World Journal of Educational Research
ISSN 2375-9771 (Print) ISSN 2333-5998 (Online)
Vol. 13, No. 5, 2026
www.scholink.org/ojs/index.php/wjer

Original Paper

From Affordance to Coupling: Reconstructing Foreign Language Pedagogy through an AI-Enabled Production-Oriented Approach

Xuan Sun[1*]

[1] College of Arts and Sciences, Kunming, China

[*] Corresponding author

Received: October 22, 2026 Accepted: November 06, 2026 Online Published: November 25, 2026

doi:10.22158/wjer.v13n5p223 URL: http://dx.doi.org/10.22158/wjer.v13n5p223

Abstract

Generative artificial intelligence (AI) is now a routine presence in foreign language classrooms, yet the gap between what the technology can do and what it actually teaches remains poorly theorised. This paper argues that the gap is structural, and that it is best analysed at the level of the pedagogic move. Taking the Production-Oriented Approach (POA) as its anchor, the paper identifies three chronic constraints in POA practice: the cost of evaluating student output, the difficulty of matching input to diagnosed gaps, and the delay between production and feedback. It then proposes the construct of an AI-enabled Production-Oriented Approach (AI-POA), which frames AI as a change in the resourcing and pacing of an existing cycle. Three affordances of generative AI (generative, dialogic, evaluative) are specified, together with four conditions under which they couple with instruction: goal alignment, temporal placement, cognitive appropriateness and teacher control. A nine-move cycle is derived, and implications for objectives, materials, assessment, teacher judgement and institutional arrangements are discussed. The framework is offered as a falsifiable middle-range theory rather than a design recipe.

Keywords

artificial intelligence, production-oriented approach, technological affordance, pedagogical reconstruction, foreign language teaching, coupling conditions

1. Introduction

A first-year reading class of fifty-two students is working through a six-hundred-word popular science text on sleep and memory. Four minutes in, most of the room has the passage open in a translation app.

The page is photographed, the text appears in Chinese, and the comprehension questions are answered from the Chinese. The lesson runs smoothly. Nearly every answer is correct and the session finishes on time. Whether anyone in the room has read the text is a separate question.

What the students did was reasonable. Given an English text pitched above their reading level, in a class too large for individual attention, the translation app solved the problem the task had set: produce answers. The task was answerable without reading, and the app found the shortest route to an answer. This pattern recurs wherever generative tools meet large classes of weaker readers. The technology has not corrupted the exercise. It has exposed what the exercise was rewarding.

Some parameters are worth stating, because the argument depends on them. The episode is a composite assembled from classroom observation, not a transcript of a single lesson. The class has fifty-two students, and roughly two-thirds entered university with an English score at or below 90 out of 150 in the national college entrance examination (gaokao). These are cohort-level descriptors of the kind that institutional research reports as a matter of course; no individual record is used and no institution is named. Classes of this shape are ordinary in provincial and private universities in western China, which is where the analysis below is anchored.

Episodes of this kind are easy to collect and hard to interpret. The familiar readings are that students are cheating, that the tool is too easily available, or that the text was set too hard. None of them explains how an exercise could be completed without the skill it was designed to develop. That is a design question, and it is the question this paper pursues. Generative AI has made cheap a set of activities that reading instruction had treated as evidence of comprehension, and an activity that becomes cheap loses its evidential value. The chain from what a student does to what a student can do is broken at the point where the doing can be outsourced. Teachers notice the break quickly. What they lack is a way of saying which part of their practice has been undermined, and therefore what to rebuild.

Public discussion of AI in language education has moved through three overlapping phases, and it is worth being explicit about them because each has left a sediment in the literature. The first asked whether the technology should be admitted at all. The second catalogued what it could be used for: generate a lesson plan, simplify a text, mark an essay, produce a dialogue. The third, where the more careful work now sits, asks under what conditions a given use improves learning. This progression is healthy, but it has left a gap. We have a large and growing stock of cases and a very thin stock of explanations. Case collections tell a teacher what someone else did in a different institution with a different cohort. They do not tell them what will happen in their own room on Thursday morning.

Three biases in the existing literature make the gap wider than it needs to be. The first is descriptive: accounts of practice tend to report the sequence of actions (the model was prompted, the output was checked, the students discussed) without modelling the mechanism by which a technical property becomes a learning effect. The second is granular: studies usually isolate one moment, usually feedback or material generation, and say little about how that moment sits inside a longer cycle. The third is

evaluative: acceptance, motivation and satisfaction are measured far more often than learning, partly because they are easier to measure and partly because they respond quickly.

The Production-Oriented Approach (POA) is taken here as the anchor, for three reasons. It is a theory with explicit stages, so a claim about where technology enters can be checked rather than asserted. It was built in response to a specific institutional constraint, namely large classes with limited teacher time, which is precisely the constraint that generative AI alters. And it already contains an unresolved tension about evaluation, which turns out to be the point at which AI bites hardest. POA is also, for better or worse, the most widely adopted indigenous framework in Chinese foreign language teaching, which means that any argument about it has immediate practical consequences for a large number of classrooms.

The paper asks three questions. What does AI actually change inside POA? Under what conditions does that change hold? And what follows for design, for teacher development, and for the institutions that employ both? The answer, in brief, is that AI changes the economics of three pedagogic moves rather than the logic of the method; that the change holds only when four conditions are jointly satisfied; and that the practical consequences fall mainly on assessment policy and on the recognition of design work, not on software choice.

A word on method. This is a conceptual paper and reports no intervention. No design proposal here is dressed in the vocabulary of evidence. Empirical claims are either cited or marked as conjectures that a later study could test, and the propositions at the end of Section 4 are phrased so that they can fail, which is as far as a conceptual paper can go.

2. The Production-Oriented Approach and Its Three Unsolved Problems

2.1 A Theory Organised around Output

POA grew out of a dissatisfaction with input-heavy instruction in Chinese university English classes, where students could parse a text and still be unable to say anything with it. Wen's formulation moved through an output-driven hypothesis, then an output-driven and input-enabled hypothesis, and finally a full architecture organised around three principles (learning-centred, learning-using integrated, whole-person education), three hypotheses (output-driven, input-enabled, selective learning) and three phases of instruction (motivating, enabling, assessing) (Wen, 2015, 2020). The English-language statement of the theory is available in Wen (2018), and the design of enabling activities is set out at length by Qiu (2020).

The taxonomy matters less here than the causal claim underneath it: that the attempt to produce something creates a gap the learner can feel, and that a felt gap organises attention in a way that a teacher's announcement of importance does not. The claim is about motivation and the allocation of working memory, and it is why POA insists that production comes first, even when the learner is not ready. The discomfort is the point.

The framework has been revised repeatedly, which reads as a sign of health. But three difficulties have persisted through every revision, and none of them is a theoretical oversight. They are capacity problems, and capacity problems are exactly the kind that technology can sometimes dissolve.

2.2 Evaluation: The Capacity Problem

Teacher-student collaborative assessment is POA's most original contribution (Wen, 2016). The idea is elegant: the teacher selects one or two representative scripts, the class evaluates them together against a shared rubric, and the evaluative competence developed on those samples transfers to the students' own work. A single teacher's attention is thereby distributed across a whole class without pretending that every script has been read.

In practice the design strains at two points. Selecting a representative sample requires the teacher to have read enough scripts to know what is typical, which is the very labour the design was meant to reduce. And the classroom discussion of a sample takes time, so it can be done for perhaps two or three samples per unit, not for thirty. Both strains grow with class size. In a class of fifty, which is ordinary in the institutions POA was designed for, the teacher is caught in a loop: to save time they sample less, and because they sample less they know less about what the class actually needs.

The consequence is a quiet retreat from the theory's own standard. Teachers still run the collaborative assessment, but they run it on scripts they suspect are representative rather than on scripts they know are. The evaluative act becomes ceremonial. The pattern is common enough in reported practice to count as a structural effect.

2.3 Input: The Fit Problem

POA requires enabling materials that are precise (aimed at the gaps just diagnosed), gradual (pitched at more than one level) and varied (in modality and in genre) (Qiu, 2020). These three requirements pull against one another. Precision wants material cut to this class's gaps. Gradualness wants several versions of the same content. Variety wants text, image, audio and video. Satisfy all three for a single unit and the preparation load exceeds what one teacher can carry across four or five courses.

The compromise that teachers actually make is to lower precision first. It is the least visible compromise: nobody in the room can tell that the text was chosen because it was available rather than because it addressed the gap that surfaced on Tuesday. Variety goes next, then gradualness. What remains is a single text, taught well, which is more or less the input-led classroom that POA set out to replace.

The point is worth dwelling on, because it explains a pattern that otherwise looks like teacher resistance. Teachers are not refusing to differentiate. They are rationing a scarce good, and they ration it in the order that produces the least immediate complaint.

2.4 Feedback: The Timing Problem

POA's cycle is organised by unit, which means feedback arrives in weeks. The problem with delayed feedback is not that it is late in a bureaucratic sense. It is that the cognitive situation in which the error occurred has evaporated. By the time a student reads a comment on their paraphrase of paragraph three,

they can no longer reconstruct what they were trying to do, so the comment attaches to nothing. Hattie and Timperley's (2007) account of feedback makes the point in general terms that feedback works when it addresses the gap between current and desired performance while the performance is still available to the learner.

There is a second-order effect. Because feedback is expensive, it is concentrated on the final product, and the final product is where revision is least useful. The drafts where a comment could still change something go unread.

2.5 A counter-example worth Taking Seriously

One finding in the POA literature is regularly cited as evidence against the approach and deserves a direct response. Zhang (2017) compared POA with a conventional method over one semester and found significant gains in listening and writing but a significant decline in reading. On the face of it this is awkward for anyone proposing to build a reading-oriented framework on POA.

The result admits another reading. In that design reading functioned as input to the writing task, with no standing of its own. Students read in order to extract material for their compositions, which means they practised search reading and gist reading and never practised the slower, more effortful reading that a comprehension test rewards. The decline is not evidence that POA damages reading; it is evidence that an unexamined input role produces unexamined reading. If reading is to improve, it needs its own production task, its own gap, and its own evaluation. That inference shapes the design in Section 5, where the reading class is reorganised so that the output is a reading-derived product rather than a writing exercise that happens to require reading.

3. Three Lines of Research on AI in Language Education

3.1 Tool Efficacy

The largest body of work asks whether a particular tool improves a particular outcome, usually with a quasi-experimental design and a comprehension post-test. The results are broadly positive and broadly fragile. Wang et al. (2024) built a personalised reading support system on a large language model and reported gains in comprehension and in reported strategy use. Shang et al. (2025) staged a competition between learners and a model and found that the competition condition outperformed both the model-alone and the learner-alone conditions, which is a more interesting result than a simple gain. Lin and Chen (2024) tested the model's capacity to generate multiple-choice items and found adequate but uneven quality, with item types differing sharply in reliability.

What this line establishes is that effects exist and can be detected. What it cannot establish is transfer, because the treatment is usually a specific tool version embedded in a specific sequence over a short period. Balantekin (2025) reports gains in both comprehension and reading motivation from model-generated texts.

3.2 Learner Experience

A second line documents acceptance, use patterns and affect, often through the technology acceptance model or self-determination theory. Alyami's (2026) study of model-mediated mind mapping is a good example of the genre done well, combining a comprehension measure with an account of how students used the tool. This work explains why students adopt a tool and how they feel about it. It says much less about what they learn, and the two are not interchangeable: a tool can be enjoyable and instructionally inert, and a tool can be irritating and effective.

3.3 Instructional Design

The third line asks how the tool is woven into teaching, and it is the closest to the argument of this paper. Wen's group has produced a series of studies on POA with AI support, covering textbook writing, textbook use and the design of enabling activities (Wen, 2025a; Zhang, 2025; Qiu, 2025). The textbook-writing study describes a six-step workflow in which the human writer sets the specification, the model drafts, and the writer reviews. The enabling study examines how model-generated scaffolding can be calibrated to diagnosed gaps. Wen (2025b) has also argued that AI makes possible a shift towards immediate communicative use, which reframes the temporal structure of the whole cycle rather than merely accelerating a stage.

These studies are detailed and honest about their limits. What they do not offer, and do not claim to offer, is a general account of why a given design works, in a form that would let a teacher in a different institution predict whether it will work for her.

3.4 The Missing Middle

The three lines share a structural omission. Each operates either above or below the level at which teaching decisions are actually made. Tool-efficacy studies operate below it, at the level of a mechanism tested in isolation. Policy and acceptance studies operate above it, at the level of institutions and attitudes. The level in between, where a teacher decides that this affordance will be used at this moment for this purpose, is thinly theorised.

A theory at that level would need to do three things. It would specify the properties of the technology in terms that connect to pedagogy, not to engineering. It would specify the conditions under which those properties produce learning rather than the appearance of learning. And it would be stated in a form that can fail.

The unit of analysis is the pedagogic move, not the tool, for a practical reason. Tools turn over in about eighteen months; the structure of a teaching cycle turns over in about half a decade. A theory built on tools has a shelf life measured in model releases. A theory built on moves has a chance of outliving the particular model that prompted it, which is the minimum condition for a contribution to a field rather than to a product cycle.

4. Affordances and the Conditions of Coupling

4.1 Three Affordances

The concept of affordance, as Gibson introduced it, refers to the action possibilities that an environment offers an organism with a particular body and a particular purpose (Gibson, 1979). Norman's (1988) later use of the term in design shifted it towards perceived possibility, which is more useful for instruction, because what matters in a classroom is what the teacher believes they can do. Both versions share a commitment worth keeping: an affordance is relational. It sits between the artefact and the user, and neither side owns it.

Read that way, three affordances of generative AI are relevant to language teaching, and they are not equally important.

The generative affordance is the capacity to synthesise multimodal material to specification and to revise it along a parameter: shorter, simpler, more concrete, pitched at a reader who knows nothing about the topic. Text production is the visible part of this. The consequence that matters is a change of regime, from scarcity to abundance, which makes it economically possible to prepare different material for different groups within one class. That possibility did not previously exist at any price a normal department could pay.

The dialogic affordance is the capacity to sustain a multi-turn exchange in ordinary language. What separates it from a search box is that the exchange can be interrogated: the learner can ask for a different explanation, a counter-example, or a challenge to their own reading, and can do so without the social cost of asking a teacher for the fourth time. This approximates the tutorial relation, which happens to be the least scalable form of instruction ever devised.

The evaluative affordance is the capacity to score and diagnose a product against stated criteria, and to do so for every item rather than a sample. Full-coverage formative assessment becomes arithmetically possible. It does not become valid. That distinction carries most of the argument in Section 6.3.

These are described deliberately as affordances rather than as functions, because functions invite a checklist ("can it do X?") whereas affordances invite a question about fit ("for whom, for what, under what arrangement?"). The checklist is what produced the current literature of feature inventories. The fit question is what produces design.

4.2 Four Conditions of Coupling

Affordances do not convert themselves into learning. Between the two there is an arrangement, and the arrangement can be wrong in ways that are invisible in outcome data. Four conditions are proposed, each necessary and, in the settings examined here, jointly sufficient. They are stated as conditions rather than as variables because they are meant to work as a diagnostic. When a design fails, at least one of them will have been violated, and the violation can usually be located.

Condition 1: goal alignment. The action the technology makes possible must serve the purpose the stage was designed to serve. Violations are easy to find once the question is posed. Generating a model essay in the driving stage looks like scaffolding and is in fact an erasure: the point of the driving stage

is that the learner attempts the task and feels the inadequacy of her resources. Give her a model first and she will not attempt, so no gap is felt, so the driving stage has not happened. The technology performed exactly as intended and the pedagogy failed. This is the commonest failure in practice, and it is caused by the plausibility of the output, which makes the substitution feel like help.

Condition 2: temporal placement. The moment of intervention must match the cognitive sequence of the task. In reading, an explanation offered before the learner has attempted inference replaces the inference; the same explanation offered after they have committed to an interpretation becomes a resource for testing it. The content is identical, the effect is opposite, and the difference is entirely a matter of when. This condition explains a puzzle in the empirical literature, where the same tool produces strong and null results across studies. The studies differ in placement far more than in technology.

Condition 3: cognitive appropriateness. The cognitive load the technology absorbs must be extraneous rather than germane. Retrieving a word, checking a collocation and confirming a historical date are extraneous to the reading task; offloading them frees working memory for comprehension, which is exactly what cognitive load theory would predict (Sweller et al., 2011). Summarising the argument, inferring the writer's stance and noticing the rhetorical structure are the learning. Offload those and nothing remains but the appearance of comprehension. The line between the two is not always obvious in advance, and it shifts with proficiency: for a first-year student, decoding a dense nominal phrase may be extraneous; for a second-year student it may be the exercise.

Condition 4: teacher control. The teacher must retain final authority over objectives, content and standards, and must be able to understand and correct what the system produces. This is sometimes presented as a matter of professional dignity. It is better understood as a validity requirement. If the teacher cannot inspect the basis of a judgement, they cannot certify it, and an uncertified judgement is not an assessment. Where the model's reasoning is opaque, the teacher's signature becomes a fiction, and the institution is awarding credit on the strength of a process nobody has examined.

The four conditions are not a hierarchy. In practice, however, the fourth is the one whose absence makes the others unenforceable, because a teacher without authority cannot refuse a misaligned use, cannot reschedule an intervention, and cannot rule that a particular load is germane. Section 6.5 returns to this, arguing that the decisive constraint on AI-POA is institutional.

4.3 Four Propositions

The framework yields four claims, stated so that they could be wrong.

P1 (matching). The pedagogic value of a technology in a given stage depends on the fit between its affordances and the demands of that stage, not on the capability of the model. A weaker model well placed will outperform a stronger model badly placed. This is testable by crossing model capability with placement in a factorial design, where a substantial interaction would be expected.

P2 (mediation). The effect of AI on language ability is carried mainly through pedagogic mechanisms, with a small direct path. Operationally, the affordances should predict ability largely through the quality

of output, input and assessment experienced by the learner. This is the proposition a companion study is currently testing with a structural model, and it is the one for which the measurement work is furthest advanced (Wen & Ye, 2014).

P3 (moderation). The effect is conditional on learner self-regulation and on the teacher's calibration skill. Students who cannot monitor their own comprehension will use the tool to terminate difficulty rather than to work through it. This predicts a disordinal interaction rather than a main effect, which is why mean differences across groups so often wash out.

P4 (evolution). The construct will change as the technology changes, and the framework should be expected to age. A theory that predicted its own stability would be suspect. What should not change is the level of description: moves, conditions and evidence, not tools and features.

P1 and P4 are conceptual claims whose evaluation requires comparison across cases. P2 and P3 are stated in a form that mediation and moderation analysis can address directly, and that asymmetry is a virtue: a framework that generated only testable claims would be a model, and models at this stage of knowledge tend to be premature.

4.4 Position Relative to CALL and Intelligent Tutoring

It helps to locate the proposal historically, because "AI in language teaching" is often discussed as if it were new. Computer-assisted language learning has been with us since the 1960s (Warschauer & Healey, 1998), and its structural logic has been remarkably stable: the machine provides practice and resources around a syllabus that the teacher owns. Intelligent tutoring systems took a different route, modelling the learner's knowledge state and attempting to supply instruction in place of a teacher where teachers are scarce. Table 1 sets out the contrast.

Table 1. Four Pedagogical Arrangements Compared

Dimension	CALL	Intelligent tutoring	AI-POA
Location of technology	Peripheral to instruction (practice, resources)	At the centre (substitutes for part of the teacher's role)	Inside the cycle (reconfigures its moves)
Pedagogical basis	Behaviourism, drill and practice	Mastery learning	Production-oriented approach
Dominant pathway	Predetermined	Adaptively predetermined	Generated
Object of assessment	Accuracy	Criterion attainment	Quality of production and process evidence
Teacher's role	Selecting resources	Operating the system	Judging and calibrating
Characteristic failure	Detachment from course goals	Narrowed objectives, reduced creativity	Cognitive offloading, invalid assessment

The difference is not one of technical sophistication. It is a difference in where the technology is placed relative to the method. CALL puts it outside the method, intelligent tutoring puts it above the method, and AI-POA puts it inside. Placement determines what the teacher must be able to do: select, operate, or judge.

The comparison also disposes of a common misreading, namely that AI-POA is simply a cleverer CALL. The two overlap in that both supply material. But CALL is judged by whether its resources are used, a question that click data can answer, whereas AI-POA is judged by whether its cycle runs, a question that only evidence of learning can answer. Confusing the two leads departments to purchase platforms and then wonder why attainment has not moved.

5. The Cycle: Three Stages, Nine Moves

Placing the three affordances inside the three stages of POA, subject to the four conditions, yields nine moves. Figure 1 presents them as a cycle rather than a sequence, for reasons that will become clear in Section 5.4.

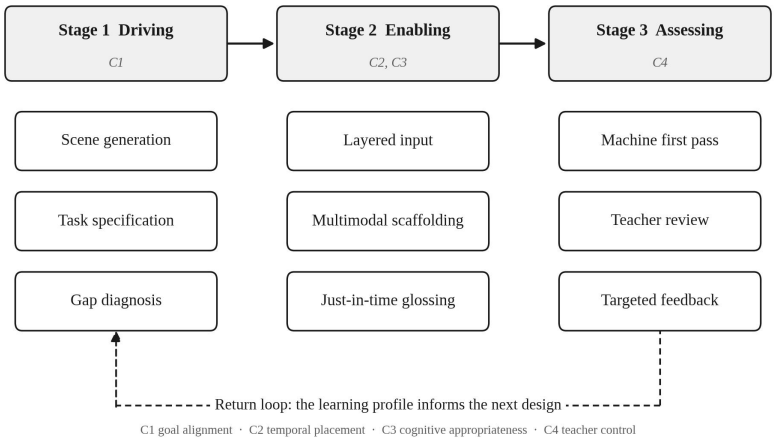

Figure 1. The Three-stage, Nine-move cycle of AI-POA

5.1 Driving

The three moves here are scene generation, task specification and gap diagnosis. The design rule is produce first, calibrate second. Students attempt the task before they are ready, the system characterises what the cohort collectively could not do, and the teacher uses that characterisation to decide what the unit will teach.

The characteristic error is to generate exemplars. An exemplar answers the question the driving stage exists to raise. A subtler error is to let the model diagnose before students have produced anything, by asking it what typically goes wrong with this task type. That produces a plausible list of generic difficulties, which is not the same as this class's difficulties, and it quietly returns the teacher to the position of guessing that POA was designed to escape.

5.2 Enabling

The three moves are layered input, multimodal scaffolding and just-in-time glossing. The design rule is intervene after blockage, not before. Glossing should be triggered by evidence that the student has attempted an interpretation, not offered continuously. In practice this means the tool is not always on, which is unpopular with students and essential to the design.

Two errors recur. The first is the permanently open translator, which violates conditions 2 and 3 simultaneously and produces fluent readers who cannot read. The second is scaffolding that does the task rather than supporting it: a model that summarises the passage when the student was supposed to summarise it has replaced the exercise, however helpful the summary.

5.3 Assessing

The three moves are machine first pass, teacher review, and targeted feedback. The design rule is the machine does not adjudicate. The system scores against the rubric and locates problems; the teacher reviews the boundary cases and makes the final call; the student may contest a score and require an explanation. The characteristic error is to enter machine scores directly into the grade book, which converts a formative instrument into a terminal judgement and destroys the ground for appeal.

Here the capacity problem of Section 2.2 is actually solved. Full coverage becomes affordable, so the teacher's attention can be spent where judgement is needed rather than where it is merely laborious. Note that the saving is reallocated rather than eliminated: the teacher now reads fifty score reports instead of fifty essays, and reading score reports well is a skill that has to be learned.

5.4 The Return Loop

The nine moves describe a unit. The more consequential change is between units. In conventional POA the cycle closes when the unit ends, and what carries forward is the teacher's memory of how it went. With structured process data, the cycle can close at the level of the task, and what carries forward is a record: which gaps recurred, which scaffolds were used and abandoned, which rubric dimensions produced disagreement.

This is the least evidenced part of the framework, and possibly the most consequential. If language development depends on frequent short cycles of attempt, feedback and revision, then halving the cycle length may matter more than improving any single move within it. A class that completes two cycles a week instead of one is not doing the same thing twice as fast; it is doing a different kind of thing. It is offered as a conjecture, flagged as such, because the design that would test it requires a longitudinal comparison that nobody has yet run.

6. Reconstructing the Path

6.1 Objectives: From Ability Lists to Producible Goals

AI cannot act on an objective phrased as an ability. "Improve reading comprehension" gives the system nothing to work with, and gives the teacher nothing to assess. "Read an 800-word popular science text

and explain its central mechanism in 300 words to a non-specialist, with a diagram" gives the system a specification, the student a task, and the teacher a rubric.

This is sometimes dismissed as a bureaucratic refinement. It is not. The move from an ability statement to a producible goal is what makes the rest of the design derivable: the driving task is the goal performed badly, the enabling material is whatever closes the gap between the bad performance and the goal, and the rubric is the goal described as criteria. A unit without a producible goal has to be assembled from materials, which is how most units are in fact assembled.

6.2 Materials: From Selection to Generation and Calibration

The traditional verb is select. The verb that fits the new conditions is calibrate, and it presupposes generation. Three mechanisms are needed if generation is not to become a source of noise.

The first is specification: the teaching intention has to be written as an executable instruction, which is harder than it sounds because most teaching intentions are tacit.

The second is calibration. Every generated item passes four checks, for factual accuracy, for value appropriateness, for difficulty fit and for alignment with the assessment criteria.

The third is accumulation. Calibrated material goes into a departmental bank, so that the second iteration starts from a stock rather than from nothing.

The economics change here in a way that is easy to miss. Under selection, material is a consumable: used once, then discarded with the syllabus revision. Under generation and calibration, material becomes an asset that appreciates as it is corrected. A department that runs this loop for two years ends up with something a new teacher can inherit, which is not true of a stack of photocopies.

6.3 Assessment: From Judgement to Evidence

Full-coverage formative assessment is now affordable, but affordability is not the problem. The problem is whether what gets produced counts as evidence. Three properties are required.

The rubric must be public, so that students know what counts as good before they produce. The evidence must be traceable, so that any score can be referred back to a specific feature of a specific performance. And the gain must be computable, so that a claim of improvement can be checked against a prior measure rather than asserted.

Assessment that satisfies these three properties stops being an administrative instrument and becomes a learning instrument. Assessment that does not satisfy them becomes more plentiful without becoming more useful, which is the likely outcome of most current experimentation: more scores, no better warrants.

6.4 Teachers: From Operation to Judgement

The competence that matters is shifting, and it is not a competence in using software. Four judgements recur, and they map onto the four conditions. Is this material appropriate to this purpose (condition 1)? Is this the moment to intervene (condition 2)? Is this load extraneous or germane (condition 3)? And is the teacher still in a position to overrule the system (condition 4)?

None of these can be taught in a workshop on prompting. They are developed by designing units, watching them fail, and revising, which means that teacher development has to be organised around the production of artefacts. A workshop in which teachers bring a unit and leave with a revised unit will do more than a lecture on the affordances of large language models, and it will take three times as long.

6.5 Institutions: The Constraint that Decides

A design that cannot be run within existing budgets is a proposal, not a reform. Three institutional arrangements determine whether AI-POA survives its first year.

Assessment policy comes first. If the final grade rests on a closed-book examination, a teacher who invests in process evidence is investing in something the institution does not count. The rational response is to minimise process and maximise drill, and no amount of training will change that. What is required is a grade structure in which process evidence carries real weight, with examinations kept but reweighted.

Workload accounting comes second, and it is the one most often overlooked. AI-POA shifts effort from marking to design: more time on scenes, calibration and rubrics, less on correction. If workload is counted in contact hours and scripts marked, the teacher's ledger shows a reduction while her actual hours rise. The arrangement is then experienced as a penalty, and it will be abandoned by the able rather than the unable.

Collective capacity comes third. Calibration and review are too heavy for individuals. They need a teaching group that meets, argues about rubric boundaries, and maintains the material bank. The change is in the organisation of teaching labour, from isolated preparation to shared production, and it is the part of the reform that institutions are least prepared for because it does not appear in any budget line.

Taken together, these three arrangements support a claim that runs against the grain of most writing on educational technology: the binding constraint is institutional. A department with a generous assessment policy and a functioning teaching group will make good use of a modest model. A department with a closed examination and isolated teachers will waste an excellent one.

7. Boundaries and Risks

Theoretical boundary. The framework is built on POA and inherits its scope conditions. For teaching orientations that do not centre on production, such as the appreciative reading of literature where the point is receptivity rather than output, the nine moves would need separate argument. They would probably survive in modified form, with the assessing stage carrying more weight, but this has not been shown.

Cognitive offloading. When a system can supply an interpretation, a student may skip interpretation. The risk is not hypothetical; it is the default behaviour of a well-designed tool used without constraint. Condition 3 is the guard, but condition 3 depends on a judgement that the student is not equipped to make and the teacher cannot always supervise.

Validity of machine scoring. If agreement between machine and human scoring is low and stays low, the evidence base of the whole arrangement collapses. Current systems are reasonably consistent on surface features and much weaker on stance, coherence and appropriateness, which are precisely the dimensions that matter at advanced levels.

Digital divide. Affordances are not equally distributed. Paid tiers are better than free tiers, and a student who can pay for a better model has, in effect, a better tutor. In a system that increasingly assesses process, this converts an economic difference into an academic one.

Erosion of teacher agency. If design can be generated, the teacher's distinctive contribution may be redescribed as curation. The counter-argument, that judgement cannot be generated, is true but not self-enforcing; it has to be institutionalised in workload, in assessment policy and in the right to refuse.

Ethics. Three issues need prior agreement. Data: the scope, retention and secondary use of learning analytics require informed consent and the minimum necessary collection. Copyright: the status of generated material is unsettled, and it matters most for course packages that are sold or published. Integrity: the defensible line is that AI may be used in the process but the product must be defensible by the student, which means assessment has to examine the student's capacity to explain what she submitted, not only the submission.

Limitations. The paper is conceptual throughout and offers no intervention data, and the propositions in Section 4.3 are stated for others to test. The reading focus is deliberate but narrows the argument's reach; the extension to speaking, where interaction is synchronous and the relevant affordances differ, is begun but not completed in Wen's (2025b) account of an immediate communicative turn. The framework should also be expected to date. Its claim to survive rests on a level of description durable enough to be revised. The content will not be.

8. Conclusion

The argument can be put in one sentence. AI does not replace the teacher in language teaching. It changes which designs become affordable: the highly individualised assessment, the tightly fitted input and the immediate feedback that POA has always demanded and that practice has always compromised.

That reframing has consequences. If the constraint was capacity, then the arrival of capacity should be visible as a change in what teachers do, not as a change in what they believe. And it is visible, in the small number of classrooms where the cycle has been rebuilt, and notably not visible in the larger number where a tool has been added to an unchanged cycle.

The four conditions are an attempt to say why. They locate the human in the arrangement as the holder of four specific judgements that the system cannot make and that the institution must therefore protect. That is a structural position, not a sentimental one. A teacher who cannot align, time, discriminate and overrule is not a teacher in this framework, whatever she is called in the timetable.

The framework is tidy, and tidiness is a warning sign. It has four conditions because four survived contact with actual classrooms, not because four is a meaningful number. Others will find a fifth, or will collapse two into one. That would be progress. The value of a construct like AI-POA lies in its being specific enough to be argued with, since a claim that cannot be argued with cannot be improved.

Acknowledgement

This paper arises from a commissioned research project on AI-enabled reading instruction in Chinese higher education. I am grateful to colleagues in the teaching group who allowed their unit designs to be examined and revised, and to two anonymous reviewers whose objections improved Section 4.

References

Alyami, N. (2026). Investigating ChatGPT-mediated mind mapping to facilitate EFL learners' reading comprehension. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0336185

Balantekin, Y. (2025). Improving students' reading, comprehension, and reading motivation through AI-generated texts. *Humanities and Social Sciences Communications*. https://doi.org/10.1057/s41599-025-06265-x

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, *123*, 103344. https://doi.org/10.1016/j.system.2024.103344

Norman, D. A. (1988). *The design of everyday things*. Basic Books.

Qiu, L. (2020). *Designing enabling activities in the production-oriented approach*. Foreign Language Teaching and Research Press. [in Chinese]

Qiu, L. (2025). Research on AI-enabled enabling activities in the production-oriented approach. *Modern Foreign Languages*, *48*(5). [in Chinese]

Shang, J., Huang, Y., Xu, M., Huang, Y., & Shen, X. (2025). Competition between human learners and ChatGPT: Enhancing university EFL students' reading comprehension. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2025.2577356

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Wang, X., Zhong, Y., Huang, C., & Huang, X. (2024). ChatPRCS: A personalized support system for English reading comprehension based on ChatGPT. *IEEE Transactions on Learning Technologies*. https://doi.org/10.1109/TLT.2024.3405747

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, *31*(2), 57-71. https://doi.org/10.1017/S0261444800013161

Wen, Q. (2015). Constructing the theoretical system of the production-oriented approach. *Foreign Language Teaching and Research*, *47*(4), 547-558. [in Chinese]

Wen, Q. (2016). Teacher-student collaborative assessment: A new assessment form created by the production-oriented approach. *Foreign Languages World*, (5), 37-43. [in Chinese]

Wen, Q. (2018). The production-oriented approach to teaching university students English in China. *Language Teaching*, *51*(4), 526-540. https://doi.org/10.1017/S026144481600001X

Wen, Q. (2020). *The production-oriented approach: Exploring a Chinese theory of foreign language education*. Foreign Language Teaching and Research Press. [in Chinese]

Wen, Q. (2025a). AI-enabled textbook writing for the production-oriented approach. *Modern Foreign Languages, 48*(5). [in Chinese] https://doi.org/10.1007/978-3-031-65080-2

Wen, Q. (2025b). AI-enabled POA reform: An exploration of the immediate communicative turn. *Foreign Language Teaching Theory and Practice*, (4), 40-49. [in Chinese]

Wen, Z., & Ye, B. (2014). Analyses of mediating effects: The development of methods and models. *Advances in Psychological Science*, *22*(5), 731-745. [in Chinese] https://doi.org/10.3724/SP.J.1042.2014.00731

Zhang, L. (2017). A study of the teaching effectiveness of the production-oriented approach. *Modern Foreign Languages*, *40*(3), 369-376. [in Chinese]

Zhang, W. (2025). Classroom practice in AI-enabled use of POA textbooks. *Modern Foreign Languages*, *48*(5). [in Chinese]

Notes

Note 1. The classroom episode in Section 1 is drawn from the author's own teaching and is presented as an illustration rather than as data; the class size and task type are representative of the institutional setting described in Section 6.5.

Note 2. "AI-POA" is used throughout as a convenient label for the arrangement described in Section 4, not as the name of a method. The distinction matters for citation: work on POA proper should be cited to Wen and colleagues, and the present framework should not be treated as their position.