



Journal of AI and International Chinese Language Education

Delegation boundaries: A multiple-case study of teacher pedagogical content knowledge for generative AI in Chinese-as-a-foreign-language classrooms

Chuangdong Li^{1*}, Rafael Martín Rodríguez²

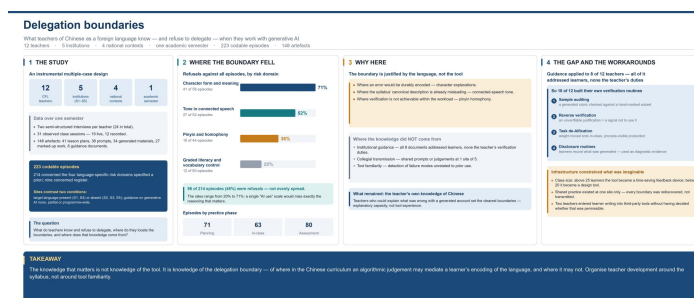
Keywords:

Chinese as a foreign language; teacher knowledge; pedagogical content knowledge; generative AI; human-AI collaboration; teacher development

Citation:

Li & Rodríguez. Delegation boundaries: A multiple-case study of teacher pedagogical content knowledge for generative AI in Chinese-as-a-foreign-language classrooms. JAIICLE 2026, 1(1), 15–27.

Graphical Abstract



Academic Editor:

Justy Chuang

Copy Editor:

Jennifer Mahham

Production Editor:

Jun Wang

Abstract

Generative AI has entered Chinese-language classrooms more as a new participant than a syllabus component, shifting teachers' core task from delivering content to orchestrating a content-generating system. This study asks what knowledge Chinese as a foreign language (CFL) teachers retain rather than delegate when working with generative AI, and whether that knowledge is specific to Chinese as a target language. In an instrumental multiple-case design, we tracked twelve CFL teachers across five institutions in four national contexts for one academic semester, drawing on 24 interviews, 31 observed class sessions, and 148 teaching artefacts. Analysis combined inductive thematic coding with a deductive TPACK-based framework, yielding four cross-case patterns: teachers moved from prompt crafting to task architecture, sequencing activities so generative output appeared only where they kept interpretive decision-making power; they set boundaries for Chinese-distinctive features, refusing to delegate such judgments even when AI output was plausible; assessment demanded the most improvisation, as teachers built verification routines beyond existing institutional guidance; and incompatible infrastructure -workload models, class sizes, absent shared practices-constrained pedagogical possibilities. We propose that AI-specific pedagogical content knowledge for CFL is best defined as knowledge of delegation boundaries, and argue teacher development should center on this concept rather than tool familiarity.

¹ International School of Chinese Studies, Northeast Normal University, China

² Faculty of Philology, Translation and Communication, University of Valencia, Spain

*Correspondence to: Dr. Chuangdong Li (lichuangdong123@gmail.com)

1. INTRODUCTION

For most of the history of computer-assisted language learning, the technology arrived with a designed pedagogy attached. A drill programme was a drill programme; the teacher's decision was whether to use it and when. Generative AI does not arrive with a pedagogy attached. It arrives as a participant that can produce a dialogue, explain a grammar point, generate a graded reading passage, mark a composition, and answer a student's question at eleven o'clock at night. The pedagogical decision is no longer "shall I use the tool" but "for which part of the learning of this language, at this moment, with these learners, should the tool's judgement be allowed to stand".

That decision is a knowledge problem, and it is the teacher's problem. Reviews of AI in education have repeatedly found that the teacher is the least examined element in the technology literature [1], and the early empirical work on generative AI in language teaching has largely examined tools, tasks, and learner perceptions rather than what teachers know and how they come to know it [2, 3]. Where teachers do appear, they typically appear as adopters or resisters, along a diffusion curve, rather than as professionals making situated judgements about the epistemic status of a system's output in relation to a particular curriculum.

This study takes the teacher as its object. Specifically, it asks what teachers of Chinese as a foreign language know about where generative AI may and may not be allowed to mediate learning, and whether that knowledge is bound up with the specific properties of Chinese as a target language. The question is worth asking with Chinese in view for three reasons. First, Chinese is a language in which several central learning objects are precisely the ones on which language models are known to be least reliable: the relation between the form of a character and the morpheme it writes, the disambiguation of homophones in pinyin, the tone of a syllable in connected speech, and the graded sequencing of literacy. Second, a substantial proportion of CFL teaching worldwide takes place in non-target-language environments with learners who read and write Chinese as their principal route into the language; errors that pass unnoticed in a listening-oriented classroom become consequential when literacy is the object of instruction. Third, the CFL teacher workforce is distinctive in the distribution of its own expertise: it includes native speakers with limited formal training in second-language pedagogy, non-native speakers with strong metalinguistic knowledge, and heritage-language teachers without certification, all of whom relate to an authoritative model's output differently [4].

The contribution of the paper is a description, grounded in classroom practice, of an AI-specific pedagogical content knowledge for CFL, and an argument about what its central object is.

2. BACKGROUND

2.1 From TPACK to an AI-specific teacher knowledge

Technological pedagogical content knowledge (TPACK) remains the most used framework for teacher knowledge involving technology [5]. It derives from the earlier argument that teachers possess a knowledge of subject matter for teaching that is distinct from subject matter itself [6, 7]. It has been productive but it has a well-known weakness: it describes knowledge at a level of abstraction that does not predict what a teacher does next [8]. Knowing that one possesses "TPACK" does not tell an observer whether a teacher will accept a generated character decomposition, or whether she will regenerate it. Recent attempts to specify teacher knowledge for AI have moved in two directions. One direction distinguishes knowledge of AI from knowledge for AI-mediated practice, arguing that teachers need to understand a system's epistemic characteristics — its training regime, its failure modes, its tendency to produce fluent but unwarranted output — in order to decide what to do with it [9, 10]. The other direction stays closer to the classroom and describes knowledge-in-practice: the routines by which teachers interrogate, repair, and contextualise generated material [11, 12].

For language teaching, the most promising specification is the one that ties the delegation decision to a property of the language. Content knowledge in language teaching is not a body of facts; it is an internalised model of the target language system, including the parts of that system that are irregular, in flux, or variable across speech styles. Those are precisely the parts on which a statistical model is likely to produce a fluent approximation of the

majority pattern. A teacher who knows this in general still has to know it *for her language*, and to locate it in her syllabus.

2.2 What is expected of the CFL teacher

The CFL literature supplies specific candidates for the parts of the system where delegation is risky.

First, the writing system. Chinese characters are not pictures of their meanings. A minority of characters are genuinely pictographic or ideographic; the large majority are phono-semantic compounds in which one component suggests a semantic field and another hints at a pronunciation that is often only approximately recoverable in modern Mandarin. Generated explanations that present characters as meaning-pictures are pedagogically destructive because they produce a durable but false encoding strategy, and they are hard to dislodge once learners have used them to store hundreds of characters. Character-learning research has consistently emphasised the value of systematic knowledge of component functions and of the distinction between components that carry meaning and those that hint at sound [13, 14].

Second, tone in connected speech. The citation-form contours that appear in textbooks are rare in natural speech; Tone 3 in particular is usually realised as a low tone, and tone sandhi applies obligatorily in some contexts (notably Tone 3 followed by Tone 3) and variably in others [15]. Any system that classifies or corrects tones on the basis of a canonical model will misdescribe a large proportion of ordinary speech, and a teacher who cannot recognise the nature of that error will either transmit it or abandon the tool.

Third, pinyin and homophony. Because pinyin is a romanisation of a heavily homophonous lexicon, generated romanised output requires disambiguation that is easy to fake and hard to verify. A teacher's ability to spot a plausible but wrong homophone is a specific form of content knowledge that matters more in AI-mediated work than it did in textbook-mediated work.

Fourth, graded literacy. Chinese language curricula typically control both vocabulary and character introduction, and the sequencing of character introduction is a curricular artefact with real consequences for reading development. Generated texts do not respect that sequencing unless they are constrained to do so, and the constraint has to be supplied by someone who knows the syllabus.

2.3 Human–AI collaboration as a division of labour

The concept of human–AI collaboration in education is often framed as augmentation: the machine handles routine work, the human handles judgement. In practice, the boundary is not given; it is negotiated, and it is negotiated by teachers who are accountable for outcomes. Research on teacher cognition has established that the judgements of practitioners are governed by knowledge that is largely tacit and is rebuilt in practice [16, 17]. Research on automation in professional work has shown that the boundary between human and machine judgement is drawn through situated experimentation rather than specified in advance, and that the reliance of practitioners on an automated system depends on the trust they place in it rather than on its technical reliability alone [18]. For language teachers, the negotiation is complicated by the fact that the machine's output is fluent, and fluency is easily mistaken for authority. We therefore treat the delegation boundary not as a design parameter of the tool but as an achievement of the teacher — which is what makes it a knowledge object.

2.4 Research questions

- 1) **RQ1.** How do CFL teachers enact human–AI collaboration in planning, instruction, and assessment across one semester?
- 2) **RQ2.** Where do they locate the boundaries of acceptable delegation, and how do they justify those boundaries in terms of the target language?
- 3) **RQ3.** What knowledge do teachers need in order to draw those boundaries, and where does that knowledge come from?

3. METHODS

3.1 Design

The study is an instrumental multiple-case study. The case is the individual teacher's AI-mediated practice over one semester; the instrument is the delegation boundary. A multiple-case design was chosen because the object of interest is a form of practical knowledge that becomes visible only in the accumulation of decisions, and because the cross-case comparison is what allows a claim about what is language-specific and what is generic.

3.2 Settings and participants

Twelve teachers in five institutions across four national contexts (mainland China, the United Kingdom, Spain, and Singapore) participated. Sites were selected for contrast on two dimensions expected to condition practice: whether the target language is present in the surrounding society, and whether guidance on generative AI applied to the programme in which the teacher worked. Table 1 summarises the participant and site characteristics.

Table 1. Participant and site characteristics.

Teacher	Site	Setting	Prior AI use	Class	Learners per class	Guidance on generative AI
T1	S1	Target language present	None	First-year university	24	None
T2	S1	Target language present	Informal	Second-year university	31	None
T3	S1	Target language present	Tool-based	Elective (mixed)	46	Partial
T4	S2	Target language absent	None	First-year university	18	Programme-wide
T5	S2	Target language absent	Tool-based	Second-year university	21	Programme-wide
T6	S2	Target language absent	Extensive	Heritage-track	15	Programme-wide
T7	S3	Target language absent	Informal	Community school	12	None
T8	S3	Target language absent	Tool-based	First-year university	27	Partial
T9	S3	Target language absent	Extensive	Advanced university	19	Programme-wide
T10	S4	Target language present	Informal	Secondary school	22	Partial
T11	S4	Target language present	None	Secondary school	30	Partial
T12	S5	Target language absent	Extensive	Adult evening class	14	None

Note. All names and site identifiers are pseudonyms; to protect institutional anonymity, sites are not mapped to the four national contexts named in Section 3.2. "Prior AI use" refers to the teacher's self-reported use of generative AI in teaching before the study began. Guidance on generative AI is recorded at the level of the programme in which the teacher worked; teachers within one institution could therefore be subject to different documentation. Figure 1 arrays the twelve cases on the two dimensions.

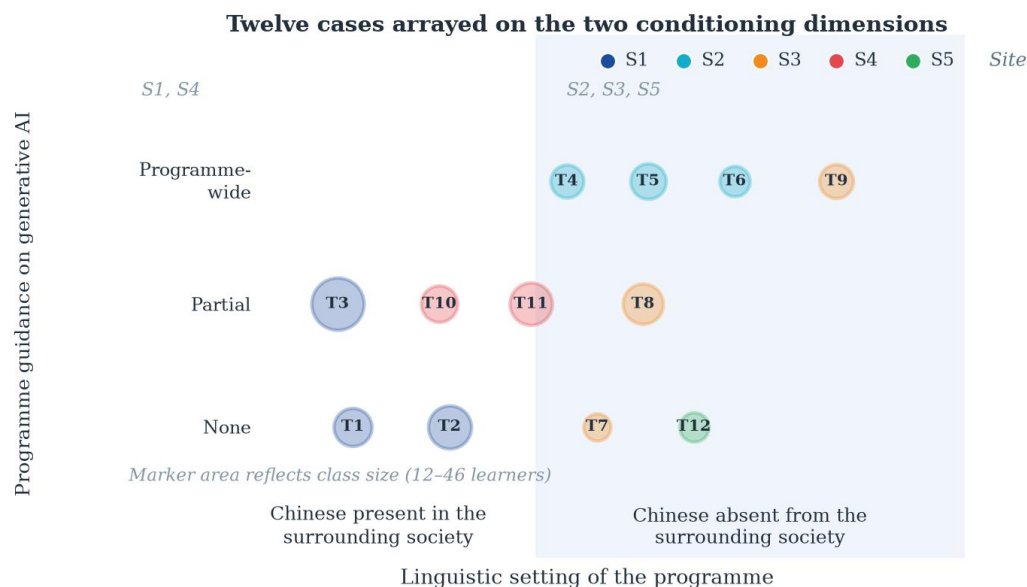


Figure 1. The twelve cases, arrayed on the two dimensions the sites were selected to contrast (linguistic setting; guidance on generative AI), with marker area showing class size.

3.3 Data sources

- 1) **Interviews.** Each teacher was interviewed twice: at the beginning of the semester, covering their prior use, their view of what the tool could be trusted with, and their plans; and in the final three weeks, covering what they had actually done, what they had changed, and what they would not repeat. Interviews lasted 55–80 minutes and were conducted in the language the teacher chose (Mandarin, English, or Spanish), transcribed, and, where necessary, translated into English for coding by the first author with a bilingual second coder checking a subset.
- 2) **Observation.** Thirty-one class sessions were observed, of which 19 were observed live and 12 were recorded with consent. The observation protocol recorded episodes in which the teacher used, modified, rejected, or explained generative output in front of learners.
- 3) **Artefacts.** A corpus of 148 documents was collected: lesson plans ($n = 41$), the prompts or instructions teachers used with generative tools ($n = 38$), generated teaching materials ($n = 34$), marked-up student work ($n = 27$), and departmental or institutional guidance ($n = 8$).

3.4 Analysis

Coding proceeded in two cycles. The inductive first cycle used open coding at the level of the episode (a decision, a refusal, a repair, an explanation). The deductive second cycle mapped codes onto the four candidate language-specific risk domains identified in Section 2.2 and, at a more general level, onto the components of the TPACK framework; the domain-level results are reported in Section 4, and the TPACK-level reading of the patterns is developed in Section 5. Cross-case analysis produced a matrix in which each teacher's episodes were arrayed against the four domains and against the three phases of practice (planning, instruction, assessment). Trustworthiness was addressed by coding a 15 per cent sample of episodes with a second coder (agreement on the first-cycle codes was 0.81 by Cohen's κ), by member checking with eight of the twelve teachers, and by maintaining an audit trail of analytic decisions.

3.5 Ethics, data governance, and positionality

Ethics approval was obtained from the lead institution's research ethics committee and, where required, from the participating institutions. All participants gave written informed consent, including consent for the observation of teaching and the collection of teaching artefacts. Interviews were pseudonymised at transcription, and the pseudonym key was stored separately from the data. Artefacts containing learner names were redacted at the point of collection. Because three authors are themselves CFL teachers, we kept a reflexive journal in which we

recorded instances where our own practice made us read a teacher's decision as more or less defensible than the data warranted; two such instances are reported in Section 6.

3.6 Generative AI disclosure

Generative AI was used in this study in a single role: it was the object of study. It was not used to transcribe or translate interviews, to code data, or to draft the manuscript. The tools and versions teachers used are reported in Table S1 of the supplementary material; because the study ran over one semester, three teachers changed product or version mid-semester, and where this happened the change is treated in the analysis as a distinct episode.

4. FINDINGS

The findings below report the cross-case patterns that emerged from the analysis of the codable episodes, the observed sessions, and the teaching artefacts collected over the semester. All teacher and site identifiers are pseudonyms.

4.1 Overview

Across the semester the twelve teachers produced 223 codable episodes involving generative AI, of which 214 concerned the four language-specific risk domains specified a priori in Section 2.2; the remaining nine concerned register and are reported in Section 4.6. Within the four domains, the distribution across practice phases was uneven: 71 episodes concerned planning, 63 concerned in-class instruction, and 80 concerned assessment or feedback on learner work. Refusals — episodes in which a teacher declined to use generated material or overrode it — accounted for 96 of the 214 episodes, that is, 45 per cent. Refusals were not distributed evenly across the four language-specific risk domains (Table 2; Figure 2).

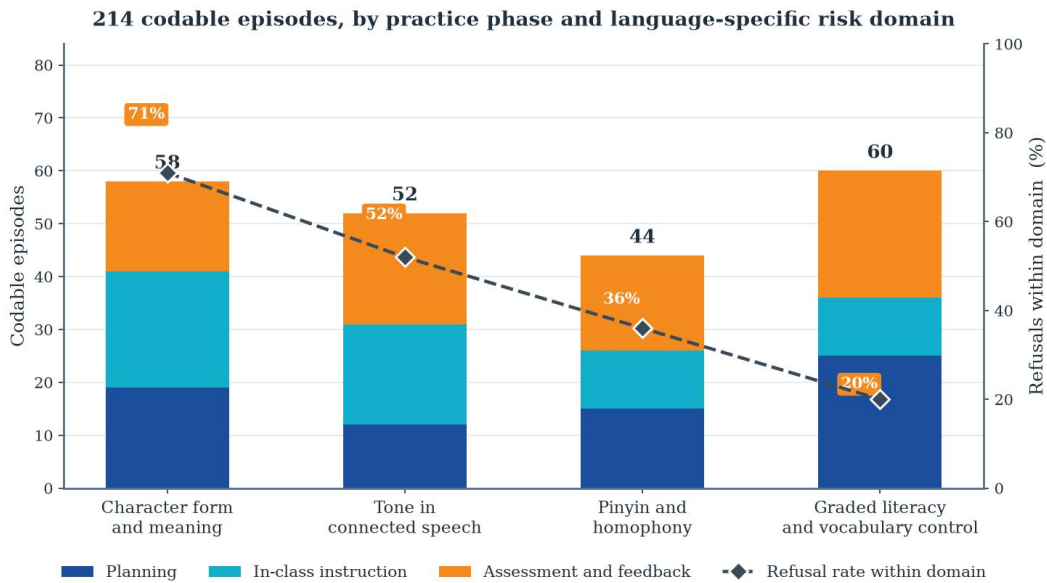


Figure 2. Codable episodes by practice phase within each language-specific risk domain (stacked bars), with the refusal rate within each domain. Note. Values are taken from Table 2; Figure 3 reports the refusal counts of the same domains against their episode totals.

Table 2. Distribution of codable episodes by practice phase and risk domain.

Risk domain	Planning	In-class instruction	Assessment and feedback	Total	Refusals within domain
Character form and meaning	19	22	17	58	41 (71%)
Tone in connected speech	12	19	21	52	27 (52%)

Pinyin and homophony	15	11	18	44	16 (36%)
Graded literacy and vocabulary control	25	11	24	60	12 (20%)
Total	71	63	80	214	96 (45%)

Note. Episodes are decision points, not minutes; a single lesson could generate several episodes. The table covers the 214 episodes coded into the four domains specified a priori; nine further episodes concerned register and are reported in Section 4.6.

4.2 Pattern 1: from prompt-craft to task architecture

Every teacher began the semester by treating generative AI as a production tool — something that makes materials. Ten of the twelve described an early period of what might be called prompt-craft: trying to write instructions that would produce a usable worksheet, dialogue, or explanation. Nine of those ten reported that the results were fluent, plausible, and pedagogically unusable, most often because they did not fit the class's current vocabulary or because the register was wrong.

What changed over the semester was not the prompts. It was the position of the tool in the activity sequence. T5's account is representative:

In September I was asking it to write a dialogue and then I was fixing the dialogue. By November I was asking it to write eight candidate dialogues, choosing two, deleting the endings, and giving the endings to the students to write. The work moved from the material to the choice.

This move from producing artefacts to architecting tasks in which generated artefacts are inputs appeared in nine of the twelve cases and was the single most common developmental trajectory in the data. It has a specific consequence for what the teacher needs to know. Producing materials requires content knowledge and prompt skill. Architecting tasks requires knowing which step of the learning sequence is appropriate for mediated judgement, which is a curricular judgement. Teachers who made the move described their own contribution as having shifted from content to sequence; teachers who did not make it described themselves as increasingly frustrated with the tool's quality. By the second interview ten of the twelve teachers reported using generative AI more than they had at the start of the semester, including all three who had reported no prior use.

4.3 Pattern 2: the boundary around character form and connected-speech tone

The clearest and most stable finding is that teachers refused to delegate judgement on the relationship between character form and morpheme. Forty-one of the 58 episodes coded in that domain were refusals, and the refusals were unusually uniform across settings, levels, and teacher profiles. T1, a native-speaker teacher with eighteen years of experience, put the objection in terms of the durability of the error:

If I let it explain 妈 as "a woman next to a horse, so mother", the student will remember it, and in three years I will have to remove it. I would rather spend two hours preparing the character list myself.

T8, a non-native teacher with strong metalinguistic training, made the same refusal on different grounds: the generated explanations were not wrong so much as unfalsifiable, and she had no way to certify them quickly enough to use them in class. The refusal, in both cases, was not a refusal of the tool. Both teachers used it extensively for other purposes. It was a refusal within a domain, and it was justified by a property of the language — the fact that character explanations become retrieval structures for learners and are therefore expensive to revise.

The pattern for tone in connected speech was similar in kind but weaker and more variable. Twenty-seven of 52 episodes were refusals. Six teachers described a specific incident that reorganised their practice, in which a tool judged a connected-speech Tone 3 token as Tone 2. Three of those six responded by abandoning the tool for pronunciation work entirely; three responded by keeping it but adding a verbal framing for learners, along the lines of T9's routine:

I now say, before we start: in real speech, third tone is low, it is not the valley you learned in the textbook. The app will say "rising" for things you know as third tone. You are not wrong and the app is not wrong; the model is reading the textbook description.

This framing is an act of content knowledge deployed to repair a technical artefact's pedagogical effect. It also shows how a language-specific risk domain generates a specific pedagogical move, which is the substance of the claim that this is a knowledge about Chinese rather than a knowledge about AI. Figure 4 places the four domains on the three-zone account of delegation developed in Section 5.1.

Notably, teachers without formal training in Chinese linguistics were not less likely to detect the problem; they were less likely to be able to name it. Detection was likewise unrelated to prior tool experience: teachers whose self-reported prior use was none or informal refused in this domain as often as those whose prior use was extensive. Where they detected it and could not name it, their repair was usually to stop using the tool for that purpose. Figure 3 sets the refusal share of each domain against its total episode count.

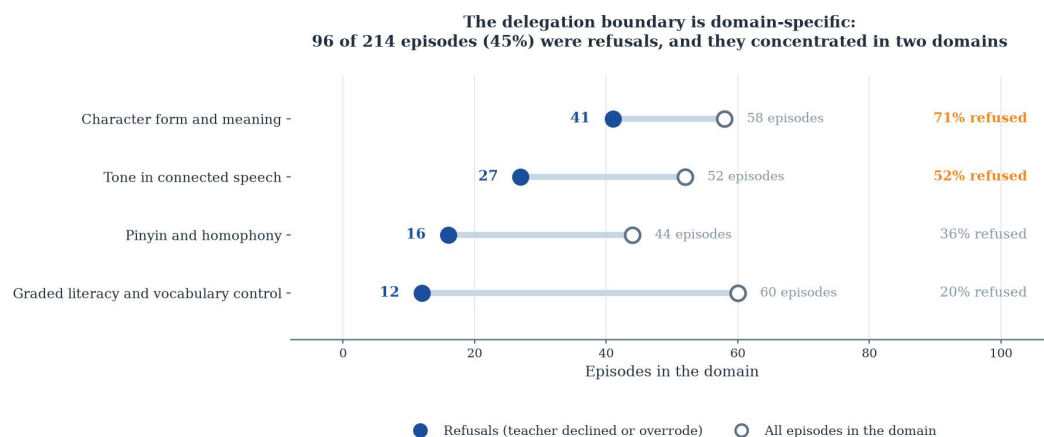


Figure 3. Where the delegation boundary fell: refusals against all episodes, by risk domain. Note. Values are taken from Table 2.

4.4 Pattern 3: assessment as the site of improvisation

Assessment produced the largest number of episodes and the widest gap between what teachers did and what they had been told to do. Guidance on generative AI applied to the programmes of eight of the twelve teachers (programme-wide in four cases and partial in four), but in every case it addressed the use of the tool by learners rather than the verification obligations of the teacher. Ten of the twelve teachers had therefore constructed their own verification practices, which clustered into four types.

First, *sample auditing*: checking a generated marking rubric or a set of graded texts against a small hand-marked subset before using it. Second, *reverse verification*: asking the tool to justify a judgement and treating an unverifiable justification as a signal not to use it. Third, *task de-AIfication*: redesigning assessment so that the tool's output could not substitute for the learner's, most often by moving weight onto in-class, process-visible production. Fourth, *disclosure routines*: requiring learners to record which parts of a submission had been generated, with an explicit statement of what the teacher would do with the information — in T4's case, adjust feedback rather than penalise.

The fourth type deserves comment because it was the least anticipated and the most consequential. T4, whose learners were preparing for a university-level exit examination with a handwritten component, described the disclosure routine as the only way to keep the AI part of the learners' study visible to her:

I do not want to police it. I want to know what they used it for, because that tells me what they cannot yet do. If they used it to make the dialogue more natural, fine. If they used it to write the character introduction, that is a signal for me.

This is an assessment argument: the disclosure is used as diagnostic information about learner competence rather than as a compliance device. It also shows a teacher treating the tool's presence in the learner's workflow as a source of assessment evidence, which is a genuinely new assessment design problem.

4.5 Pattern 4: infrastructure as a constraint on pedagogically imaginable practice

The final cross-case pattern concerns what teachers could do rather than what they knew. Three features of the working environment recurred across all five institutions.

Class size and workload. Teachers with classes above 25 learners described their AI use as a time-saving device for feedback, while teachers with classes below 20 described it as a design tool. T2, with 31 learners in a second-year course, reported generating individualised vocabulary lists for every learner each week — a practice that would be unthinkable by hand and whose verification she had no time to perform, as she acknowledged herself.

Absence of shared practice. In only one site did two or more teachers share prompts or share judgements about output quality. Teachers elsewhere worked alone, which meant that each teacher's boundary was rediscovered rather than transmitted. The three teachers who changed tool or version mid-semester had no institutional route to ask whether the change affected the behaviour they had learned to rely on.

Guidance written for learners, not teachers. All eight guidance documents collected in this study addressed learner use. None addressed the teacher's own verification obligations, the conditions under which generated materials may be distributed to learners, or what to do about data from learner work being entered into a third-party system. Teachers filled this gap individually and sometimes inconsistently; two teachers described entering learner writing into a tool for feedback without having considered whether that was permissible, and one subsequently stopped.

4.6 Variation across settings

The four cross-case patterns were not evenly distributed across the five sites, and two contrasts are worth reporting because they bear on how generalisable the delegation-boundary claim is.

The first contrast concerns the presence of the target language in the surrounding society. Teachers at S1 and S4, where Chinese is present in the surrounding environment, used generative AI most heavily for explanation — producing alternative accounts of a grammar point, generating example sentences for a character already met, answering a learner's question that the teacher had not anticipated. Teachers at S2, S3, and S5, in non-target-language settings, used it most heavily for input generation — graded texts, dialogues, listening scripts. The difference is intelligible in terms of what each setting lacks. Where ambient input is available, the scarce resource is individual explanation; where it is absent, the scarce resource is text at the right level. It also implies that the same tool addresses different pedagogical problems in the two settings, and therefore that the delegation boundary will fall in different places: the input-generation use puts weight on the teacher's ability to police vocabulary and character control, whereas the explanation use puts weight on the ability to detect a plausible but wrong account of the system.

The second contrast concerns the heritage track. T6, teaching fifteen learners in a heritage-track course at S2, diverged from every other case in the dataset in one respect: she described her principal use of generative AI as generating *oral* practice and *register contrast* material, because her learners' need was to move from conversational Chinese toward written Chinese, and she could not find published materials pitched at that transition. In her setting the risk domain that dominated was neither character form nor tone but register, a domain that barely appeared in the coding of the other eleven cases (nine episodes in total, seven of them hers). This is a single case and cannot support a general claim. It does, however, suggest that the set of language-specific risk domains is a function of the learner population as much as of the language, and that a programme whose population includes heritage learners will face a delegation problem that a programme without them never encounters.

5. DISCUSSION

5.1 The object of an AI-specific PCK for CFL is the delegation boundary

The findings suggest that the most useful description of what the teachers in this study knew is not a list of skills and not a level of technology confidence, but a set of boundaries and the reasons for them. Two properties of these boundaries are worth emphasising.

First, they are *domain-specific and asymmetric*. Refusal rates ranged from 20 per cent in graded literacy and vocabulary control to 71 per cent in character form and meaning. A framework that treats "AI use" as a single behaviour would score a teacher who uses the tool for vocabulary lists and refuses it for character explanations as

a partial adopter. On our account she is doing precisely the right thing, and the interesting question is the reasoning behind the boundary, not its location on a scale.

Second, the boundaries are *justified by properties of the target language* rather than by properties of the tool. Teachers refused where an error would be durably encoded (character explanations), where the canonical description in their own syllabus was already misleading (connected-speech tone), and where verification was not practically achievable within their workload (pinyin homophony). This is the sense in which the knowledge is a pedagogical *content* knowledge: it joins an understanding of what learners do with Chinese to an understanding of how the model produces it.

RQ3 also asked where the knowledge needed to draw these boundaries comes from. In this dataset it came from none of the sources that the surrounding infrastructure assumes. It did not come from institutional guidance, since all eight guidance documents collected addressed learner use and none addressed the teacher's verification obligations (Section 4.5). It did not come from collegial transmission, since in only one site did teachers share prompts or judgements about output quality (Section 4.5). And it did not come from tool familiarity, since detection of the language-specific failure modes was unrelated to prior use (Section 4.3). What remained was the teachers' own disciplinary knowledge of Chinese, exercised through a semester of trial, error, and repair: the teachers who could explain what was wrong with a generated account were the ones who set the clearest boundaries (Section 4.3), and it is that explanatory capacity, rather than tool experience, that development would need to strengthen (Section 5.5).

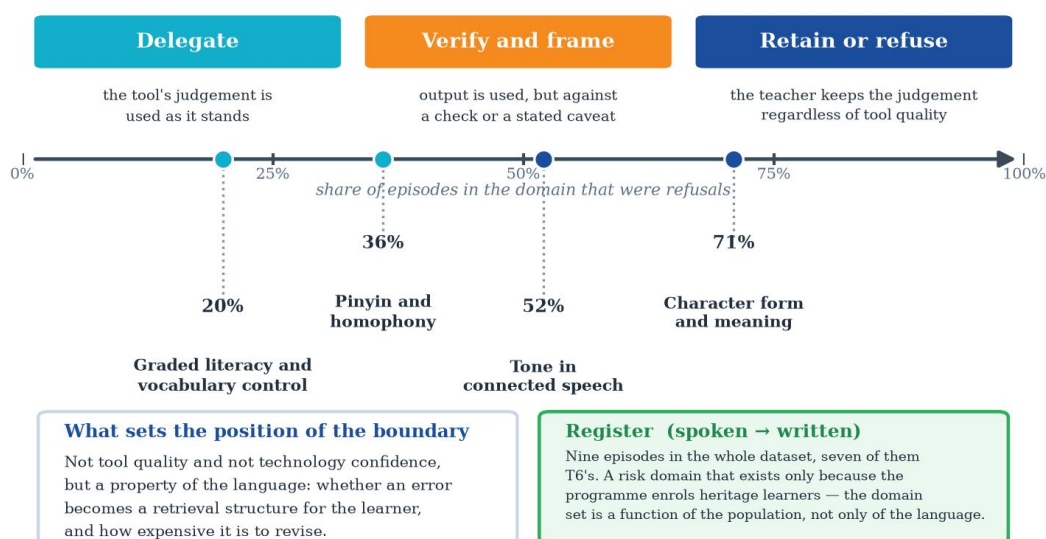


Figure 4. The delegation boundary as a three-zone account, with the four language-specific risk domains placed at their observed refusal shares, and the register domain that exists only where the programme enrolls heritage learners.

5.2 Refusal as professional knowledge, not resistance

The teachers in this study were not technophobes. Ten of twelve made more use of generative AI by the end of the semester than at the start (Section 4.2), and the teachers who refused most often were among the most frequent users of the tool (Section 4.3). Their refusals were targeted and articulate. This matters for how teacher development is conceived. Much current professional development aims at adoption: familiarity with tools, prompt writing, classroom examples. If the operative knowledge is boundary-setting, then adoption-focused development addresses the wrong thing, and may even work against the target, because it supplies the confidence to use a tool without supplying the content knowledge required to know where its output must not stand. A development programme built around the delegation boundary would instead begin from the syllabus, identify the points at which algorithmic judgement would mediate a durable learner encoding, and work outward from there.

5.3 Human–AI collaboration as a division of labour that teachers author

The second implication concerns the increasing use of the phrase human–AI collaboration in policy and in the literature. In this dataset, the collaboration was not a division of labour that had been designed and then implemented; it was authored by teachers, one decision at a time, under conditions of uncertainty and without institutional support. The consequences are visible in the assessment pattern: teachers improvised verification routines, and the improvisations were of very uneven quality, ranging from a careful audit protocol (T9) to an unexamined practice of entering learner writing into a third-party system, which two teachers reported having adopted without considering whether it was permissible (Section 4.5). The pattern is not a failure of individual professionalism; it is the predictable result of asking practitioners to author a division of labour that has not been specified by anyone with the relevant knowledge. Institutional guidance that addresses only learner conduct leaves exactly the teacher's verification and data obligations undefined, which is where the consequential decisions lie.

5.4 What this means for the CFL-specific focus of AI research

The study also speaks to a general question about whether there is anything specifically Chinese about AI-mediated language education. Two of the four risk domains we identified — character form and meaning, and tone in connected speech — have no direct analogue in the teaching of a language written alphabetically and without lexical tone. Learners' dependence on character explanations as retrieval structures, and the mismatch between canonical tone descriptions and connected-speech realisations, generate pedagogical problems that a general framework for AI in language education cannot see. The claim that Chinese language education needs its own AI scholarship is therefore not a claim about the language's difficulty, and it is not a claim about the distinctiveness of foreign language teaching in general [19], but a claim about where the failures of an algorithmic mediator land. They land on the writing system and on the phonology of connected speech, and those are the places where teachers in this study stopped delegating.

5.5 What teacher development would need to look like if this account is right

If the operative knowledge is knowledge of delegation boundaries, then the design of teacher development follows from the object rather than from the tools. Four features would distinguish a programme built on this account from an adoption-oriented one.

First, it would begin from the syllabus rather than from the tool [20]. The productive unit of work is not "here are five things generative AI can do" but "here is where in this course an algorithmic judgement would mediate a learner's encoding of the language, and here is how we would know". A department working through its own syllabus, point by point, would produce a boundary map specific to its curriculum, and the map — not the tool list — is the durable output of the training.

Second, it would treat the teacher's own disciplinary knowledge as the resource being developed, not the obstacle being overcome. The teachers in this study who set the clearest boundaries were not those with the most technology experience but those who could explain what was wrong with a generated character explanation (Section 4.3). Development activity that strengthens that explanatory capacity — work on component functions, on the phonology of connected speech, on the register distinction between written and spoken Chinese — is, in this account, AI-related professional development, even though it involves no AI.

Third, it would be collective. The absence of shared practice documented in Section 4.5 is the single most remediable deficit in the dataset: in only one site did teachers share prompts or shared judgements about output quality, which meant that every boundary had to be rediscovered. A department-level boundary map, maintained in one place and revised when a tool changes, converts individual discoveries into institutional knowledge, in the way that collective, inquiry-based teacher learning has long been understood to work [21]. This also addresses the version-change problem directly: a shared map makes visible which practices depend on behaviour that a version change could alter.

Fourth, it would include the assessment and data-governance questions that institutional guidance currently omits. Teachers in this study improvised verification routines and, in two cases, transmitted learner work to third-party systems without having decided whether that was acceptable. Both are questions with determinate answers that could be specified in a short institutional document, and the absence of that document is not a gap in teachers' knowledge but in the system that expects them to act.

6. LIMITATIONS

Four limitations should be borne in mind when reading these findings. First, twelve cases cannot support claims about prevalence; the design supports claims about the range and reasoning of practice, and the distributional statements in Section 4 should be read accordingly. Second, all data concern one semester, which is too short to observe whether boundaries stabilise or drift; the three cases of mid-semester tool change hint that they are less stable than they appear. Third, we did not collect learner outcomes, so we cannot say whether teachers' boundaries produced better learning; the boundary is a knowledge claim, and its validity is a separate empirical question that research on assessment in CFL's scope is well placed to address. Fourth, our own position as CFL teachers shaped the coding in the two instances recorded in the reflexive journal, both of which involved reading a teacher's abandonment of a tool as a defensible linguistic judgement where the data also supported reading it as a workload adaptation.

7. CONCLUSION

This study asked what teachers of Chinese as a foreign language know and refuse when they work with generative AI. Across the twelve cases, teachers did not distribute their trust evenly across the Chinese curriculum. They drew a stable, well-reasoned boundary around the features that make Chinese distinctive, most sharply around character form and meaning and around tone in connected speech, and they justified that boundary by appeal to how learners store characters and how the target language actually sounds. Alongside this, they improvised verification practices for assessment that no institutional guidance covered, and they enacted their knowledge inside working conditions that constrained what was imaginable. We have argued that this pattern is best captured by describing an AI-specific pedagogical content knowledge for CFL as knowledge of delegation boundaries, and that teacher development should be organised around the syllabus, asking at each point where an algorithmic judgement may safely mediate learning, rather than around the tools themselves.

DECLARATIONS

Ethics approval. Approved by the research ethics committee of the lead institution and by participating institutions where required. All participants gave written informed consent, including consent for observation of teaching and collection of teaching artefacts.

Data governance. Interviews were pseudonymised at transcription; the pseudonym key was held separately. Learner names in teaching artefacts were redacted at collection. Quotations were checked with participants during member checking.

Generative AI disclosure. Generative AI was the object of this study. It was not used to transcribe, translate, code, or draft any part of this manuscript. Tool names and versions used by participating teachers are reported in the supplementary material.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors declare no competing interests.

Author contributions. All authors contributed to the conception and design of the study, the collection, analysis, and interpretation of data, and the drafting and critical revision of the manuscript, and approved the final version for submission.

Data availability. The interview transcripts, observation records, and teaching artefacts cannot be shared publicly because they contain information that could identify participants; de-identified excerpts are available from the corresponding author on reasonable request.

REFERENCES

- [1] Zawacki-Richter, O.; Marín, V. I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16* (1), 39. <https://doi.org/10.1186/s41239-019-0171-0>.
- [2] Barrot, J. S. Using ChatGPT for Second Language Writing: Pitfalls and Potentials. *Assess. Writ.* **2023**, *57*, 100745. <https://doi.org/10.1016/j.asw.2023.100745>.
- [3] Kohnke, L.; Moorhouse, B. L.; Zou, D. ChatGPT for Language Teaching and Learning. *RELC J.* **2023**, *54* (2), 537–550. <https://doi.org/10.1177/00336882231162868>.
- [4] Orton, J. Educating Chinese Language Teachers—Some Fundamentals. In *Teaching and Learning Chinese in Global Contexts: Multimodality and Literacy in the New Media Age*; Tsung, L., Cruickshank, K., Eds.; Continuum: London, 2011; pp 151–164. <https://doi.org/10.5040/9781474212373.ch-010>.
- [5] Mishra, P.; Koehler, M. J. Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teach. Coll. Rec.* **2006**, *108* (6), 1017–1054. <https://doi.org/10.1177/016146810610800610>.
- [6] Shulman, L. S. Those Who Understand: Knowledge Growth in Teaching. *Educ. Res.* **1986**, *15* (2), 4–14. <https://doi.org/10.3102/0013189X015002004>.
- [7] Shulman, L. S. Knowledge and Teaching: Foundations of the New Reform. *Harv. Educ. Rev.* **1987**, *57* (1), 1–23. <https://doi.org/10.17763/haer.57.1.j463w79r56455411>.
- [8] Tsai, C.-C.; Chai, C. S. The “Third”-Order Barrier for Technology-Integration Instruction: Implications for Teacher Education. *Australas. J. Educ. Technol.* **2012**, *28* (6), 1057–1070. <https://doi.org/10.14742/ajet.810>.
- [9] Holmes, W.; Porayska-Pomsta, K.; Holstein, K.; Sutherland, E.; Baker, T.; Buckingham Shum, S.; Santos, O. C.; Rodrigo, M. M. T.; Cukurova, M.; Bittencourt, I. I.; Koedinger, K. R. Ethics of AI in Education: Towards a Community-Wide Framework. *Int. J. Artif. Intell. Educ.* **2022**, *32* (3), 504–526. <https://doi.org/10.1007/s40593-021-00239-1>.
- [10] Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, E.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>.
- [11] Godwin-Jones, R. Emerging Spaces for Language Learning: AI Bots, Ambient Intelligence, and the Metaverse. *Lang. Learn. Technol.* **2023**, *27* (2), 6–27. <https://doi.org/10.64152/10125/73501>.
- [12] Hafner, C. A.; Chik, A.; Jones, R. H. Digital Literacies and Language Learning. *Lang. Learn. Technol.* **2015**, *19* (3), 1–7. <https://doi.org/10.64152/10125/44426>.
- [13] Ke, C. Effects of Language Background on the Learning of Chinese Characters among Foreign Language Students. *Foreign Lang. Ann.* **1998**, *31* (1), 91–102. <https://doi.org/10.1111/j.1944-9720.1998.tb01335.x>.
- [14] Shen, H. H.; Ke, C. Radical Awareness and Word Acquisition among Nonnative Learners of Chinese. *Mod. Lang. J.* **2007**, *91* (1), 97–111. <https://doi.org/10.1111/j.1540-4781.2007.00511.x>.
- [15] Xing, J. Z. *Teaching and Learning Chinese as a Foreign Language: A Pedagogical Grammar*; Hong Kong University Press: Hong Kong, 2006. <https://doi.org/10.5790/hongkong/9789622097629.001.0001>.
- [16] Borg, S. Teacher Cognition in Language Teaching: A Review of Research on What Language Teachers Think, Know, Believe, and Do. *Lang. Teach.* **2003**, *36* (2), 81–109. <https://doi.org/10.1017/S0261444803001903>.
- [17] Borg, S. *Teacher Cognition and Language Education: Research and Practice*; Bloomsbury Academic: London, 2015. <https://doi.org/10.5040/9781474219983>.
- [18] Parasuraman, R.; Riley, V. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* **1997**, *39* (2), 230–253. <https://doi.org/10.1518/001872097778543886>.
- [19] Kramsch, C. Teaching Foreign Languages in an Era of Globalization: Introduction. *Mod. Lang. J.* **2014**, *98* (1), 296–311. <https://doi.org/10.1111/j.1540-4781.2014.12057.x>.
- [20] Hubbard, P. CALL and the Future of Language Teacher Education. *CALICO J.* **2008**, *25* (2), 175–188. <https://doi.org/10.1558/cj.v25i2.175-188>.
- [21] Cochran-Smith, M.; Lytle, S. L. Relationships of Knowledge and Practice: Teacher Learning in Communities. *Rev. Res. Educ.* **1999**, *24*, 249–305. <https://doi.org/10.2307/1167272>.