Article

From Cognitive Offloading to Cognitive Outsourcing: The Differentiating Effects and Conditional Mechanisms of Generative AI on the Acquisition Pathways of Learners of Chinese as a Second Language

Wenlu Wang[1,*]

[1] Shandong University of Finance and Economics, Jinan, China
* Correspondence: Wenlu Wang, Shandong University of Finance and Economics, Jinan, China

Abstract: Generative AI has become ordinary equipment for anyone studying Chinese outside China. Learners put vocabulary questions to a chatbot at midnight, ask it to mark their compositions, rehearse a visa interview with it, and now routinely ask it to write the composition as well. Research has followed, though unevenly: most of it asks what teachers can do with these tools, and comparatively little asks what the tools do to the learner. This paper takes the second question. I distinguish cognitive offloading, the delegation of cognitive operations to an external system, from selective filtering, the evaluative scrutiny a learner applies to what that system returns. A recent synthesis of process-tracing studies treats the two as opposite poles of a single continuum. I argue that this geometry is wrong and consequential, because on it a learner who offloads heavily and checks carefully becomes hard to distinguish from one who offloads lightly and checks carefully - precisely the distinction a teacher needs. Treating the two as independent dimensions instead yields a four-cell space, and the pedagogically live contrast inside that space is between offloading with filtering and offloading without it. I then trace how generative AI alters the input, interaction and output stages of the acquisition chain, and set out three routes - monitoring sensitivity, perceived accountability, and goal clarity - through which evaluative engagement is held or surrendered. Task level moderates how much is at stake, and the Chinese writing, tone and register systems distribute the risk unevenly rather than uniformly. The paper closes with a Production-Oriented Approach sequence built to keep evaluation in the loop, and with an honest account of what a conceptual argument of this kind cannot settle.

Keywords: generative AI; second language acquisition; international Chinese education; cognitive offloading; selective filtering; evaluative engagement; Production-Oriented Approach

Received: 14 September 2026
Revised: 06 November 2026
Accepted: 20 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The spread of large language models through Chinese-language programmes outside China has been fast enough to outrun the research that should describe it. In international Chinese education the change has been registered largely as a set of institutional and curricular transformations rather than as a change in what learners do with the language [1]. A recent review of the field's own output notes that work on AI-assisted language learning remains heavily product-oriented, organised around post-test gains, error reduction and learner satisfaction [2]. That is understandable. Programmes need to know whether the tools work, and fast. But it leaves a gap that has begun to matter: we can now say a good deal about what AI-assisted learners produce, and rather little about what they do while producing it.

The question I pursue here is what changes inside the learner once the tool is in routine use. This is a question about process, and process questions in applied linguistics are awkward, because the interesting events are brief, internal, and easy to mistake for

their outcomes. A learner who accepts nine of ten AI suggestions and one who accepts three may end up with comparable essays; a learner who produces a fluent paragraph alone and one who assembles it from four prompts may be indistinguishable on the page. If the written product cannot separate them, something other than the product has to do the work.

My argument runs in four steps. First, two constructs that current discussion tends to run together - cognitive offloading and selective filtering - are logically independent, and the received habit of placing them at opposite ends of one scale obscures the contrast that matters. Second, generative AI changes the input, interaction and output stages of the acquisition chain in ways that make this independence newly visible, because it removes the friction that used to do some of the filtering for the learner. Third, the divergence between learners who keep evaluating and learners who stop is not a matter of personality but is conditioned by specifiable mechanisms and by task level. Fourth, the risk is not spread evenly across Chinese: character writing, tone production and register choice sit at very different distances from the outsourcing threshold, and a framework that cannot register that unevenness is not much use to a teacher.

Readers should know what kind of paper this is: a conceptual one. I offer no new data, no intervention, no effect size. The empirical anchoring comes chiefly from one systematic review of process-tracing work on AI-assisted second-language writing [2], supplemented by the standard sources in second language acquisition and by a parallel literature on what has been called metacognitive laziness in AI-supported study [3]. Where I go beyond those sources I say so, and in VI I set out what would have to be true for the argument to fail. The claim is deliberately modest in scope and, I hope, useful in kind: a way of sorting observations that the current vocabulary tends to flatten.

A word on the title is in order. The "from...to" in it names a trajectory of risk that a learner can enter and, with help, leave. It does not describe a developmental sequence that everyone passes through, and II.3 explains why the two constructs involved cannot be stages of a single process. Readers who find the preposition misleading should read it as shorthand for a drift that is common and reversible rather than for a law of development.

2. Two Constructs, and Why the Received Geometry Fails

2.1. Cognitive Offloading

Cognitive offloading, in the sense that has currency since Risko and Gilbert's review, refers to the use of external means to reduce cognitive load, thereby freeing limited internal resources for higher-order processing [4]. The construct is neutral with respect to learning. A writer who does not have to hold spelling in working memory can spend that capacity on whether the argument holds together, and this is not a compromise but a reasonable division of labour. Much of what we teach beginners is, in this sense, the deliberate construction of offloading routines: we give them dictionaries, conjugation tables, and now we give them chatbots.

Note what the definition does and does not say. It concerns the allocation of cognitive operations between person and tool. It says nothing about whether the person inspects what the tool returns. In practice the two get folded together, because Risko and Gilbert's formulation includes the phrase "retains control," and control is ambiguous: it can mean control over what is delegated, or control over what is accepted once it comes back. The first is nearly definitional - if you did not choose to delegate it, you did not offload it. The second is an empirical matter, varies enormously between learners, and is where the pedagogical action is. Much of the confusion in the current literature, I think, comes from letting the first sense of control stand in for the second.

2.2. Selective Filtering

Selective filtering names the second thing: the evaluative scrutiny a learner applies to external input before incorporating it. The term is not mine. Alghamdi and Alghizzi, synthesising thirty-three process-tracing studies of AI-mediated second-language writing, report two recurring profiles. Higher-regulation learners show selective uptake, recursive

evaluation, and extended processing of feedback; lower-regulation learners show rapid acceptance and reduced evaluative engagement [2].

An example may sharpen this. A learner has written yinwei tianqi bu hao, suoyi women mei qu Changcheng ("because the weather was bad, we did not go to the Great Wall"), and the model suggests replacing yinwei ("because") with youyu (a more formal "because"). Accepting the suggestion is not in itself evidence of anything, because youyu is in fact the more formal option and the substitution may be entirely correct. What separates the two profiles is what happens next: whether the learner asks whether the register of the surrounding text supports the shift, notices that the youyu...suoyi frame is awkward in formal writing and that the suoyi ("therefore") should probably go too, and decides. Filtering, on this account, is not the rejection rate. A learner can accept nine suggestions out of ten and still be filtering hard, provided the tenth was rejected for a reason. A learner can reject half of them out of general suspicion and be filtering not at all.

This is why filtering has to be defined by the quality of engagement rather than by its behavioural trace. It also explains why the construct is hard to measure, a problem I return to in VI.

2.3. The Disagreement: Continuum or Orthogonal?

Here I part company with the source I have been relying on. Alghamdi and Alghizzi place the two at the ends of a single scale, metacognitive filtering at one pole and cognitive offloading at the other [2]. The image is intuitive and, I think, mistaken.

A continuum presupposes a single latent quantity along which both constructs can be ordered. If filtering and offloading sit at opposite poles, then more of one entails less of the other, and the model has one degree of freedom. But the two constructs answer different questions. Offloading asks how much of the operation was delegated. Filtering asks how hard the returned material was checked. Nothing in the first constrains the second. The case that breaks the continuum is not exotic; it is the strong student who uses the model as a sparring partner - generating six alternative openings for an essay, reading all six against her own intention, discarding five, and writing the seventh herself. She offloads a great deal and filters hard. On a continuum she has to be placed somewhere in the middle, which is where the model also places a weak student who writes everything himself and checks nothing. That collapse is the cost of the geometry.

Reading the constructs as independent gives four cells. Offloading with filtering is the mode a teacher would want to cultivate: the learner delegates low-level operations and keeps authority over judgement. Offloading without filtering is the risky one, and it is where the outsourcing of the title lives. Working without offloading but with filtering is the classical independent learner, the model most assessment regimes were designed to reward. Working without offloading and without filtering is a learner producing alone and uncritically, reproducing the same errors term after term; I include the cell for completeness, but I have no evidence bearing on it and would not defend it as more than a placeholder.

The practical difference between the two geometries is not academic. On a continuum, the sensible advice to a worried teacher is to reduce AI use, since use and dependence are the same axis. On the orthogonal reading, the advice is different and harder: the amount of use may be left alone, and the instructional work goes into what happens after the output arrives. Teachers I have spoken with tend to reach for the first advice because it is available. The second is what the evidence seems to support.

I should concede something to the continuum view, because I think I know where it comes from. Across a population, the two dimensions correlate: weaker learners tend both to delegate more and to check less, so a scatter plot of real learners will show a diagonal band, and a diagonal band looks like a continuum. The correlation is real. It is also not a definition, and treating a population-level correlation as a conceptual identity is the kind of move that makes a framework unable to describe its own exceptions.

3. The Acquisition Chain under Generative AI

3.1. Input: I+1 When Input Is Infinite

Krashen's input hypothesis holds that acquisition proceeds on comprehensible input pitched slightly above the learner's current level [5]. The formulation is famous mostly for its second half, but the first half carries an assumption that has gone unexamined because it was never in dispute: input is scarce, and getting it costs something. A learner in 1985 who wanted reading material at the right level had to find it, and the finding was work. That work was not incidental to the theory. It is what made i+1 a target rather than a setting.

Generative models remove the cost. A learner can now request unlimited Chinese text on any topic, at any length, in any register, at any hour. The interesting consequence is not that input has become abundant, which is by now banal, but that the difficulty of the input has become a parameter the learner sets. Ask for a text about huanjing baohu ("environmental protection") at too hard a level and you get tan paifang, tan zhonghe, lüse fazhan ("carbon emissions, carbon neutrality, green development"); ask again more gently and you get women yao baohu huanjing, yinwei huanjing hen zhongyao ("we must protect the environment, because the environment is very important"). Both are Chinese. Both are about the environment. The distance between them is a turn of phrase in the prompt.

Two problems follow. The first is mechanical: the model does not know the learner's i. It infers a level from the prompt and from the conversation so far, and it infers badly, because learners are poor at reporting their own proficiency and because fluent output from the model is a misleading signal. In my experience the typical failure is oscillation - a text that is far too hard, followed by an over-correction into something a learner could have read a year ago.

The second problem is theoretical and more interesting. If input can always be tuned to the edge of comprehension, the learner may never be pushed past it. The literature on desirable difficulties is relevant here: conditions that make study feel harder and slower frequently produce better retention than conditions that make it feel smooth [6]. A model that reduces every text to the point of comfortable understanding is, in those terms, an efficient machine for producing the feeling of comprehension without the durable kind. This suggests a shift in what input theory should care about in the AI era. Quantity of input is no longer the binding constraint, and comprehensibility is now trivially achievable. What remains scarce is input that imposes a demand the learner has to meet - text that is reachable, but only with effort. That, rather than abundance, is the variable worth theorising.

3.2. Interaction: What the Machine Cannot Supply

The interaction hypothesis holds that acquisition is driven by the adjustments interlocutors make when understanding fails: requests for clarification, confirmation checks, recasts, and the modified input that results [7]. The engine of the account is breakdown. Something goes wrong, the learner notices, and the repair work is where learning happens. Meta-analytic work on corrective feedback supports the general picture while complicating it: feedback helps, and its effects vary by type and by the conditions of delivery [8].

Human-machine interaction reproduces the surface of this and not the mechanism. A learner can rehearse ordering food, complain about a late delivery, or defend an opinion to a model that is infinitely patient and never judges. The practice volume this makes available is genuinely new and probably valuable. But the model does not fail to understand. It repairs the learner's Chinese because it has been trained to, not because communication broke down. The learner therefore never encounters the moment the interaction hypothesis depends on: the experience of having said something that did not land, and having to find another way to say it.

Register is where this shows most clearly. Chinese encodes social relations in choices that a model handles inconsistently. Nin ("the honorific you") and ni ("the plain you") are not interchangeable; Lao Wang (an informal, familiar form of address, "old" + surname)

and Wang laoshi ("Teacher Wang," a title-based, respectful form of address) are not two ways of naming the same colleague; guixing ("your honored surname," a set phrase for asking a stranger's name) belongs in an introduction and nowhere else; a jianyi ("suggestion") letter that opens with wo xiang ni yinggai ("I think you should") reads, to a Chinese reader, as though it had been translated. Models produce these forms fluently and often wrongly, and because the output is grammatically impeccable the error does not announce itself. A learner without the social knowledge to catch it will not catch it.

I should flag the evidential weakness here. This section is inference. The systematic review I have been drawing on covers writing, not speaking [2], and process data on spoken human-machine interaction is thin. The claim that divergence in the interaction stage mirrors divergence in the output stage is a prediction, not a finding, and it is the part of this paper most likely to be wrong.

3.3. Output: Filtering, or the Ghostwriter

Swain's output hypothesis holds that production is not merely a display of what has been learned but a driver of learning, because the attempt to express something exposes the gap between what one means and what one can say, and the gap is where development occurs [9].

Generative AI intervenes at exactly this point, and more decisively than at the other two. Alghamdi and Alghizzi's two profiles [2] map onto two ways of handling the gap. The filtering learner treats the model as an advisor: it checks grammar, suggests alternatives, flags an awkward collocation, and the learner decides. The outsourcing learner hands over the gap itself. The essay may be assembled from a prompt and lightly revised, or dictated sentence by sentence, or generated wholesale and adjusted at the margins.

Both learners, note, are offloading, and in roughly equal measure. What separates them lies downstream: whether anything was evaluated after the output arrived.

The usual framing also misses something about durability. The traditional worry about AI in writing classes is that the essay will be bad: inaccurate, unidiomatic, obviously machine-made. That worry is becoming obsolete, because the essays are now good. The serious worry is the reverse. A good essay that the learner could not have written, and cannot reproduce, and cannot explain, is a more effective way of concealing non-learning than a bad one, because there is nothing in the artefact to raise the alarm. This is what Lan and Niu call cognitive bypassing: delegating a target operation without substantively attempting, evaluating, or reconstructing it [10]. Their emphasis on delayed transfer follows from it. If the only measure is the assisted draft, bypassing is invisible by construction.

The distinction is easier to see in texts than in definitions. Take the task of writing a short letter to a landlord about a heater that has stopped working. The two versions below are constructed for illustration and correspond to the two profiles the review reports [2]; they are not drawn from any particular class.

Version A: Wang xiansheng: nin hao. Wo zhu zai 302, fangjian de nuanqi shang ge xingqi huai le. Wo gei wuye da guo dianhua, tamen shuo san tian lai xiu, danshi dao jintian hai meiyou ren lai. Wanshang hen leng, wo shui bu zhao. Mafan nin bang wo wen yi xia, hao ma? Xiexie nin.

("Mr Wang: hello. I live in 302; the heater in my room broke last week. I phoned the property management office; they said they would come to fix it in three days, but no one has come yet. It is very cold at night and I cannot sleep. Could I trouble you to ask about it for me? Thank you.")

Version B: Zunjing de Wang xiansheng: nin hao! Jianyu benren suo zuzhu zhi 302 shi gongnuan shebei zi shang zhoumo qi chixu guzhang, sui jing duoci yu wuye guanli bumen goutong, zhijin reng wei huo tuoshan chuli. Yejian shinei wendu chixu piandi, yi dui benren zhengchang xiuxi zaocheng yanzhong yingxiang. Kenqing nin yuyi guanzhu, bing xiezhu ducu xiangguan renyuan jinkuai wancheng jianxiu. Ci zhi jingli!

("Dear Mr Wang: hello! Whereas the heating equipment in the room I rent, No. 302, has been continuously out of order since last weekend, and although I have communicated with the property management department on several occasions, it has still not been properly dealt with. The indoor temperature at night has remained persistently low, which has seriously affected my normal rest. I earnestly request your attention and your assistance in urging the relevant personnel to complete the repairs as soon as possible. With best regards!")

Version A is the rougher Chinese, and a teacher would mark two or three things in it: tamen shuo san tian lai xiu wants "san tian yihou"; the danshi clause runs long; and wo shui bu zhao is blunter than the situation calls for. But every choice in it is the learner's, and the letter is pitched at the relationship it belongs to: mafan nin bang wo wen yi xia is how a tenant asks a landlord for a favour. Version B is grammatical throughout and wrong in a way that leaves no trace on the page. Jianyu, suo zuzhu zhi, wei huo tuoshan chuli and kenqing nin yuyi guanzhu belong to a legal notice; no tenant writes to a landlord like this, and a Chinese reader feels the mismatch at once without being able to point to an error. This is the register failure described in IV.3, and it is also the diagnostic one. A learner who produces A can be taught, because the text shows where she is. A learner who submits B has handed in something that cannot be assessed, because there is nothing of her in it.

4. Why Evaluative Engagement Diverges

4.1. Three Routes

If the divergence is downstream of offloading, the next question is what governs it. Lan and Niu's Calibration-Bypassing-Uptake theory provides the most developed account available: it models AI-mediated writing through five linked processes - perceived feedback affordances, dual calibration of source and self, cognitive-affective regulation, the fork between strategic uptake and bypassing, and delayed transfer [10]. Their central move is to make calibration prior to everything else: the learner's capacity to judge both the reliability of the feedback and the adequacy of their own understanding.

What I want to add is narrower than their model and, I hope, more testable. I compress the calibration stage into three routes through which metacognitive capacity shows up in observable behaviour. I present them as hypotheses, not as findings, and I would expect at least one of them to survive contact with data in a modified form.

The first is monitoring sensitivity: the capacity to notice that one is uncertain. This sounds trivial and is not. A learner who does not register that the aspect marker le is missing from wo zuotian qu tushuguan ("I went to the library yesterday") cannot evaluate a suggestion to add it, because there is nothing in her experience for the suggestion to attach to. Learners differ widely in this sensitivity, and the difference is not well captured by proficiency scores, since a learner can be accurate by rule and still be unable to feel a sentence going wrong.

The second is perceived accountability - whether the learner treats the outcome as hers. This is a disposition rather than a skill, and it is more malleable than it looks. A learner who believes the model is responsible for the Chinese will not check the Chinese. One who believes she will be asked to defend every sentence will. The instructional lever here is assessment design, not exhortation, and it is the cheapest of the three to pull.

The third is goal clarity. Evaluation requires a criterion, and a learner without one cannot filter, because there is no basis on which to prefer one version to another. Asked to improve a paragraph, a learner with a clear sense of what the assignment rewards will select suggestions that serve it; a learner without one will accept whatever reads more smoothly, which is usually whatever the model produces, since the model is optimised for smoothness. This is also, incidentally, why vague writing prompts and heavy AI use combine so badly.

4.2. Task Level as a Moderator

The stakes are not constant across tasks, and the same behaviour can be prudent in one setting and disastrous in another.

At the low end sit operations with little cognitive depth: character lookup, pinyin checking, spelling correction, word-order drills, pronunciation modelling. Delegating these is close to costless, and the case for making students do them by hand is weak. A learner who has outsourced every liangci ("measure word") decision to a model has lost almost nothing, because the knowledge involved is a lookup table.

In the middle sit tasks where form and meaning begin to interact: rewriting a sentence for a different audience, organising a paragraph, sustaining a short dialogue, writing a description. Here filtering starts to pay. A learner who checks whether the model's smoother version has changed the emphasis learns something; one who does not begins to accumulate text she cannot account for.

At the high end sit argument construction, rhetorical organisation, cultural interpretation, and academic writing in Chinese. These are the tasks where the learner's own intention is the object being developed, and where delegation does not merely skip the work but replaces the thing the work was for. A learner who asks a model to write a yilunwen ("argumentative essay") on shifou yinggai jinzhi shouji jin xiaoyuan ("whether mobile phones should be banned from school grounds") and submits the result has not written an argument; she has commissioned one. Whatever else she has learned, she has not learned to argue in Chinese, and no subsequent course will supply that, because the occasion for learning it has been spent.

4.3. Where Chinese Concentrates the Risk

Discussions of AI and language learning tend to treat "language" as a single surface. Chinese is not one. Its subsystems sit at different distances from the outsourcing threshold, and the differences are instructive precisely because they cut against any uniform verdict.

Character writing is the clearest case for concern. Chinese characters are logographic, and writing them recruits visuospatial and motor processes that recognising them does not. The evidence on hand versus keyboard is not settled - the classic imaging work found that handwriting and typewriting produce different recognition patterns for new graphemes [11], but it was conducted on non-Chinese scripts, and replications in Chinese have been mixed. What can be said with more confidence is that the integration of form, sound and meaning is what character learning consists in, and that a learner who substitutes voice input for handwriting, or generated text for her own, is removing the practice that builds it. This is the subsystem where outsourcing is easiest to accomplish and hardest to detect: the submitted essay will look better than anything she could write.

Tone is the counter-case, and it is worth dwelling on because it limits the framework. Tone acquisition is a perception-production coupling: a learner has to hear the difference between mā ("mother"), má ("hemp"), mǎ ("horse") and mà ("to scold") before she can reliably produce it, and the production has to be hers. A model can supply a model utterance, can score a recording, can repeat indefinitely. It cannot produce the tones in her mouth. The space for outsourcing here is genuinely narrow, which means the risk profile of Chinese is not uniform but strongly subsystem-dependent - and that is a testable implication, not a rhetorical flourish. If the framework in this paper is right, we should find outsourcing concentrated in writing and register and largely absent in phonology. If we find it everywhere equally, the framework is probably measuring something else.

Register and culture sit with characters on the high-risk side, for a different reason. Chinese encodes social meaning in choices that are grammatical rather than lexical: the nin/ni distinction, the use of qing ("please") and its absence, the difference between jianyi ("suggestion") and yingdang ("should/ought to"), the placement of yixia (a softening particle) as a softener, the four-character pattern that marks a certain kind of formality. A model can produce all of these. Whether it produces the right one depends on a reading of the situation that it does not have and the learner may not have either. The characteristic failure is a text that is correct everywhere and appropriate nowhere - a complaint letter to a landlord that reads like a customer-service template, or an email to a professor that is

grammatically flawless and faintly presumptuous. The learner cannot see the problem, because there is nothing on the page to see.

5. Teaching Against Outsourcing: A POA Sequence

If evaluation is the thing that has to be preserved, the instructional question becomes how to design tasks in which evaluation has something to work on. Wen Qiufang's Production-Oriented Approach offers a serviceable architecture, because its sequencing already assumes that the productive gap is the engine of learning [12].

The logic of POA is output-driven and input-enabled. Learners attempt production first, encounter the limits of what they can do, and feel the gap; targeted input is supplied afterwards, once the gap has made the input worth attending to. This ordering matters for the present problem in a way that is easy to miss. If a learner's first contact with a task is a model's completed version, no gap is ever experienced, and the evaluative machinery has nothing to engage - there is no discrepancy between what she meant and what she can say, because she never tried to say it. The model's output is not better than her intention; it replaces the intention. Put the attempt first and the same output becomes usable, because it now arrives as a candidate answer to a question she has already asked.

A concrete sequence for an intermediate writing class might run as follows. The teacher sets a task with a real addressee - a letter to a landlord about a broken heater, say - and students write a first draft alone, badly, marking with a highlighter every point where they were unsure. That marking is the crucial step and the one students skip. The marked points become the prompts: students put specific questions to the model, not general ones, asking why xiu ("to fix/repair") is right here and xiuli ("to repair") would be wrong, or whether mafan nin... ("may I trouble you...") can carry a complaint without sounding aggrieved. They revise. They then write a short account of which suggestions they took, which they refused, and on what grounds. The teacher assesses the account as seriously as the letter.

That last step is where the design does its work, because it makes filtering visible and therefore teachable and assessable. It also solves a measurement problem that I noted in II.2: if filtering is defined by the quality of engagement rather than by acceptance rates, then a reflective account is one of the few instruments that gets at it directly. It is imperfect - students can write a plausible account of evaluation they did not perform - but it is better than counting revisions, and it can be compared against the draft history if the writing is done in a platform that keeps one.

A second move is to make evaluation itself the task. The teacher brings in several model-generated Chinese texts of varying quality - some idiomatic, some grammatically clean and culturally off, one with a register error that only a native reader would catch - and asks groups to rank them and justify the ranking. This is a task students find difficult and enjoy, and it rehearses exactly the discrimination that filtering requires. It also has the incidental benefit of demonstrating that fluent Chinese is not the same as good Chinese, a lesson that transfers to their own use of the tools.

The sequence above is a design, not a finding, and it is worth being explicit about what would count as evidence for it. Three outcomes would have to be kept apart. The first is task performance: whether the letters improve. The second is the quality of evaluation, which the reflection log is meant to index - whether the student can say why a suggestion was refused, and not merely that it was. The third is delayed transfer, which is the one that matters and the one almost never measured: whether, five or six weeks later and without the tool, the same learner can write a comparable letter unaided, and whether she still notices the places where she is unsure.

A design that separates them is not elaborate. Two intact classes at the same level, the sequence taught in one and a conventional draft-and-revise cycle in the other; the same task at pre-test, immediate post-test and delayed post-test; reflection logs coded for the presence of a reason rather than for length; and, where the writing platform keeps them, draft histories to check the logs against, since students can describe evaluation they never

performed. I have not run this. Until someone does, the argument in this paper is a claim about mechanism, and it should be read as one.

6. What This Argument Cannot Do

A conceptual paper owes its reader an account of its own limits, and this one has several.

The most obvious is that there are no data. The four-cell space in II.3 is analytic, not empirical; I have not shown that learners distribute across it in any particular way, and I do not know the relative frequencies of the cells. The three routes in IV.1 are proposed, not tested, and a plausible outcome of testing them is that two collapse into one and the third turns out to be a proxy for proficiency.

The evidence base is also narrower than the argument's reach. My main empirical anchor [2] synthesises process-tracing studies of writing. The extension of the framework to interaction in III.2 is inference by analogy, and analogy is where frameworks of this kind usually break. Spoken interaction may well show a different pattern, since the real-time demands of speech leave less room for the kind of reflective monitoring that filtering requires - or more, if the social stakes make learners more attentive. I do not know which.

There is a falsifiability problem that deserves to be named. A framework that classifies every instance of AI use as either filtering or outsourcing risks becoming unfalsifiable, because any outcome can be redescribed after the fact. The remedy is operationalisation, and it is not optional. At minimum the framework needs independent measures of three things: the proportion of the target operation delegated; the presence of evaluation, indexed by revision latency, uptake ratio and justification quality; and delayed transfer, measured on an unaided task some weeks later. Without the third, the distinction between filtering and outsourcing collapses into a distinction between two styles of the same performance.

Finally, scope. The argument is about learners of Chinese, and the subsystem analysis in IV.3 is specific to Chinese. How much of it travels to languages with alphabetic orthographies and no lexical tone is an open question, though the register point probably generalises and the character point obviously does not.

7. Conclusion

Generative AI does not push second-language acquisition along one track. It removes a set of frictions that used to structure the learner's encounter with the language, and what learners do in the absence of those frictions varies enough to matter. Some keep checking. Some stop. The difference between them is not how much they use the tools - in my experience the heaviest users are often the most careful - but whether anything survives the moment of acceptance.

For teachers of Chinese this suggests a shift in where the instructional effort goes. The question of whether to permit AI is largely settled by circumstance, and answering it consumes attention that would be better spent elsewhere. The tractable questions are narrower and more concrete. Which operations may be delegated without cost, and which carry the learning with them? What in the assessment makes a student accountable for a sentence she did not compose? Where in the Chinese system - characters and register, on my reading, rather than tone - is the exposure greatest? A POA sequence answers some of these at the level of task design: attempt first, mark the difficulty, ask the model a specific question, then account for what you took.

None of this makes the tools less useful. It makes the learner's share of the work visible again, which is the condition on which the usefulness depends.

Behind the mechanism lies a plainer question: what a language course is for. The answer is older than the tools. A person becomes able to say what she means in a language not her own, and competent texts follow from that. A model that writes better than she can does not disturb the answer; it makes it harder to leave unstated. Once the text no longer needs her, the remaining question is whether she can still say what she means

without it. None of this argues against the tools. They are useful because they serve something they do not themselves supply.

References

1. R. Ma and J. Xu, "DeepSeek empowers five innovative transformations in international Chinese smart education," *Journal of Yunnan Normal University (Teaching and Research in Chinese as a Foreign Language Edition)*, vol. 23, no. 2, pp. 1–10, 2025.
2. L. H. Alghamdi and T. M. Alghizzi, "Metacognitive filtering and cognitive offloading in AI-assisted L2 writing: A PRISMA-guided process-tracing synthesis," *Behavioral Sciences*, vol. 16, no. 7, Art. no. 1229, 2026, doi: 10.3390/bs16071229.
3. Y. Fan, L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, Y. Shen, X. Li, and D. Gasevic, "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance," *British Journal of Educational Technology*, vol. 56, no. 2, pp. 489–530, 2025.
4. E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends in Cognitive Sciences*, vol. 20, no. 9, pp. 676–688, 2016.
5. S. D. Krashen, *The Input Hypothesis: Issues and Implications*. London, U.K.: Longman, 1985.
6. E. L. Bjork and R. A. Bjork, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," in *Psychology and the Real World*, M. A. Gernsbacher, R. W. Pew, L. M. Hough, and J. R. Pomerantz, Eds. New York, NY, USA: Worth Publishers, 2011, pp. 56–64.
7. M. H. Long, "Input, interaction, and second language acquisition," *Annals of the New York Academy of Sciences*, vol. 379, no. 1, pp. 259–278, 1981.
8. S. Li, "The effectiveness of corrective feedback in SLA: A meta-analysis," *Language Learning*, vol. 60, no. 2, pp. 309–365, 2010.
9. M. Swain, "Communicative competence: Some roles of comprehensible input and comprehensible output in its development," in *Input in Second Language Acquisition*, S. Gass and C. Madden, Eds. Rowley, MA, USA: Newbury House, 1985, pp. 235–253.
10. Z. Lan and C. Niu, "AI-mediated writing and durable learning: A psychological theory of calibration, bypassing and uptake," *Frontiers in Psychology*, vol. 17, Art. no. 1866062, 2026, doi: 10.3389/fpsyg.2026.1866062.
11. M. Longcamp, C. Boucard, J.-C. Gilhodes, J.-L. Anton, M. Roth, B. Nazarian, and J.-L. Velay, "Learning through hand- or typewriting influences visual recognition of new graphic shapes: Behavioral and functional imaging evidence," *Journal of Cognitive Neuroscience*, vol. 20, no. 5, pp. 802–815, 2008.
12. Q. Wen, "The output-driven, input-facilitated hypothesis: An attempt to construct a theory of university foreign language classroom instruction," *Chinese Foreign Language Education*, no. 2, pp. 3–12, 2014.

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