 mentioned, never operationalised; 2 one subdomain; 3 two or more	Derived; not coder-assigned

3.5. Coding procedure and reliability

Two automated coders of different model lineages applied the identical instrument independently: Coder A (Claude) segmented each document into activities and coded both axes; Coder B (Gemini) received Coder A's segmentation with all codes removed and recoded every activity and indicator without access to Coder A's judgments. Reliability of this dual-model coding is reported rather than assumed: Cohen's kappa is computed over all activity tier assignments and all subdomain indicators, disagreements are logged for human adjudication, and primary analyses use Coder A's codes with a sensitivity check restricted to activities on which both coders agreed. Because Gemini both generated one third of the corpus and served as Coder B, a self-preference check compares the two coders' scoring of Gemini-generated versus other plans.

3.6. Alignment quantification and analysis

For each plan, the distribution of coded activities across the three tiers (excluding outside-tier activities) forms a three-cell proportion vector X; the same vector for the official case of the same unit forms Y; Porter's alignment index $1 - (\Sigma|X - Y|)/2$ quantifies distributional alignment [15], used comparatively as Porter advises. Group comparisons across models use Kruskal-Wallis tests and across prompt conditions Mann-Whitney U tests on the transferring-creating share, the alignment index, and the four competency scores; generation stability is described by absolute differences between repeated generations within cells.

4. Results

4.1. Reliability

The two coders assigned identical tiers to 84.5% of the 1,825 activities in the full document pool (Cohen's kappa = 0.77), and agreed on 95.7% of the 624 document-level subdomain judgments (pooled kappa = 0.75); Table 2 reports each measure.

Table 2. Inter-coder reliability (Coder A: Claude; Coder B: Gemini).

Measure	kappa	% agreement	N
Activity tier (activity level)	0.7709	84.5%	1825
CA1 comparing-judging	0.8412	92.3%	52
CA2 adjusting-communicating	0.6237	80.8%	52
CA3 appreciating-internalising	0.7797	96.2%	52
LN1 passion for learning	1.0	100.0%	52
LN2 selecting-adjusting	0.3716	84.6%	52
All 12 indicators (pooled)	0.754	95.7%	624

Note: the seven indicators omitted above reached 98.1–100% raw agreement but degenerate (0.0) or undefined kappa owing to near-uniform positive prevalence; percent agreement is the informative statistic there.

Disagreements concentrated on the applying-practising boundary (108 on the LU-AP border, 115 on the AP-TC border, of 283 total) and were retained for adjudication; the prompt effect on the transferring-creating share survives a sensitivity re-analysis restricted to activities on which both coders agreed ($p = .0009$). Seven subdomain indicators showed the high-agreement, low-kappa prevalence paradox [19] noted under Table 2. Neither coder favoured plans generated by its own model family: the between-coder difference in transferring-creating share did not differ between Gemini-generated and other plans ($p = .61$) or between Claude-generated and other plans ($p = .84$).

4.2. Tier distribution and competency coverage (RQ1)

The 48 AI plans contained 1,767 activities (M = 36.8 per plan, range 15–65), about two and a half times the granularity of the official cases (11–18 activities each). Contrary to expectation, the plans were not tier-deficient: averaging per-plan tier shares across all conditions, learning-and-understanding accounted for 38.6%, applying-and-practising 38.5%, and transferring-and-creating 22.9% – the latter essentially identical to the official cases' 23.3% (Fig. 1).

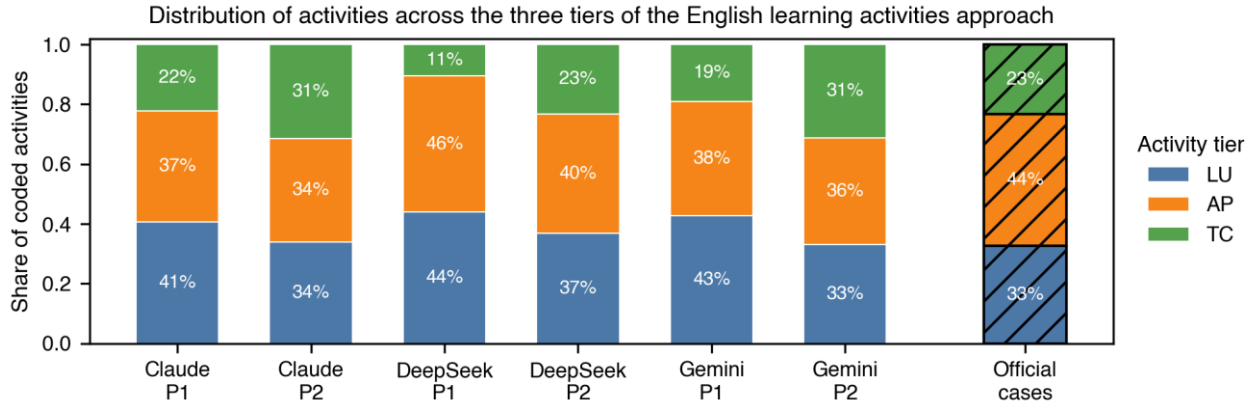


Fig. 1 Distribution of activities across the three tiers of the English learning activities approach, by model and prompt condition; hatched bar shows the official cases.

Document-level competency scores showed a pronounced ceiling: every one of the 48 plans reached the maximum score of 3 for language ability, thinking capacity, and learning ability, as did the official cases for language ability and thinking capacity (Table 3).

Table 3. Mean tier shares, Porter alignment index, and competency scores by model and prompt condition (official cases for reference).

Group	LU	AP	TC	Porter	Lang.	Cult.	Think.	Learn.
Claude P1	40.7%	37.1%	22.2%	0.818	3.00	2.50	3.00	3.00
Claude P2	34.0%	34.5%	31.5%	0.814	3.00	3.00	3.00	3.00
DeepSeek P1	44.0%	45.5%	10.5%	0.804	3.00	1.88	3.00	3.00
DeepSeek P2	37.0%	39.8%	23.2%	0.875	3.00	2.88	3.00	3.00
Gemini P1	42.8%	38.1%	19.0%	0.851	3.00	2.38	3.00	3.00
Gemini P2	33.2%	35.7%	31.2%	0.880	3.00	3.00	3.00	3.00
All AI plans	38.6%	38.5%	22.9%	0.840	3.00	2.60	3.00	3.00
Official cases	32.8%	43.9%	23.3%	–	3.00	1.50	3.00	2.75

Only cultural awareness discriminated. There, the gap under naive prompting was real: mean scores were 2.25 (P1) versus 2.96 (P2), and at subdomain level only 25% of P1 plans operationalised comparing-and-judging Chinese and foreign cultures (CA1) and 29% adjusting-and-communicating across cultures (CA2), against 96% and 67% under P2 (Fig. 2).

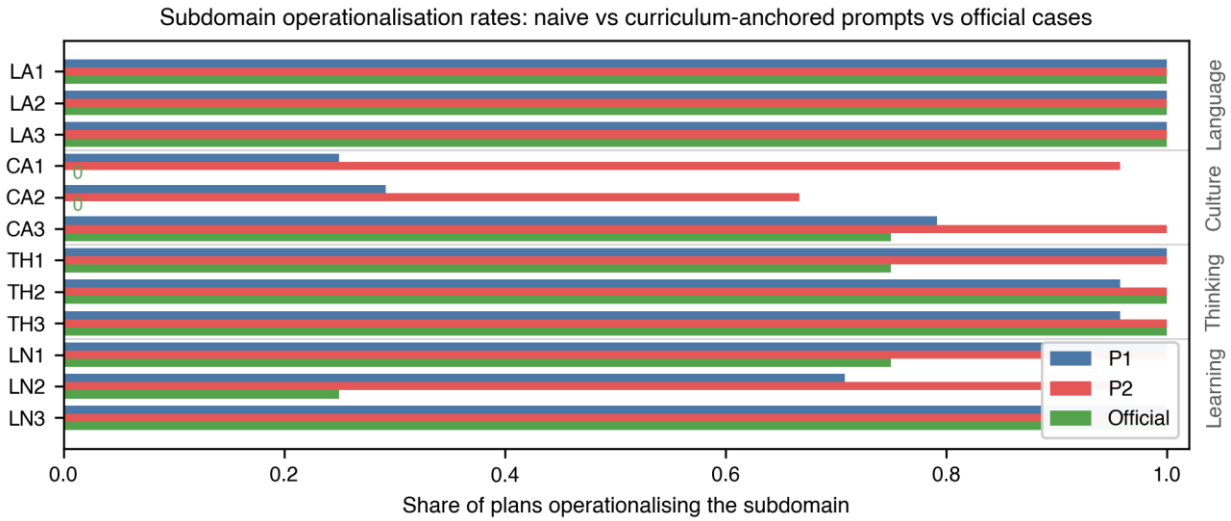


Fig. 2 Subdomain operationalisation rates under naive (P1) and curriculum-anchored (P2) prompts, with the official cases as reference; zeros are marked explicitly.

Notably, the official cases themselves operationalised neither subdomain in these four units (rate 0%), and their mean cultural-awareness score (1.5) was below even the naive-prompt plans.

4.3. Prompt and model effects (RQ2)

Curriculum-anchored prompting significantly raised the transferring-and-creating share (median 0.174 to 0.266; Mann-Whitney U = 136.0, p = .002) and the cultural-awareness score (median 2.0 to 3.0, p < .001), but not the Porter index (0.820 to 0.868, p = .112) (Fig. 3).

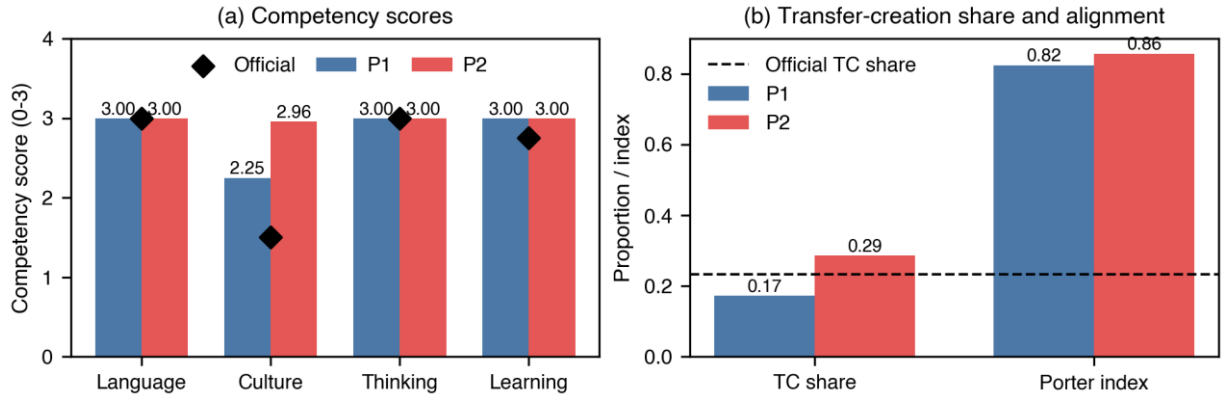


Fig. 3 Effects of curriculum-anchored prompting: (a) mean competency scores with official-case reference points; (b) mean transferring-creating share and Porter alignment index.

Differences among models were marginal (transferring-creating share: Kruskal-Wallis H = 5.34, p = .069), with DeepSeek producing the smallest transferring-creating share (mean 0.169, versus 0.268 for Claude and 0.251 for Gemini) and the lowest cultural-awareness mean (2.38; model differences in cultural-awareness scores were not significant, p = .22).

4.4. Distance from the official exemplars and generation stability (RQ3)

Porter alignment indices against the same-unit official case averaged 0.840 (range 0.705–0.965), indicating high but imperfect distributional alignment. The direction of misalignment reversed with the prompt: 17 of 24 naive-prompt plans fell below their unit's official transferring-creating share, whereas 15 of 24 curriculum-anchored plans exceeded it – anchored prompting overshoots the official distribution rather than converging on it. Repeated generation under identical prompts produced visible instability: the median within-cell change in transferring-creating share was 0.057 (maximum 0.200), and activity counts differed by a median of 5 activities (maximum 23) between two runs of the same prompt.

5. Discussion

Three findings reframe the trust question. First, the tier-collapse reported for the 2024 model generation did not replicate: where only 10% of activities in the Massachusetts civics corpus promoted higher-order thinking [5], in this corpus the transferring-and-creating share matched the official exemplars under curriculum-anchored prompts, and even under naive prompts (17.2%) it stayed well above the collapse level of the earlier generation – though not uniformly, as the DeepSeek naive cell (10.5%) came close to replicating it; coding followed the same conservative rule of ignoring the plans' self-labels. The two audits use different taxonomies and subjects, so the comparison is directional rather than exact, but it suggests that the cognitive shallowness of AI lesson plans may be a property of particular model generations rather than of the technology itself – and that audit results must be dated accordingly.

Second, the alignment risk that remains is cultural, and it is prompt-addressable. Naive prompts yielded plans that largely ignored Chinese-foreign cultural comparison and cross-cultural communication – the subdomains through which the standards carry their distinctive value orientation– while a single curriculum-anchored prompt nearly closed the gap (though cross-cultural communication still reached only two thirds of plans). Yet the same anchoring pushed plans past the official distribution (and past the official cases' own modest cultural coverage), consistent with instruction-following models treating the prompt as a checklist to be satisfied everywhere. Whether a family-chores unit genuinely benefits from inserted cultural comparison is a judgment the teacher, not the model, must make; over-compliance is the new face of surface alignment, replacing the label-without-substance pattern the scoring rule was designed to catch.

Third, the audit instrument matters as much as the audit verdict: document-level presence scores saturated for three of four competencies, and the discriminating signal lived at activity and subdomain level. Combined with the observed generation instability – the same prompt shifting a plan's transferring-creating share by up to 20 percentage points, echoing earlier reports of extreme variability [6] – this cautions against one-shot, document-level evaluations of AI planning tools, by researchers or by school procurement processes.

Teachers should anchor prompts in the standards' own terms, curate cultural content deliberately, and regenerate at least once before selecting a plan. Four limitations bound these claims. The corpus covers four units whose official cases are culture-light, so the cultural findings describe these topics, not the curriculum; the coders were large language models, with reliability demonstrated statistically but human adjudication of the 283 flagged disagreements still pending; models were accessed through command-line interfaces whose vendor harness prompts are absent from consumer chat apps; and the novelty claim rests on title-level screening of English databases plus CNKI (China National Knowledge Infrastructure).

6. Conclusion

This study provided a criterion-referenced audit of generative-AI lesson planning against the core competencies of China's 2022 compulsory-education English curriculum, benchmarking 48 plans from three models against the standards' own official teaching cases with a dual-coder, standards-derived instrument. The current model generation produced structurally competent plans; the substantive gap lay in cultural awareness under naive prompting, and curriculum-anchored prompting overshot rather than matched the official distribution. AI planning support in this context is therefore neither trustworthy by default nor dismissible: it is a competent drafting instrument whose curriculum fit depends on how it is prompted and audited, with the teacher retaining responsibility for cultural substance and final selection.

References

[1] Ministry of Education of the People's Republic of China. Compulsory Education English Curriculum Standards (2022 Edition). Beijing: Beijing Normal University Publishing Group, 2022. (In Chinese; official English translation consulted.)

[2] Zhou A, He H. English curriculum reform in China and teachers' preparedness to implement the latest curriculum standards. In: Oxford Research Encyclopedia of Education. Oxford: Oxford University Press, 2025.

[3] Lo C K, Yu P L H, Xu S, Ng D T K, Jong M S Y. Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. Smart Learning Environments, 2024, 11: 50.

[4] van den Berg G, du Plessis E. ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. Education Sciences, 2023, 13(10): 998.

[5] Trust T, Maloy R, Xu C, Pelletier K. Civic education in the age of AI: Should we trust AI-generated lesson plans? Contemporary Issues in Technology and Teacher Education, 2025, 25(3): 418–442.

[6] Dornburg A, Davin K J. ChatGPT in foreign language lesson plan creation: Trends, variability, and historical biases. ReCALL, 2025, 37(3): 332–347.

[7] Lammert C, DeJulio S, Grote-Garcia S, Fraga L M. Better than nothing? An analysis of AI-generated lesson plans using the Universal Design for Learning and transition frameworks. The Clearing House, 2024, 97(5): 168–175.

[8] Baytak A. The content analysis of the lesson plans created by ChatGPT and Google Gemini. Research in Social Sciences and Technology, 2024, 9(1): 329–350.

[9] Powell W, Courchesne S. Opportunities and risks involved in using ChatGPT to create first grade science lesson plans. PLoS ONE, 2024, 19(6): e0305337.

[10] Cooper P K. Music teachers' labeling accuracy and quality ratings of lesson plans by artificial intelligence (AI) and humans. International Journal of Music Education, 2024, 44(1): 23–36.

[11] Karaman M R, Goksu I. Are lesson plans created by ChatGPT more effective? An experimental study. International Journal of Technology in Education, 2024, 7(1): 107–127.

[12] Tekin S, Atar C. "Hey, ChatGPT! Prepare me a lesson plan...": How EFL teacher candidates enact AI-generated lesson plans. Journal of Learning and Teaching in Digital Age, 2026, 11(3): 220–236.

[13] Nyaaba M, Zhai X. Developing a theory-grounded AI tool for the generation of culturally responsive lesson plans. Computers and Education: Artificial Intelligence, 2025, 9: 100474.

[14] Lan X, Yao W. Quality assessment and feasibility analysis of AIGC-assisted lesson plan development. Journal of Guangxi Vocational and Technical College, 2025, 18(6): 7–13. (In Chinese)

[15] Porter A C. Measuring the content of instruction: Uses in research and practice. Educational Researcher, 2002, 31(7): 3–14.

[16] Webb N L. Issues related to judging the alignment of curriculum standards and assessments. Applied Measurement in Education, 2007, 20(1): 7–25.

[17] Li S, Deocampo M F. An analysis of senior high school English textbooks' core competencies using China's English curriculum standard framework. Arab World English Journal, 2022, 13(4): 162–182.

[18] Näsström G, Henriksson W. Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. Electronic Journal of Research in Educational Psychology, 2008, 6(3), No. 16: 667–690.

[19] Feinstein A R, Cicchetti D V. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 1990, 43(6): 543–549.