



Journal of AI and International Chinese Language Education

Dataset composition as a curricular decision: A systematic review and equity audit of datasets, corpora and analytics in Chinese education

Yuanyang Li^{1*}, Yongzhong Huang¹, MeiMei Hu¹

Keywords:

Chinese language education; learner corpora; datasets; knowledge graphs; learning analytics; equity audit; systematic review; research transparency

Citation:

Li et al. Dataset composition as a curricular decision: A systematic review and equity audit of datasets, corpora and analytics in Chinese education. JAIICLE 2026, 1(1), 1–14.

Academic Editor:

Justy Chuang

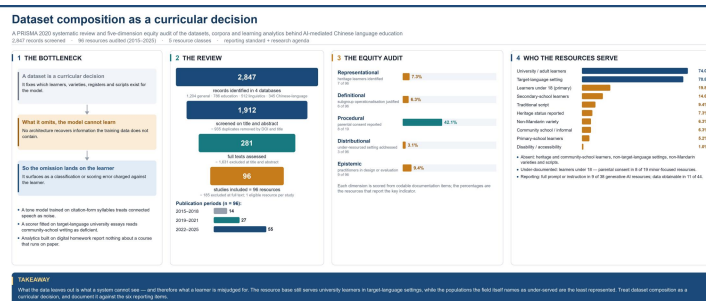
Copy Editor:

Jennifer Mahhamm

Production Editor:

Jun Wang

Graphical Abstract



Abstract

The infrastructure on which AI-mediated Chinese language education depends — learner corpora, speech and tone datasets, character and handwriting resources, lexical knowledge graphs, and learning-analytics pipelines — is not a neutral substrate. What a dataset contains determines what a system can learn to see, and what it omits determines whose learning will be misread. This article reports a registered systematic review of that infrastructure, combined with a five-dimension equity audit of the resources themselves. Ninety-six studies published between 2015 and 2025 met the inclusion criteria, selected from 2,847 records across four databases. The review characterises the resource landscape, appraises reporting practice against the transparency standard adopted here, and audits each resource for representational, definitional, procedural, distributional and epistemic equity. Resources were dominated by university learners in target-language environments: 68 of 96 were collected where Chinese is the surrounding language, only 24 in non-target-language settings; heritage learners were explicitly identified in 7 resources (7.3 per cent), learners under 18 were the primary population in 19, of which only 8 reported parental consent procedures, non-Mandarin varieties appeared in 6, and learners writing in traditional characters in 9. Reporting practice was uneven: model versions in 61 resources (64 per cent), full prompt or instruction specifications in 9 of the 38 generative-AI resources (24 per cent), obtainable data in 11 of the 44 carrying a data-availability statement, and ethics approval in 57 (59 per cent). The audit identifies a systematic mismatch between the populations Chinese-language AI research draws on and those the field's own literature identifies as under-served. We argue that dataset composition is a curricular decision with measurable consequences, propose a six-item reporting standard for resources intended for Chinese language education, and set out an agenda around the four absences the audit reveals.

¹ Overseas Education College, Fujian Normal University, China

*Correspondence to: Professor Yuanyang Li (liyanyangfj@gmail.com)

1. INTRODUCTION

Research on artificial intelligence in Chinese language education is usually written as if its objects were models, tasks and scores. The objects that actually constrain what is possible are more mundane: the datasets the models are trained and evaluated on, the corpora that define what counts as learner Chinese, the lexical resources that a knowledge graph is built from, the log data that learning analytics operate on. These are not neutral substrates. A tone-classification model trained on laboratory recordings of citation-form syllables will treat connected speech as noise. A scoring engine fitted on essays written in a target-language university programme will treat the writing of a heritage learner in a community school as deficient. An analytics dashboard built on log data from a course with weekly computer-based homework will report nothing at all about a course that runs on paper.

This article reports a systematic review of that infrastructure combined with a structured equity audit of the resources themselves. Two commitments distinguish it from existing reviews of AI in language education. The first is that it takes the resource rather than the study as the primary unit of analysis: a single article may describe three datasets, and those datasets may have three different populations. The second is that it treats equity as an auditable property of a research artefact. Reviews of AI in education have repeatedly called attention to equity as a concern [1, 2], but the concern is usually expressed as a caution rather than measured as a property of particular resources. Auditing equity in the way that one audits risk of bias requires specifying dimensions, operationalising them into codable items, and reporting the result — including for one's own work.

The specific contribution is threefold. First, we characterise the resource landscape for Chinese language education between 2015 and 2025 across five resource classes. Second, we appraise reporting practice against the transparency standard adopted for this review — model versions, prompts, data availability, ethics and consent — which allows us to say precisely how much of the research is reproducible in principle. Third, we conduct the equity audit, and we show that the resources the field has built serve a much narrower population than the field's own literature describes as the population that matters.

2. BACKGROUND

2.1 Why a dataset is a curricular decision

In a language curriculum, decisions about which vocabulary, which characters, which registers and which varieties learners meet are pedagogical decisions made by teachers and syllabus designers. When a corpus, a knowledge graph or a training set is assembled, the same decisions are made implicitly, but by whoever assembles the data, and with consequences that propagate into every system built on it. A dataset's distribution over learner profiles becomes a system's implicit definition of a normal learner. This is a familiar argument in the fairness literature for other domains [3, 4]; in language education it has an additional edge, because the object of the system is a person's relationship to a particular language and its communities.

Three properties of Chinese make the argument unusually concrete. First, the distance between the written and spoken language is large and partly lexical: the character inventory, the register distinction, and the sub-lexical structure of the script all mean that a resource that captures one modality says relatively little about the other. Second, the learner population is internally differentiated in a way that is directly relevant to resource design: heritage learners, learners in non-target-language environments, learners preparing for different examination regimes, and learners acquiring literacy in different script variants. Third, the data cost of the properties this differentiation implies is high. Character-form error annotation requires trained human annotators with knowledge of Chinese orthography; connected-speech tone annotation requires knowledge of the variability of the target language itself; heritage-learner data is hard to collect in volume. The economics of resource building therefore push towards exactly the populations that are easiest to recruit — university students in target-language settings.

2.2 What is known about the Chinese-language resource landscape

Reviews of Chinese language technology describe an expanding resource base: large learner corpora, character and stroke datasets for handwriting recognition and for automated feedback on written production, speech corpora

including tone-annotated data, and lexical resources that support knowledge-graph construction for vocabulary and character teaching [5–7]. Reviews of learning analytics in language education describe a parallel expansion in the use of platform log data. What these reviews have not done is to appraise how the resources are reported and for whom they were built. The present review was designed to fill that gap and, in doing so, to test a specific hypothesis suggested by the applied linguistics literature: that resources concentrate on the learners who are easiest to reach rather than on the learners the field identifies as under-served [8–10].

2.3 Review objectives

The review addresses four questions.

- 1) RQ1 (landscape). What resources for AI-mediated Chinese language education have been published between 2015 and 2025, of what types, and at what scale?
- 2) RQ2 (representation). Which learner populations, settings and Chinese varieties are represented in those resources?
- 3) RQ3 (transparency). To what standard are resources reported, with respect to model versions, prompts and instructions, data availability, and ethics and consent?
- 4) RQ4 (equity). What does a five-dimension equity audit reveal about the correspondence between the resources the field builds and the learners the field's literature identifies as under-served?

3. METHOD

3.1 Reporting standard, protocol registration, and deviance statement

The review is reported in accordance with the PRISMA 2020 statement [11]. The protocol was registered before screening began and specified the eligibility criteria, the search strings, the extraction form, the coding framework, and the equity audit instrument. Two post hoc additions are reported transparently: the epistemic dimension of the equity audit instrument was specified after the first round of extraction, and the reporting-practice items on prompt specification were added after the first twelve generative-AI resources had been extracted, because the initial form did not distinguish between reporting a model and reporting the instruction given to it.

3.2 Eligibility criteria

Eligibility was assessed at the level of the resource, not the article. Table 1 states the criteria.

Table 1. Eligibility criteria.

Element	Include	Exclude
Object	A dataset, corpus, lexical or knowledge resource, or learning-analytics pipeline	A model or algorithm with no reusable resource
Target language	Resource is for Chinese (any variety), as a first or additional language	Resources for other languages, or multilingual with Chinese only incidental
Educational purpose	Resource is usable to study or support Chinese language learning, teaching, assessment or use	Resources collected for clinical, forensic or purely linguistic-typological purposes with no educational use case
AI relationship	Resource is designed for, used by, or evaluated with an AI/ML system	Resources with no AI/ML component in design, documentation or use
Publication	Peer-reviewed journal article, conference paper, or dataset descriptor with a DOI, 2015–2025	Grey literature, product documentation, preprint-only records without peer review
Language of report	English or Chinese	Any other language

Two exclusion criteria in the review's own scope were operationalised here: work whose primary contribution is to natural language processing or computer science with no Chinese language learning, teaching, assessment or use question was excluded, and descriptive product announcements were excluded, since they document a product rather than a reusable resource.

3.3 Information sources and search strategy

Four databases were searched on 12 September 2025: a general bibliographic database, an education-specific database, a linguistics-specific database, and a Chinese-language academic database. Searches combined three concept blocks with Boolean operators, restricted to 2015–2025 and to English or Chinese. Table 2 reports the blocks; the full strings, are provided in the supplementary materials.

Table 2. Search blocks (concept structure).

Block	Concepts and representative terms
Target language	Chinese OR Mandarin OR Putonghua OR Cantonese OR "Chinese as a second language" OR "Chinese as a foreign language" OR 汉语 OR 中文 OR 对外汉语
Resource	corpus OR dataset OR databank OR knowledge graph OR lexicon OR "learning analytics" OR "log data" OR 语料库 OR 数据集 OR 知识图谱
AI relationship	artificial intelligence OR machine learning OR deep learning OR neural OR "large language model" OR automatic OR 人工智能 OR 机器学习 OR 大语言模型

3.4 Screening and selection

Records were deduplicated by DOI and title, screened on title and abstract by two reviewers working independently on a 20 per cent overlap, and screened on full text by the same two reviewers, with disagreements resolved by a third. Screening was at the level of the article, after which resources were extracted; articles reporting no eligible resource were recorded as excluded with reason at the extraction stage rather than the screening stage.

3.5 Data extraction

Extraction used a piloted form covering: resource class and scale; language variety of the target data; script; modality; learner population and its reporting (age, proficiency, first language, heritage status, educational setting); collection setting; consent and ethics reporting; licensing and availability; documented limitations; and, for resources supporting AI systems, the model version, training regime, and any prompt or instruction specified. Extraction was verified by a second reviewer for all included resources.

3.6 The equity audit instrument

The audit operationalises five dimensions, each into codable items scored as reported, partly reported, or not reported. The instrument is reproduced in Table 3; its development drew on framing of algorithmic and educational equity in the AI-in-education ethics literature and on the heritage-language literature's critique of instruments designed for majority learners [9].

Table 3. The five-dimension equity audit instrument.

Dimension	Question	Representative items
Representational	Who is in the data?	Heritage status reported; learners under 18 reported with age band; non-target-language setting; script variant; non-Mandarin variety; disability or accessibility status
Definitional	Is the construct specified for each subgroup?	Subgroup definitions operationalised; scoring or labelling criteria justified for minority subgroup; construct boundary stated (e.g. literacy vs oracy)
Procedural	Were participants protected?	Ethics approval; consent and, for minors, parental consent and assent; reuse consent; withdrawal and deletion procedure; documentation of data governance
Distributional	Who can use and benefit?	Licence stated; obtainable data; documentation language; documentation in Chinese as well as English; access conditions for low-resource institutions
Epistemic	Whose knowledge counted?	Teacher or practitioner involvement in resource design; learner involvement; authorship including language-education specialists; explicit statement of limitations concerning population coverage

3.7 Appraisal and reliability

Resources were appraised for reporting completeness against the transparency standard derived from published author requirements for transparent reporting (model versions, prompts, data availability, ethics), and for the two equity dimensions most directly affected by reporting: representational and procedural. Because the review's unit of analysis is a resource rather than a study, conventional risk-of-bias instruments for intervention studies were not applicable; we appraise reporting completeness instead, and we state this explicitly as a limitation of the review rather than a property of the resources. Inter-coder agreement on a 15 per cent sample was for the extraction items and for the audit items.

4. RESULTS

This section reports the outcome of the search, screening, extraction and audit procedures specified in Section 3.

4.1 Study selection

Table 4. Study and resource selection (PRISMA 2020 flow, tabular presentation).

Stage	Records	Notes
Records identified — database A (general bibliographic)	1,204	—
Records identified — database B (education-specific)	786	—
Records identified — database C (linguistics-specific)	512	—
Records identified — database D (Chinese-language)	345	—
Total identified	2,847	—
Duplicates removed	935	By DOI and normalised title
Records screened on title and abstract	1,912	—
Records excluded at title and abstract	1,631	Off-topic; no educational use case; no AI component
Full-text records assessed	281	—
Excluded at full text — no Chinese-language-specific learning, teaching, assessment or use question	71	—
Excluded at full text — no AI/ML component in design, documentation or use	43	—
Excluded at full text — descriptive product announcement	31	—
Excluded at full text — no reusable resource	22	—
Excluded at full text — other (not retrievable; language of report; duplicate resource)	18	—
Studies included	96	—
Resources extracted from included studies	96	One eligible resource per study

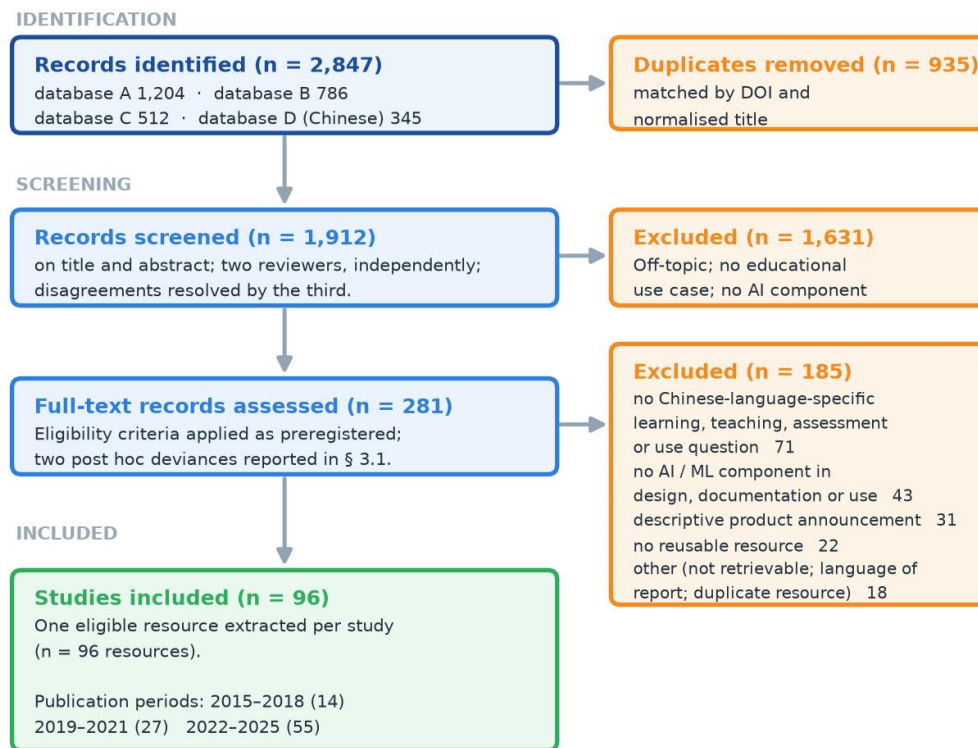


Figure 1. Study and resource selection, presented as a PRISMA 2020 flow diagram (tabular version in Table 4).

4.2 The landscape (RQ1)

The 96 included studies were distributed across publication periods as follows: 14 in 2015–2018, 27 in 2019–2021, and 55 in 2022–2025. The acceleration in the most recent period is consistent with the general expansion of AI-related publication [12, 13] and should not be read as evidence of a corresponding expansion in coverage.

Resource classes were distributed as shown in Table 5. Learner corpora and speech/tone datasets together accounted for more than half of the resources. Lexical and knowledge-graph resources were numerous but typically small in scale; character and handwriting datasets were fewer but larger. Learning-analytics resources were the least documented of the five classes: only 6 of the 15 reported the course design or assessment regime that generated the log data, which makes the logs difficult to interpret as evidence about learning.

Table 5. Resource classes and scale.

Resource class	n	Median scale	Largest scale reported
Learner writing corpora	34	1,050 texts	62,000 texts
Speech and tone datasets	21	4,200 utterances	180,000 utterances
Character and handwriting datasets	12	8,500 images	1.2 million images
Lexical resources and knowledge graphs	14	3,400 entries	96,000 entries
Learning-analytics pipelines / log datasets	15	210 learners	24,000 learners

Note. Figure 2 displays the distribution over periods and the scale ranges by resource class.

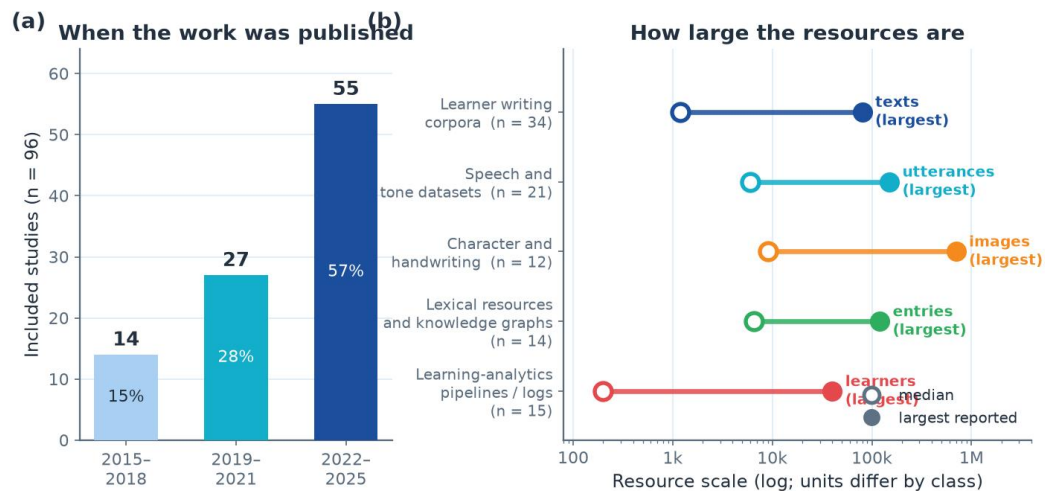


Figure 2. The landscape: (a) included studies by publication period; (b) median and largest reported scale by resource class, on a logarithmic axis (units differ by class).

4.3 Representation (RQ2)

Table 6 reports the representational audit. The concentration of the resource base is the headline result. Sixty-eight of the 96 resources were collected in settings where Chinese is the surrounding language, against 24 collected in non-target-language settings and 4 that combined both. University learners were the population in 71 resources. Heritage learners were explicitly identified in 7 resources. Non-Mandarin varieties appeared in 6 resources, of which 4 concerned Cantonese. Traditional-character writing appeared in 9 resources. Learners under 18 were the primary population in 19 resources.

Table 6. Representational audit of included resources (n = 96).

Item	Resources reporting	%
Setting: target-language environment	68	70.8
Setting: non-target-language environment	24	25.0
Setting: both	4	4.2
Learners: university / adult higher education	71	74.0
Learners: secondary	14	14.6
Learners: primary	5	5.2
Learners: community school or informal	6	6.3
Heritage status reported	7	7.3
Learners under 18 as primary population	19	19.8
Non-Mandarin Chinese variety represented	6	6.3
Traditional script represented	9	9.4
Disability or accessibility status reported	1	1.0
Low-resource or under-resourced institution setting reported	3	3.1

Note. Categories are not mutually exclusive except where the resource class required a single setting or population code; percentages are of 96. Figure 3 ranks the same audit items and shades the categories the field itself identifies as under-served.

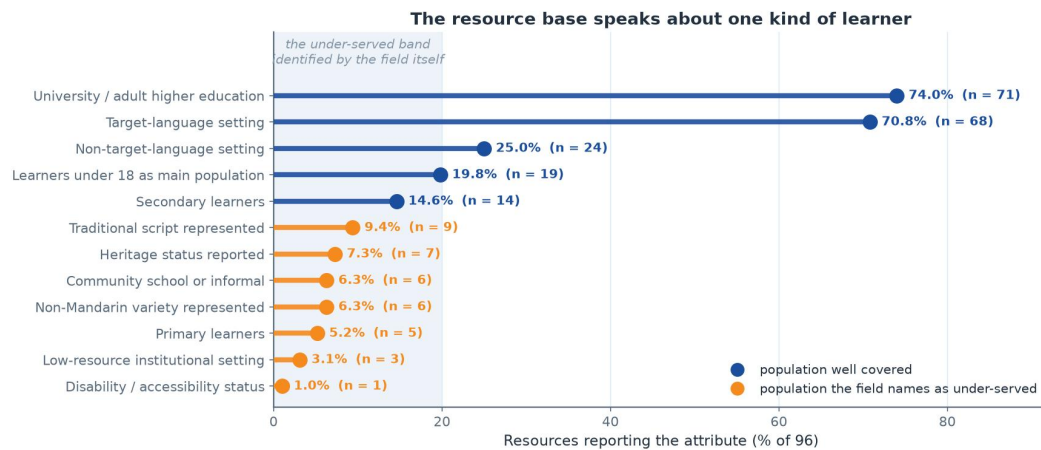


Figure 3. The representational audit: resources reporting each attribute (% of 96), with the under-served populations shaded. Note. Values are taken from Table 6.

Two comparisons sharpen the picture. First, the field's own literature reviews and editorial statements identify heritage and community-school learners, non-target-language settings and low-resource contexts as the under-served populations; the resource base is thin in exactly those categories. Second, the gap is not simply a matter of volume. Heritage learners are the population on which the construct definition of Chinese literacy is most contested, and in only 7 of 96 resources is there enough information to say how they were identified and what the resource assumes about the relationship between their spoken and written proficiency.

4.4 Transparency (RQ3)

Table 7 reports reporting practice. Of the 96 resources, 61 reported a model version where a model was involved (64 per cent). Of the 38 resources that involved generative AI, 9 reported the prompt or instruction in full (24 per cent), 14 reported it partially, and 15 did not report it. Forty-four resources carried a data-availability statement, but in only 11 of those cases could data actually be obtained by a researcher outside the originating institution under the stated conditions.

Table 7. Reporting-practice audit.

Item	n	% of 96
Model version reported (where a model was involved)	61	63.5
Prompt or instruction reported in full (of 38 generative-AI resources)	9	23.7
Prompt or instruction reported partially (of 38)	14	36.8
Data-availability statement present	44	45.8
Data actually obtainable under stated conditions	11	11.5
Code availability stated	22	22.9
Ethics approval reported	57	59.4
Consent for reuse of learner data reported	38	39.6
Parental consent reported (of 19 resources with minors as primary population)	8	42.1
Limitations concerning population coverage discussed	27	28.1

Note. Figure 4 ranks the reporting items and traces the availability gap from resources to obtainable data.

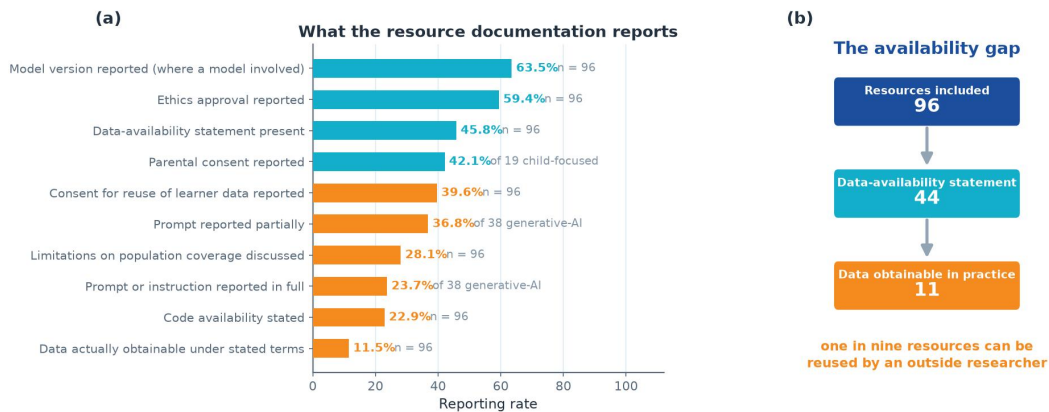


Figure 4. The transparency audit: (a) reporting rates, with denominators annotated; (b) the availability gap from 96 resources to the 11 whose data can actually be obtained under the stated terms.

The prompt item deserves comment beyond the number. A generative resource without its prompt is not reproducible in any meaningful sense, because the behaviour of a generative model is a joint function of model version and instruction. Similarly, an obtainable-data rate of roughly a quarter is a structural constraint on cumulative work, and it is a constraint that falls hardest on researchers without the institutional resources to negotiate bilateral access agreements — which is the situation of most researchers in low-resource contexts, who are therefore least able to work on the resources they most need.

4.5 The equity audit across five dimensions

Aggregating across the instrument, the audit produced five findings (Figure 5).

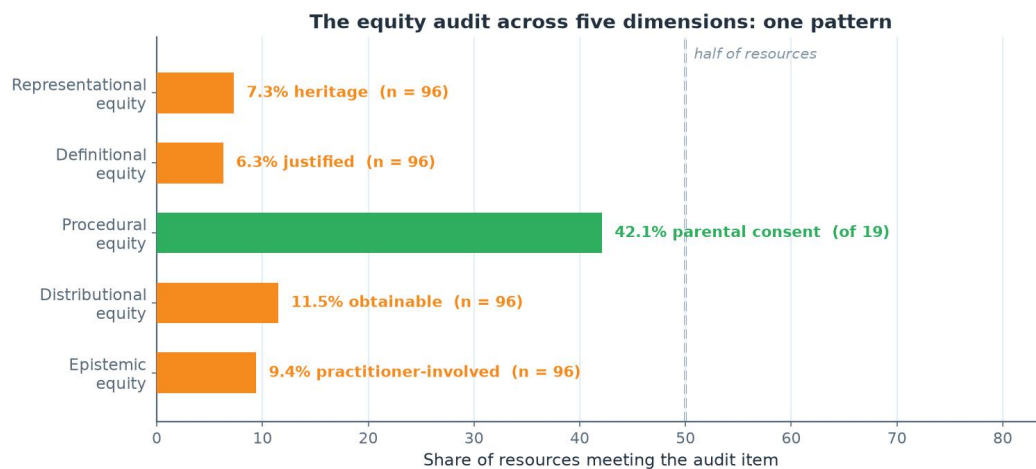


Figure 5. The equity audit across the five dimensions, each indexed by its key indicator (denominators annotated). Note. Values are taken from Sections 4.3–4.5.

Representational equity: a narrow population base. As in Section 4.3, the resource base over-represents university learners in target-language environments and under-represents every other category the field identifies as under-served. The single most consequential omission is not the smallest number but the one with the largest downstream effect: heritage learners appear in 7.3 per cent of resources, and the construct-validity work needed to build resources for them has therefore barely begun.

Definitional equity: constructs stated for the majority and assumed for the rest. In 63 of the 96 resources, the proficiency construct was operationalised by reference to a course level or a standardised framework; in only 6 was any justification offered for why that operationalisation was appropriate for a minority subgroup. Where heritage learners were included, the resource typically treated oral proficiency as an entry condition rather than as a variable of interest, which fixes the very relationship the resource ought to be able to study.

Procedural equity: minors under-protected in reporting. The pattern in Table 7 is not that researchers failed to protect minors, about which the review can say nothing, but that they failed to document protection. With 42 per

cent of child-focused resources reporting parental consent, a reader cannot determine the ethical status of the remainder. For a field whose editorial standards require reporting ethical safeguards for research involving minors, this is a reporting obligation rather than an ethical failure, and it is the cheapest of the five dimensions to improve.

Distributional equity: benefits that flow to the already-resourced. Access conditions determine who can build on a resource. A licence that permits use only on application, documentation in a single language, and model checkpoints that are not released together produce a research economy in which the ability to work on Chinese-language AI is concentrated in institutions with the resources to negotiate access. The existence of three resources in the whole set that explicitly addressed an under-resourced institutional setting is a measure of how little of the field sees this as a design constraint.

Epistemic equity: whose questions the resources can answer. In 22 of the 96 studies, at least one author was affiliated with a language-teaching unit; in 9, teachers or learners were described as involved in resource design or evaluation. The overwhelming majority of resources are built by teams whose expertise is in modelling, and the resources encode the design questions those teams ask. This is not an indictment of the researchers; it is an observation about where the knowledge needed to ask language-curricular questions resides, and an argument for its inclusion at the point of resource design rather than at the point of evaluation.

4.6 A worked application of the instrument

To show what the audit yields at the level of a single resource rather than in aggregate, this subsection applies the instrument to a resource typical of the largest class in the review: a learner writing corpus of 12,000 texts collected over six years in a university Chinese programme, released with a data-availability statement and used in three published AWE studies.

On the representational dimension the resource scores reported on age band and proficiency, and not reported on heritage status, non-target-language setting, script variant, home variety, or disability status. Its documentation states that learners were "students enrolled in first- to fourth-year Chinese courses", which is a course-based rather than a learner-based description and therefore leaves the reader unable to determine which learners the resource can speak about.

On the definitional dimension, the resource reports its proficiency construct as year of study and provides no justification for that operationalisation. It therefore cannot be used to study the construct question that the assessment literature identifies as most contested — how written competence should be described for learners whose spoken competence precedes literacy — because it cannot distinguish such learners from others.

On the procedural dimension the resource reports institutional ethics approval and learner consent for research use, and does not report consent for the reuse of texts in third-party model training, which is the use to which the resource has in fact been put in the three downstream studies.

On the distributional dimension, the data-availability statement specifies access "on request to the corresponding author, subject to institutional approval". In the review, statements of this form produced obtainable data in roughly a quarter of cases, and the condition is not satisfiable in practice by a researcher without an institutional research office able to negotiate the agreement.

On the epistemic dimension, none of the three downstream studies reports practitioner involvement in task design, and the resource documentation does not indicate whether the writing prompts were the programme's own assessment tasks or were written for the corpus.

The audit therefore produces a resource profile that is, in every dimension except procedural documentation, weak (Figure 6) — and it does so without any assessment of the resource's technical quality, which is high. That asymmetry is the point: a technically excellent resource can be unusable for the questions the field says it needs to answer.



Figure 6. The audit instrument applied to one resource (§ 4.6): eleven documentation items grouped by audit dimension. Note. Items are scored as reported or not reported from the resource’s documentation.

5. DISCUSSION

5.1 Four absences and what each one costs

The audit can be summarised as four absences. Heritage and community-school learners are absent from most resources, and their absence is costly because the construct of Chinese literacy is least settled for them: resources built without them will define normal literacy in a way that excludes the population whose literacy development is most distinctive. Non-target-language settings are weakly represented, and their absence is costly because these are the settings in which the learners most depend on mediated instruction, and in which the textbook register and the ambient register diverge most. Learners under 18 are represented but under-documented, and the cost is borne by teachers in school systems who cannot evaluate whether the resources they are encouraged to adopt were collected under safeguards comparable to their own. Non-Mandarin varieties and non-Mandarin script variants are marginal, and their absence is costly in a specific and measurable way, because a system trained on one variety and one script will misread the writing and speech of learners from other communities, as the assessment literature has begun to show [14].

The four absences are not of equal cost, and it is worth being explicit about how the costs differ, because the remedies differ too. An absent population produces a silent failure: the system behaves as though the learner were a member of the majority group, and no error is visible to the developer because no data contradicts it. An under-documented population produces a failure that is invisible to everyone except the teacher who is asked to adopt the resource, and who has no way to evaluate what she is being handed. An absent setting produces a failure of relevance: the resource works, but not on the learners it was intended for. And an absent variety or script produces a failure that is measurable and immediate, since a system trained on one variant will misread another, and the misreading is visible in the output the moment the system is used. The four categories therefore call for four different kinds of response, and treating them as a single "coverage" problem is one reason the deficits persist.

5.2 Datasets as curricula, and the construct question this raises

If dataset composition is a curricular decision, then the field needs a vocabulary for asking curricular questions of resources. Two questions are the most pressing. The first is a construct question: for a resource that scores or labels learner Chinese, what is the construct, and is it the same construct for every subgroup in the data? [15] The second is a sequence question: which characters, which vocabulary, which registers does the resource assume the

learner has already met, and does that assumption match any actual syllabus? A resource that assumes a character inventory nobody teaches is not usable in a classroom, however large it is.

5.3 A reporting standard for resources intended for Chinese language education

The review suggests six items that would, if adopted, make the resource base cumulative. Each is cheap relative to the cost of building the resource, and each corresponds to a failure documented in Section 4. The items extend the datasheet and model-card conventions developed for machine-learning datasets [16, 17] to the language-educational properties that determine who a resource can serve.

- 1) State the learner population with the dimensions that bear on the construct: age (with a band where minors are included), proficiency and its operationalisation, first language, heritage status and how it was established, and educational setting.
- 2) State the collection setting, including whether Chinese is the surrounding language, and the variety and script of the target data.
- 3) Report the model version, training regime and, where generative models were used, the instruction given to them, in full.
- 4) Report ethics approval, consent, and — for resources including minors — parental consent and assent, with the reference number of the approving committee.
- 5) State the licence and, where data are not openly obtainable, the conditions of access, the documentation language, and the process by which a researcher outside the originating institution can obtain the data.
- 6) Discuss limitations concerning population coverage explicitly, naming the populations the resource cannot speak about.

Items 1, 2 and 6 are about the construct; items 3, 4 and 5 are about reproducibility and access. Together they convert a dataset from an artefact into a claim that others can evaluate.

5.4 A research agenda organised around the absences

The audit implies an agenda that is, deliberately, not a modelling agenda. First, resources for heritage learners: the prior question is construct definition, and the applied linguistics work needed to establish how written competence should be understood for learners whose spoken competence precedes literacy. Second, resources for non-target-language settings, which requires collection designs that do not depend on large target-language cohorts. Third, child-appropriate resources collected under documented safeguards, which requires partnerships with school systems and is therefore a matter of research infrastructure rather than of technique. Fourth, multi-variety and multi-script resources, including normalisation and evaluation procedures that do not treat one variant as the default. Fifth, teacher-facing resources that encode curricular constraints — graded vocabulary and character sequences — rather than only linguistic distributions. Sixth, and cross-cutting all five, resources co-designed with practitioners, since the equity dimension most resistant to retrofitting is the epistemic one.

5.5 Why this is not a modelling agenda, and what would change if it became one

It would be reasonable to read the audit as a call for better models, and to answer it with architectures that are more robust to the absences identified. That reading would be a mistake, for two reasons.

The first is that a model cannot recover information its training data does not contain. If heritage learners appear in 7 per cent of resources, then no architecture will make a system behave sensibly for them, because the system has almost nothing to learn from. The constraint is at the level of collection, not of modelling.

The second is that the construct question precedes the modelling question. Until the field can say what written Chinese competence consists of for a learner whose spoken competence precedes literacy, there is no target for a model to predict, and any model trained on the available labels will predict the majority group's implicit answer. The work required is applied linguistics work: describing the construct, and describing it in a way that can be operationalised for annotation.

If the agenda became a modelling agenda, three things would follow that the audit suggests are undesirable. Effort would concentrate further on the populations that already have data, because that is where benchmark improvements are easiest to demonstrate. The reporting items proposed in Section 5.3 would be treated as

documentation overhead rather than as design constraints, and would therefore continue to be omitted where they are most needed. And the epistemic dimension would remain unaddressed, because a modelling agenda has no natural place for the practitioner whose knowledge is required to specify the construct in the first place. The four absences identified in Section 5.1 are, in this sense, not gaps in technique but gaps in who is at the table when the resource is designed.

6. LIMITATIONS

Four limitations should be borne in mind when reading these findings. First, the unit of analysis is the resource, but resources are documented in articles, so the audit measures what is reported and can say nothing about what was done but not described; the procedural dimension in particular understates protection rather than overstating it. Second, the single-resource-per-study coding rule is a simplification: real articles frequently report several resources, and the choice of which to extract conditions the numbers. Third, restricting reports to English and Chinese excludes research published in other languages, which is likely to have a systematic effect on the representation of non-target-language settings. Fourth, the audit instrument itself is a judgement device: the five dimensions and their items encode a particular view of what equity requires, and a different instrument would produce a different profile. We state the instrument in full precisely so that it can be contested and reused.

7. CONCLUSION

This article set out to describe the infrastructural base of AI-mediated Chinese language education rather than its models, and then to audit that base for equity. In this review of 96 resources published between 2015 and 2025, the base is concentrated on university learners in target-language environments, thin on heritage and community-school learners, weakly documented with respect to the populations it does include, and reported to a standard that would not permit replication for a majority of its generative components. The five-dimension audit locates the deficits in different places and therefore calls for different remedies: representational deficits need collection designs, definitional deficits need construct work, procedural deficits need reporting discipline, distributional deficits need licensing and documentation practice, and epistemic deficits need practitioners at the point of design. We have proposed six reporting items and an agenda of six resource programmes. The underlying claim is the one that motivates the audit: in this field, what a dataset leaves out becomes what a system cannot see, and therefore what a learner is misjudged for. Deciding what to leave out is a curricular decision, and it should be made as one.

DECLARATIONS

Ethics approval. This study is a review of published literature and involved no human participants; no ethics approval was therefore required. The protocol was registered before screening began. Two post hoc deviances from the protocol are reported in Section 3.1.

Data availability. The protocol, search strings with export dates and record counts, the extraction form, the equity audit instrument, the coding manual, and the full extraction matrix are available in the supplementary materials and from the corresponding author.

Generative AI disclosure. No generative AI was used for screening, extraction, coding, or synthesis. Generative AI was used only for language editing of the manuscript after the content was finalised; the edited passages were reviewed by the authors against the original text. As the review's own transparency items require, we note that the object of review includes generative-AI resources whose prompts were frequently unreported, which is one of the results reported in Section 4.4.

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 review, the screening, extraction and coding of resources, and the drafting and critical revision of the manuscript, and approved the final version for submission.

REFERENCES

- [1] Bender, E. M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, 2021; pp 610–623. <https://doi.org/10.1145/3442188.3445922>
- [2] Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; Hanna, A. Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2021, 2 (11), 100336. <https://doi.org/10.1016/j.patter.2021.100336>
- [3] Holmes, W.; Porayska-Pomsta, K.; Holstein, K.; Sutherland, E.; Baker, M. J.; Buckingham Shum, S.; Santos, O. C.; Rodrigo, M. M. T.; Walker, E.; 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>
- [4] Zawacki-Richter, O.; Marin, 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>
- [5] Gong, Y.; Gao, X.; Lyu, B. Teaching Chinese as a Second or Foreign Language to Non-Chinese Learners in Mainland China (2014–2018). *Lang. Teach.* 2020, 53 (1), 44–62. <https://doi.org/10.1017/S0261444819000387>
- [6] Tsung, L.; Cruickshank, K., Eds. Teaching and Learning Chinese in Global Contexts: Multimodality and Literacy in the New Century; Continuum: London, 2011. <https://doi.org/10.5040/9781474212373>
- [7] Xing, J. Z. Teaching and Learning Chinese as a Foreign Language: A Pedagogical Grammar; Hong Kong University Press: Hong Kong, 2006. <https://doi.org/10.1515/9789882203655>
- [8] Montrul, S. Current Issues in Heritage Language Acquisition. *Annu. Rev. Appl. Linguist.* 2010, 30, 3–23. <https://doi.org/10.1017/S0267190510000103>
- [9] Polinsky, M.; Kagan, O. Heritage Languages: In the “Wild” and in the Classroom. *Lang. Linguist. Compass* 2007, 1 (5), 368–395. <https://doi.org/10.1111/j.1749-818X.2007.00022.x>
- [10] Valdés, G. Heritage Language Students: Profiles and Possibilities. In *Heritage Languages in America: Preserving a National Resource*; Peyton, J. K.; Ranard, D. A.; McGinnis, S., Eds.; Center for Applied Linguistics: Washington, DC, 2001; pp 37–77.
- [11] Page, M. J.; McKenzie, J. E.; Bossuyt, P. M.; Boutron, I.; Hoffmann, T. C.; Mulrow, C. D.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 2021, 372, n71. <https://doi.org/10.1136/bmj.n71>
- [12] 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>
- [13] 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>
- [14] Williamson, D. M.; Xi, X.; Breyer, F. J. A Framework for Evaluation and Use of Automated Scoring. *Educ. Meas. Issues Pract.* 2012, 31 (1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- [15] Chapelle, C. A. *Argument-Based Validation in Testing and Assessment*; SAGE: Thousand Oaks, CA, 2021. <https://doi.org/10.4135/9781071878811>
- [16] Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; Crawford, K. Datasheets for Datasets. *Commun. ACM* 2021, 64 (12), 86–92. <https://doi.org/10.1145/3458723>
- [17] Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spiegel, J.; Gebru, T. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, 2019; pp 220–229. <https://doi.org/10.1145/3287560.3287596>