International Chinese Language Education Communications
国际中文教育通讯
ISSN 3078-3348

International Chinese Language Education Communications
Volume 3, Issue 1, 1-26
https://doi.org/10.46451/iclec.20251101

Received: 25 June, 2025
Accepted: 30 October, 2025
Published: 12 November, 2025

Evaluating Current Tools for Pinyin Transcription: Can Customising a Chatbot Lead the Way Forward?

Sara Rovira-Esteva
Departament de Traducció, Interpretació i Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona, Spain

Abstract

The Chinese government introduced the "Chinese Phonetic Notation Plan" (known as Pinyin) in 1958 to combat illiteracy, eventually formalizing it as a standardised transcription system in 2012. The correct application of Pinyin orthographic rules is essential for language learning, international communication, and digitization. This research is driven by the belief that accurate transcription of Chinese text into Pinyin is crucial, while acknowledging that the process can be difficult and tedious when done manually. Therefore, this study aims to assess the performance of various Pinyin automatic transcription tools, identify problematic aspects in transcription, and determine whether customised systems can improve results while reducing user effort. The study employs a multi-phase methodology, including the analysis of representative transcription tools, comparison of errors, and the customisation of a chatbot for enhanced performance. The results reveal that most dedicated tools segment transcriptions at the character level rather than by word. General GenAI systems perform better than specific tools, but none followed the rules consistently. Common problems were identified in reduplication, punctuation, neutral tone, and word identification. Although DeepSeek had better initial performance, the customised and trained version of ChatGPT-4 achieved superior results in adherence to Pinyin rules, though perfect accuracy proved unattainable. This research highlights the challenges faced in automated transcription and offers insights into future improvements for systems aimed at assisting users with Pinyin transcription.

Keywords

Pinyin, Chinese transcription, Pinyin converters, ChatGPT-4, DeepSeek

现有拼音转写工具评估：人工智能能否引领下一代技术？

罗飒岚
翻译、口译与东亚研究系, 巴塞罗那自治大学，西班牙

摘要
1958 年，中国政府颁布《汉语拼音方案》以推动扫盲运动；2012 年，该方案被进一步确立为标准化转写体系。拼音正词法规则的准确应用直接影响语言学习、国际传播与信息处理。鉴于人工标注耗时费力，而汉语文本的拼音转写在上述场合重要且必须，本研究评估了当前主流自动拼音转写工具的表现，以识别现存问题，并探讨定制化系统能否在减轻人工负荷的同时提高转写精度。研究采用多阶段方法，包括典型工具测评、错误对比以及定制聊天机器人实验。结果显示，（1）大多数专用转写工具仍以单字为转写单元，未能实现词语级切分；（2）通用生成式人工智能系统的表现虽优于部分转写工具，但仍难以稳定地遵循正词法规则；（3）常见误差集中于叠词、标点符号、中性声调及词语切分。虽然 DeepSeek 在初始测试中暂居优势，经定制与训练后的 ChatGPT-4 在遵循拼音规则方面却更胜一筹，然仍未达到完全准确。本研究呈现了自动转写实践中的主要挑战，并为后续系统的优化提供了实证参考。

关键词
拼音，汉字转写，拼音转写工具，ChatGPT-4，DeepSeek

Introduction
Chinese has a morphosyllabic writing system. Although most characters today are pictophonetic, meaning they contain both semantic and phonetic information, they do not provide phonetic information in a systematic manner. As a result, being able to pronounce them requires years of learning and practice. In the late 19th and early 20th centuries, Chinese reformers considered romanising Chinese to facilitate the learning of writing and reduce the high illiteracy rates in the country. After several decades of trial and error and the creation of hundreds of transcription systems for Chinese, the “Chinese Phonetic Notation Plan” (汉语拼音方案, *Hànyǔ Pīnyīn fāng'àn*) was promulgated in 1958. This plan, abbreviated as Pinyin (拼音, pīnyīn), received unprecedented support from the Chinese government as an auxiliary transcription system, though without any intention of replacing the character-based writing system, as was originally planned. Hànyǔ Pīnyīn was recognised as an international standard by the International Organization for Standardization (ISO) in 1982, under ISO 7098, while the United Nations adopted it in 1986 as the exclusive system for transcribing Chinese geographical and personal names in all its publications.

After almost four decades of use, in 1996, the Chinese government published the *Basic rules of the Chinese phonetic alphabet orthography* (《汉语拼音正词法基本规则》, *Hànyǔ pīnyīn zhèngcífǎ jīběn guīzé*), an updated and more detailed version of the original 1958 rules, aimed at providing a practical framework for standardising spelling issues related to the Pinyin transcription system. This document further refined and expanded the application of Pinyin, addressing issues such as tone marking, punctuation, and the transcription of proper names. In 2012, these rules were revised and updated based on the experience gained from years of implementation, becoming the new national standard (GB/T 16159-2012) and providing the current and most authoritative guidelines for Pinyin transcription. Each of these documents

 ISSN 3078-3348

aimed at contributing to the evolution and standardisation of Pinyin orthography, not only in China but also in making it the universal transcription system for Chinese worldwide.

The correct application of transcription rules for Chinese is crucial for several reasons. First, it helps learners pronounce unfamiliar words and characters. Second, it assists foreign students in identifying word units and consolidating new vocabulary, thereby enhancing reading comprehension among Chinese as a foreign language (CFL) learners (Xiao et al., 2020). Third, depending on their personal or professional needs, learners such as tourists, businesspeople, reporters, diplomatic support staff, or participants in short-term courses for senior citizens may not need to learn Chinese characters (Kubler, 2022, p. 78). In other words, they will need to rely on Pinyin to navigate and communicate. Fourth, both foreign students and Chinese citizens are increasingly using Pinyin to digitally input characters on their mobile devices or computers. Finally, it enables people worldwide, with no knowledge of Chinese, to speak or write about Chinese proper names or concepts in a standardised way in any context where transcription is needed, even if they do not know exactly how Pinyin letters are pronounced. Thus, the applications of Pinyin, both inside and outside China, are numerous, ranging from library catalogues, product labelling, street signs, sorting entries in dictionaries, textbooks for standard Chinese learning, news agencies, transportation systems, Braille, Chinese sign language, and, last but not least, digital Chinese text processing. Mair and Hu (2024, p. 39) highlight the importance of Pinyin with the following statement:

> *Pinyin has had a transformative effect on Chinese language teaching and learning and on Chinese society overall. It greatly helps people around the world to learn Mandarin much more easily and quickly than before. It has also enabled China to transition to the digital age: people use Pinyin to easily type Chinese characters on smartphones and computers, which makes communication between people convenient and fast.*

Although Pinyin is not strictly a writing system, it is crucial to adhere to its orthographic rules. We agree with Kubler (2022, p. 75) when he affirms that "Hànyǔ Pīnyīn must be written correctly, with connection of syllables into words, use of the apostrophe where needed, correct placement of tone marks, and proper capitalization". Doing so not only aids in the standardisation of Pinyin but also enhances the clarity and comprehension of texts for readers, since a spelling system that adequately represents the character combinations forming words in a text can resolve potential ambiguities (Hincha, 2004, pp. 17–18). A practical example of these benefits can be found in Arsenault (2001), who demonstrated that the correct application of Pinyin conventions —particularly using the word rather than the individual character as the basis for transcription—enhances both the precision and efficiency of data retrieval in bibliographic library catalogues.

However, despite being a national standard, the official documents discussed above have not been widely disseminated, nor have the authorities effectively conveyed their importance to the public. As Hannas (1997, p. 274) observes, although educated speakers may be familiar with romanisation rules, the lack of consistent reinforcement and ongoing debates among linguists mean that Chinese users often fail to apply Pinyin rules correctly, particularly in areas such as word division, tone marking, and segmentation. This demonstrates that the measures implemented to enforce the application of Pinyin rules have been insufficient, which in turn helps explain why many users—including teachers of Chinese as a second language and governmental bodies—are either unaware of these rules or fail to apply them correctly. As a result, errors are commonly found in all kinds of texts and contexts. This research is, thus, motivated by the conviction that it is important to correctly transcribe Chinese text into Pinyin,

 ISSN 3078-3348

while also recognizing that it can be a difficult and tedious task for users to do manually. Therefore, the present study pursues the following three main objectives:

1. To analyse a representative sample of different Pinyin automatic transcription tools to determine which are more accurate and reliable and thus require less user intervention.
2. To identify which aspects of applying the official transcription rules are most problematic for these tools, in order to provide users with guidance on where to pay special attention.
3. To find a system that can transcribe large amounts of Chinese text into Pinyin as accurately as possible in accordance with Pinyin orthographic rules, thereby minimising the time users spend revising the resulting transcription.

The working hypotheses guiding this research are:

1. Most automatic tools for transcribing Chinese characters fail to apply official orthographic rules correctly, although their accuracy varies.
2. These systems perform better in some transcription aspects than in others.
3. Given that most chatbots are trained primarily on large volumes of English linguistic data—introducing a well-documented Anglophone bias (Stockwell, 2024, p. 5)—it is reasonable to hypothesise that a Chinese-based GenAI system may outperform Western systems in tasks specific to the Chinese language.
4. Using a customised GenAI system that combines fine-tuning with a Retrieval-Augmented Generation (RAG) approach, and is trained with specific instructions, can improve results and reduce the workload for users when correcting the resulting transcription.

In addition to this introductory section, the article includes a conceptual framework summarising the key official Pinyin transcription rules and reference documents used in the analysis; a methodology section outlining the research process; and a results and discussion section presenting the findings. The conclusions highlight the main results, the limitations of the study, and avenues for future research. The bibliography lists all references cited in the study, while the data collected during the analysis are provided in a supplementary Excel file.

Conceptual Framework

Since this study aims to investigate to what extent the transcription tools available to users provide an accurate transcription according to the official spelling rules, the conceptual framework of this work will mainly present the key aspects of this national standard. The *Basic Rules of Chinese Pinyin Orthography* published by the Chinese government in 1996, aimed to provide a useful tool for standardising orthographic issues related to this transcription system. In 2012, an updated version of these rules was published based on the experience accumulated over the years of application. Due to space limitations, the entire translated document cannot be included in this article.[1] Therefore, below is a summary of the key aspects of the official orthographic rules for Pinyin transcription that are relevant to this study, with a focus on teaching Chinese as a foreign language. These rules have been selected either because they often raise questions or because of their high occurrence rate. We have included the item number in parentheses at the end of each rule to facilitate reference to the original document in case the reader wishes to consult the examples, should there be any doubts.

General principles

- The word is the basic unit of transcription. To determine what constitutes a word and what does not, we must consider its grammatical category, as well as aspects such as phonetics, semantics, and length (5.1).

 ISSN 3078-3348

- Two- or three-syllable structures that express a single concept should be written together (5.2).
- Four-syllable terms or longer that express a single concept should be transcribed separately according to the words or elements they are composed of (for example, if they are separated by pauses in speech). They should only be transcribed together if it is not possible to divide them into words (5.3).
- Monosyllabic reduplicated words should be transcribed together. Disyllabic reduplicated words of the ABAB type should be transcribed separately. Reduplicated words of the AABB type should be transcribed together (5.4).
- Monosyllabic prefixes and suffixes should be transcribed together with the word they form part of (5.5).
- In certain juxtaposed structures, a hyphen can be added between morphemes or abbreviations to facilitate reading and comprehension (5.6).
- Only the original tones should be indicated, that is, tone sandhi[2] is not marked unless justified for pedagogical reasons (6.5.2).

Basic rules

Nouns

- Nouns should be transcribed separately from the locatives they modify; if these co-occurrences are lexicalised, the elements should be transcribed together (6.1.1).

Verbs and adjectives

- The verb should be transcribed separately from its object. Verbs that are morphologically formed by a verb and object should be transcribed separately when other words are inserted (6.1.2.2).
- If both the verb (or adjective) and its verbal complement (resultative, potential, etc.) are monosyllabic, they should be transcribed together, while in other cases, they should be transcribed separately (6.1.2.3).
- Monosyllabic adjectives and their reduplicated prefix or suffix should be transcribed together (6.1.3.1).

Numerals

- Numbers and measure words should be transcribed separately (6.1.5.6).
- Numbers from eleven to ninety-nine should be transcribed together (6.1.5.2).
- Numbers *bǎi* (百), *qiān* (千), *wàn* (万), and *yì*(亿) should be transcribed together with the unit that precedes them, although *wàn* (万) and *yì*(亿) should be transcribed separately from the number that precedes them if it consists of more than one digit. If the number preceding them is *shí*(十), they can also be transcribed together (6.1.5.3).
- A hyphen should be used between the ordinal prefix *dì*(第) and the numeral it modifies (6.1.5.4).

Particles

- The aspectual particles *zhe* (着), *le* (了), and *guo* (过) should be transcribed together with the verb they modify, while the modal particle *le* (了) at the end of a sentence should be transcribed separately (6.1.2.1).
- Structural particles such as *de* (的), *de* (地), *de* (得), *zhī* (之), and *suǒ* (所) should be transcribed separately from the rest of the words. In the case of the particles *de* (的), *de*

 ISSN 3078-3348

(地), and *de* (得), if the word preceding them is monosyllabic, they can also be transcribed together with that word (6.1.9.1).
- Modal particles should be transcribed separately from the rest of the words (6.1.9.2).

Other parts of speech
- All kinds of pronouns (6.1.4.1, 6.1.4.2, 6.1.4.3), adverbs (6.1.6.), prepositions (6.1.7), conjunctions (6.1.8), interjections (6.1.10) and onomatopoeic words (6.1.11) should be transcribed separately from the rest of words in the sentence.

Idiomatic expressions and fixed phrases
- Idiomatic expressions, mainly composed of four-character expressions from Classical Chinese, often function independently in speech. If, from a structural point of view, they can be divided into two disyllabic elements, they should be separated with a hyphen. Expressions that cannot be divided into two disyllabic groups should be transcribed together (6.1.12.1).
- Fixed expressions that are not four-character phrases, as well as other fixed expressions, should be separated according to the words that form them (6.1.12.2).

Transcription rules for anthroponyms and toponyms
- In Chinese anthroponyms, the surname and first name should be transcribed separately, with the initial letter of both capitalised. The surname comes first, followed by the given name. Surnames consisting of more than one syllable should be transcribed together. Nicknames or pseudonyms should be transcribed following the same criteria (6.2.1.1).
- Proper names should be transcribed separately from titles or other forms of address, which should be written in lowercase (6.2.1.2).
- Prefixes that form honorifics such as *lǎo* (老), *xiǎo* (小), *dà* (大), or *ā* (阿) should be written separately from the noun and with an initial capital letter (6.2.1.3). In cases where honorifics have become part of the proper name, they should be transcribed together with the name, with the initial letter capitalised (6.2.1.4).
- Chinese toponyms should be transcribed separately from the common nouns they modify, and the initial letters of both elements should be capitalised (6.2.2.1). Monosyllabic prefixes or suffixes should be transcribed together with the proper or common nouns they attach to, and the initial letter should be capitalised (6.2.2.2). If the co-occurrence has become lexicalised, toponyms should be transcribed together with the common noun that follows them (6.2.2.3).

Orthotypographic rules
- The first letter of a sentence and a verse should be capitalised (6.3.1).
- The first letter of proper names should be capitalised. If the proper name consists of more than one word, the initial letter of each word should be capitalised (6.3.2).
- If a proper noun is combined with a common noun, the first letter of the proper noun should be capitalised (6.3.3). However, if the word has become a common noun, the first letter should be written in lowercase (6.4).
- When transcribing in Pinyin, the full stop (。) is transcribed as a small dot (.), the em dash (—) is written as a short dash (-), the six ellipsis dots (……) are written as three dots (…), and the pause sign (、) can be replaced with a comma (,) (6.7).
- In cases where a non-initial syllable of a word starts with *a*, *e*, or *o*, it should be separated from the previous syllable by an apostrophe (e.g., *Tian'anmen*).

 ISSN 3078-3348

As the official document only provides general schematic guidelines, we have also relied on the interpretive criteria of two works that, following their publication, aimed to develop the standard with numerous concrete examples to cover a wide range of cases. These works are Yin and Felley's (1990) and Shangwu Yinshuguan Cishu Yanjiu Zhongxin's (2002). Furthermore, in cases where there were contradictions between the different works or doubts about whether a particular item is considered a word in Modern Standard Chinese, we also consulted the normative dictionary 现代汉语词典 (*Xiàndài Hànyǔ Cídiǎn*). In short, our conceptual reference framework for conducting the analysis is based on the following five works:

- 汉语拼音正词法基本规则 (*Hànyǔ pīnyīn zhèngcífǎ jīběn guīzé*) / *Basic rules of the Chinese phonetic alphabet orthography* (2012)
- 中国人名汉语拼音字母拼写规则 (*Zhōngguó rénmíng Hànyǔ pīnyīn zìmǔ pīnxiě guīzé*) / *The Chinese phonetic alphabet spelling rules for Chinese names* (2011)
- 汉语拼音和正词法 (*Hànyǔ Pīnyīn hé Zhèngcífǎ*) / *Chinese romanization pronunciation & orthography* (1990)
- 新华拼写词典 (*Xīnhuá Pīnxiě Cídiǎn*) [Xinhua Pinyin Dictionary] (2002)
- 现代汉语词典 (*Xiàndài Hànyǔ Cídiǎn*) [Modern Chinese Dictionary] (2016)

Methodology

The methodology followed in this study can be divided into five phases: 1) selection of tools for Pinyin transcription; 2) contrastive analysis of errors in the application of official orthographic rules based on the transcription of the same text using the tools selected in Phase 1; 3) performance analysis based on a selection of orthographic rules using two of the systems tested in Phase 2; 4) training the two GenAI systems selected in Phase 2 to assess whether it improves their transcription accuracy; and 5) creation of a customised chatbot for Chinese–Pinyin transcription, enhanced through fine-tuning and a RAG approach, to evaluate whether this leads to further improvements in transcription accuracy. We will now explain each of these phases in more detail.

Phase 1. Selection of Pinyin transcription tools

To obtain a representative sample of transcription systems for analysing their performance in terms of adherence to Pinyin rules, we first gathered a range of systems from two different sources. On one hand, we asked colleagues about the systems they used for transcription; on the other hand, we consulted the e Chinese Tools resources database (Rovira-Esteva et al., 2021–2025) using the combined filter *tool* and *Pinyin*. These two sources of data resulted in a list of resources of different types, which we will discuss in more detail in the results section.

Phase 2. Contrastive analysis of orthographic rule application

The tools that passed the initial screening, as they segment transcriptions at the word level rather than by character and allow the input of large amounts of text to be transcribed, were Google Translate (which automatically provides a Pinyin transcription when Chinese text is entered) and five GenAI systems. Since GenAI systems are being used for a myriad of functions for which they were not originally designed, we thought it would be a good opportunity to assess their potential in relation to our study's focus.[3] Therefore, a more detailed analysis of their performance in transcribing a text was carried out on a representative sample of these systems, which included ChatGPT-4, DeepSeek, Claude, Gemini 2.0 Flash, and Copilot. We could not find any other machine translation system with a transcription feature comparable to that of Google Translate, which is why it is the only resource of its kind in this list.

 ISSN 3078-3348

To carry out this analysis, on 17th March 2025, the following prompt was entered into each of these GenAI systems: "Transcribe into Pinyin according to the official spelling the following text", followed by a passage from the Chinese text of "My Old Home" (故乡, *Guxiang*) by Lu Xun. We chose this text because it had a standardised transcription by Yin and Felley (1990) that could be used as a reference, although it should be borne in mind that this publication predates the amended Pinyin rules (2012).

The resulting transcription was copied and pasted into a spreadsheet, where all cases that did not conform to the standard were marked in red (purple was also used when there were two different errors in the same word, to facilitate counting) and labelled according to the Pinyin transcription rules. The number of errors made by each system was also quantified, counting each occurrence independently, regardless of whether the same word or kind of error appeared more than once in the text. Two of the systems were selected to proceed to the next phase. We will explain which ones and why in the results and discussion section.

Phase 3. Performance analysis of two systems

After carefully reviewing the official Pinyin orthography, we selected 37 rules, either due to their frequent occurrence or because they commonly raise questions (see Table 2), each accompanied by at least one example, mostly taken from the official guidelines (List 1). The specific list of these rules can be found in the corresponding sheet of the Excel file used for data annotation and analysis (see Supplementary File). We then asked ChatGPT-4 and DeepSeek to transcribe the list of words, expressions, and sentences with the following prompt: "Transcribe into Pinyin the following list of items. It is very important that you adhere to the official orthographic rules". We again recorded in a new spreadsheet whether the system adhered (Y) or not (N) to the norm.

Obviously, it is possible that more than one rule needs to be applied simultaneously within the same sentence or expression. However, to facilitate the analysis, the transcription was considered correct if the system correctly applied the rule being analysed at each moment. It was marked with an asterisk (Y*) if the rule in question was correctly applied, but the system made an error not directly related to that rule.

Phase 4. Training for enhanced transcription accuracy

The study's fourth phase comprised two steps. The first step (4.1) involved enhancing the chatbots through fine-tuning and a Retrieval-Augmented Generation (RAG) approach, aimed at improving the results obtained in the previous phase (List 1). Both systems were given the same prompt, together with two attached PDFs containing the official Pinyin orthographic rules:

> *Your transcription has some errors according to the official orthographic rules for Pinyin. I am attaching several related documents for you to consider, and kindly re-transcribe the cases where you have not followed the rules in the list above. Please also indicate the cases where you have made changes.*

The prompt included two attached PDFs: 汉语拼音正词法基本规则 / *Basic rules of the Chinese phonetic alphabet orthography* (2012) and 中国人名汉语拼音字母拼写规则 / *The Chinese phonetic alphabet spelling rules for Chinese names* (2011). In the second step of this phase (4.2), we asked the system to re-transcribe the Lu Xun fragment used in Phase 2 to assess whether there had been any improvement, given that the systems had already been trained with specific materials. The prompt used was as follows: "Now transcribe this text, taking into account the reference materials I have attached, which correspond to the official Pinyin orthography".

 ISSN 3078-3348

Phase 5. Customizing a version of ChatGPT specifically for transcription

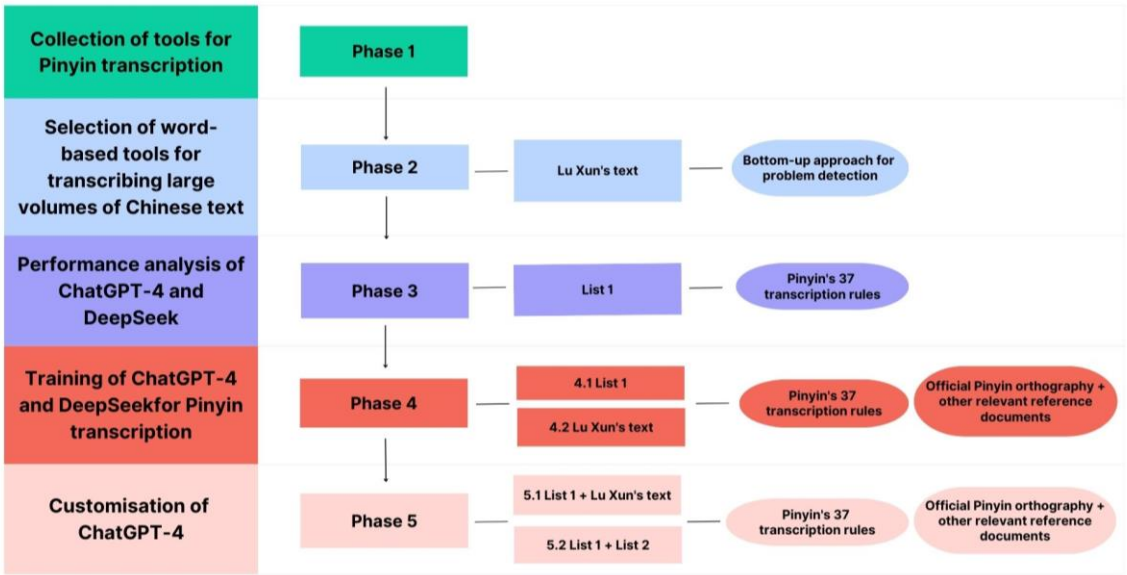
The fifth phase of the study also involved two steps. The first one (5.1.) consisted of creating a customised version of ChatGPT called "Chinese-Pinyin Transcriber", specifically designed for transcription.[4] Its description reads: "This chatbot specialises in transcribing Chinese characters (both simplified and traditional) into Pinyin according to official orthographic rules". We also gave it the following instructions:

> *This GPT is a Pinyin transcriber that follows the official orthographic rules of the Pinyin system for converting Chinese characters into Pinyin. It ensures accurate and proper transcription according to official guidelines, maintaining correct tone markings and spelling conventions whenever requested. Please refer to the attached documents for reference to avoid any mistakes when transcribing a word, expression, or text. Make sure to consult the reference materials provided for your training. If you are uncertain about how to transcribe a word, expression, or fragment, you should notify the user. Additionally, you should alert the user to cases where certain character combinations are not yet fully lexicalized as words or where there is no consensus on the matter.*

We insisted that the chatbot consult the documents we attached when performing the transcription. Furthermore, we provided the customised chatbot with the five reference materials in PDF format mentioned at the end of the section on the conceptual framework, ensuring that the system had all the necessary information to adhere to the official orthography and avoid mistakes, and at the same time taking a step toward developing a more robust RAG-based solution. Then, we asked this specialised chatbot to transcribe, on the one hand, the same Lu Xun text from Phase 2 with the simple prompt, "Please transcribe the following text", and, on the other hand, the list of words, phrases, and sentences from Phase 3 (List 1) with the following prompt: "Please transcribe the following list of items following the Pinyin transcription rules. Do not capitalise all the items, only those that are strictly necessary to adhere to the official orthography". This additional instruction was necessary because, in the previous round, the chatbot had capitalised the first letter of all the items in the list, which made it difficult for us to accurately determine whether this was due to its adherence to the Pinyin rules or other factors.

Figure 1

Research Phases and Materials Used



In the second step (5.2), we provided the "Chinese-Pinyin Transcriber" with a completely new list of items (List 2) to transcribe, in order to assess whether the training was successful and if the system was able to consolidate its knowledge of transcription for the selected 37 rules.

An Excel spreadsheet was used as a tool for data collection and analysis, with a separate tab for each phase of the analysis. The results can be found in the Supplementary Excel file. Figure 1 above summarises the different phases of this study and their main features.

Results and Discussion

Phase 1. Selection of Pinyin transcription tools

The first phase was the shortest of all. After informally asking our colleagues, who are Chinese language teachers, about the tools they used for Pinyin transcription and conducting a specific search in the e Chinese Tools database (Rovira-Esteva et al., 2021-2025), we obtained a diverse list of resources with various types of tools. The list included text processors (Microsoft Word), Chrome extensions (Zhongwen Chinese Popup Dictionary, Chinese Tools), online dictionaries (Yellowbridge, Han Dian), online conversion tools (Purple Culture, Arch Chinese, Hanyu Pinyin, Chinese Gratis, Chinese Boost), mobile applications (Pleco, Pinyiner), machine translation systems (Google Translate), and GenAI systems (ChatGPT, DeepSeek, Gemini, Copilot, Claude). After an exploratory analysis of these systems, we discarded those that were not functioning, those that do not use the word as the basis for transcription and instead transcribe sinogram by sinogram (Microsoft Word, Purple Culture, Han Dian), and those whose input method does not allow for transcribing large amounts of text (Pleco, Yellowbridge, Pinyiner, Chinese Tools).

This first screening made us realise three important issues. First, some colleagues responded that they did not use any specific system but rather transcribed a word themselves when they occasionally needed it. Secondly, it also helped us confirm that teachers do not appear to need to transcribe large volumes of Chinese text. Finally, we were also surprised to find that most of the tools supposedly dedicated to Pinyin transcription, especially the converters, transcribe sinogram by sinogram rather than by word, thus violating the first rule of the official guidelines. This final finding prompted reflection on its underlying causes: the issue could stem from a technical limitation, such as the difficulty of integrating a function within these tools that accurately detects and segments text at the word level. Alternatively, it may result from insufficient knowledge of the official orthographic rules or, more critically, negligence in applying them due to a perceived lack of relevance.

It is worth noting that some of the discarded tools in this phase offer interesting functionalities, such as placing the transcription above the character (Microsoft Word), or providing different transcription options, such as marking tones with diacritics, numbers, or without tone markings (Chinese Gratis, Hanyu Pinyin), marking tones with different colours (Purple Culture), or even offering pronunciation by simply placing the cursor above the selected character (such as the Chrome extensions Zhongwen Chinese Popup Dictionary and Chinese Tools). However, since they were not useful for meeting the objectives of this study, they had to be excluded from the next phase of the analysis.

Phase 2. Contrastive analysis of orthographic rule application

As already mentioned in the methodology section, to carry out the second phase of the analysis, the different systems selected (Google Translate, ChatGPT-4, DeepSeek, Claude, Gemini 2.0 Flash, and Copilot) were tasked with transcribing a passage from the text *Guxiang* (故乡) "My Old Home" by Lu Xun in accordance with the official Pinyin rules. The selection of this

specific text to be transcribed by these systems was motivated by the availability of a full transcription of the text (Yin and Felley, 1991, pp. 528-531).

However, we detected a few errors in Yin and Felley's (1991, p. 529-530) transcription of this fragment. First, in two cases, sinograms that should be transcribed together because they form words according to the *Xiàndài Hànyǔ Cídiǎn* were transcribed as two separate units (之后 and 大雪). Second, two monosyllabic verbs were transcribed separately from their monosyllabic resultative complements (that is, 缚在 and 罩在 should be transcribed as *fùzài* and *zhàozài*, respectively). Third, several tone-marking errors were identified, including a fourth-tone syllable transcribed with a neutral tone (*知道, *zhīdao*),[5] a first-tone syllable erroneously marked with a third tone (*撒, *să*), and a reduplicated adjective that should be in the third tone but appeared in the second (*明晃晃, *mínghuánghuáng*). Fourth, the transcription did not match the standard pronunciation of the sinogram (e.g., 批 was transcribed as *bǐ* instead of *pī*). Finally, two different words were transcribed as a single unit (e.g., 都有 should be transcribed as *dōu yǒu* instead of *dōuyǒu*). Therefore, our analysis used their transcription as a starting point but applied the Pinyin rules whenever we found that they had not been followed correctly.

The resulting transcriptions of the text object of study were included in a spreadsheet, in which all cases that did not conform to the standard were marked in red and labelled. As already mentioned, purple was also used when there were two different errors in the same word, to facilitate counting. Although not all the systems made the same mistakes, the problems detected were quite widespread and can be summarised in the following points, in the order they appeared in the analysis:

1. All systems except Copilot failed to correctly transcribe reduplicated nouns as a single unit (e.g., 日日, *rìrì*).
2. Except for Google Translate and ChatGPT-4, most systems failed to identify person names by capitalising the first letter (e.g., 闰土, *Rùntǔ*).
3. Only Google Translate transcribed aspectual particles together with the verb they modify, as the other systems transcribed them as separate units (e.g., 到了, *dàole*; 见过, *jiànguo*). Since in some cases the aspectual particle *le* at the end of a sentence can be considered both an aspectual and modal marker, these cases were not counted as incorrect whether transcribed together with the preceding verb or as two separate units.
4. None of the systems succeeded in transcribing resultative or directional complements together with the one-syllable verb they modify, and these were transcribed as separate units (e.g., 套住, *tàozhù*; 撒下, *sāxià*; 罩在, *zhàozài*; 扫出, *sǎochū*).
5. Although only word-based transcription systems were included in this phase of the analysis, none of the systems succeeded in identifying all the words in the text, meaning some of the morphemes forming words were transcribed as separate units (e.g., 飞跑, *fēipǎo*; 毡帽, *zhānmào*).
6. Numerals should be transcribed separately from the measure words they precede. However, they were found to be transcribed as a single unit (e.g., 一个, *yī gè*; 一块, *yī kuài*).
7. None of the systems transcribed all reduplicated adjectives in the text as a single unit (e.g., 明晃晃, *mínghuǎnghuǎng*, 远远, *yuǎnyuǎn*).
8. All systems except DeepSeek transcribed some different words that should be transcribed separately as a single unit (e.g., 不能, *bùnéng*, 也有, *yě yǒu*).
9. Tone sandhi should not be marked, but characters pronounced in the neutral tone in given morphological combinations or linguistic contexts should be transcribed

 ISSN 3078-3348

accordingly. However, all systems failed to do so, transcribing those syllables according to the morpheme's original tone (e.g., 母亲, *mǔqin*; 见过, *jiànguo*).

10. Ordinal prefixes should be transcribed with a hyphen after them, preceding the numeral (e.g., 第二, *dì-èr*). However, all systems failed to adhere to this rule.
11. Common nouns should not be capitalised, even if they derive from a proper noun, but Google Translate and Copilot failed to do so (e.g., 观音手, *guānyīnshǒu*).
12. The transcription did not correspond to the pronunciation of that sinogram in the given context, either because it was an error (for example, 半日 was transcribed as *bàrì* instead of *bànrì*, or 批谷, which was transcribed as *bǐ gǔ* instead of *pīgǔ*) or because the system failed to recognise it as polyphonic sinogram (transcribing 地 in 远远地 as *di* instead of *de*).

The qualitative analysis of the results shows that there are aspects of the transcription rules that all systems ignore, while others are taken into account. For example, reduplication, the use of punctuation marks, the neutral tone, and, in general, the ability to distinguish what constitutes a word and what does not are common difficulties across all systems.

Table 1 summarises each system's performance in transcribing Lu Xun's text based on 12 Pinyin orthographic rules. The systems exhibited varying degrees of accuracy: ChatGPT-4 failed to apply any rules correctly, Google Translate adhered to only two, and DeepSeek—a system of Chinese origin—correctly followed five rules. Errors were quantified by counting each occurrence, even when the same mistake appeared multiple times. In the 338-character, 249-word text, the error counts (in ascending order) were: DeepSeek (44), Copilot (46), Claude (47), ChatGPT-4 (52), Gemini (53), and Google Translate (57). As Table 1 illustrates, no consistent pattern emerged across the six systems, nor was there a direct correlation between rule adherence and total errors, as a single rule could apply to varying numbers of cases in the text.

This approach can be described as a bottom-up method, as we inferred the transcription rules violated by the system after analysing the problems detected in a given text. Therefore, the issues that arose are not necessarily representative of all the challenges these systems face in applying the official Pinyin transcription rules as a whole. However, it was useful to assess the baseline performance of some of the most widely used generalist GenAI systems and to realise that they do not present significant differences in this regard. To proceed to the next phase, we selected DeepSeek as the top-performing system and ChatGPT-4 for its paid version's support of customised chatbots, which we expected could leverage fine-tuning and RAG—capabilities shown in the literature (Guo, 2024; Balaguer et al., 2024) to substantially enhance GenAI accuracy and reliability across tasks—to achieve a higher level of performance.

Phase 3. Systematic performance analysis of ChatGPT-4 and DeepSeek

To complement the bottom-up method used in the previous phase (moving from the example to the rule), in the third phase, we adopted a top-down approach (starting from the rule to observe how it is practically applied). To this end, we selected 37 rules (see Table 2), some of which are quite general, while others are more specific. These rules aimed to cover different parts of speech, as well as a broad and representative range of cases, including not only linguistic issues but also orthotypographic conventions. After asking ChatGPT-4 and DeepSeek to transcribe a list of example words, expressions, and sentences (List 1) according to the official orthographic rules, we recorded their responses in a new spreadsheet and assessed their adherence to the norm: "yes" (in green) and "no" (in yellow). However, in some cases

 ISSN 3078-3348

(identified with an asterisk), the specific norm could be considered to have been correctly applied by the systems, but the resulting transcription contained other types of errors.

Table 1
Adherence to Pinyin Rules and Number of Errors by the Different Automatic Transcription Systems Included in Phase 2 (in Alphabetical Order)

Pinyin rule	**ChatGPT4**	**Claude**	**Copilot**	**DeepSeek**	**Gemini**	**Google Translate**
Reduplication of nouns	X	X	√	X	X	X
Capitalisation of proper names	X	√	√	√	√	X
Attachment of aspectual markers to verbs	X	X	X	X	X	√
Attachment of resultative complements to verbs	X	X	X	X	X	X
Correct identification of lexicalised combination of characters as words	X	X	X	X	X	X
Separation of numerals from measure words	X	√	√	√	X	X
Reduplication of adjectives	X	X	X	X	X	X
Marking of neutral tone	X	X	X	X	X	X
Use of hyphen after ordinal prefix	X	X	X	X	X	X
Separation of all words	X	X	X	√	X	X
All syllables correctly transcribed	X	X	X	√	√	X
Common nouns are not capitalised, even if they derive from a proper noun	X	√	X	√	√	√
Rules correctly applied	**0**	**3**	**3**	**5**	**3**	**2**
Total number of errors	**52**	**47**	**46**	**44**	**53**	**57**

Table 2 shows the performance results of the two systems for each of the selected rules. As can be seen, while DeepSeek succeeded in 83.78% of the cases (31/37), ChatGPT-4 failed to correctly transcribe almost all the words, expressions, or sentences provided. As a result, only five out of 37 items (13.51%) showed no transcription errors for that specific rule. While

 ISSN 3078-3348

DeepSeek's performance can be considered quite outstanding in comparison to ChatGPT-4's, it still falls short of being a reliable and accurate transcription tool, particularly considering it is a China-based system.

Table 2
Adherence to 37 Different Pinyin Transcription Rules of the Two Systems Included in Phase 3

Pinyin rule	**ChatGPT-4**	**DeepSeek**
Different characters that form a word are transcribed together.	N	Y
Syllables pronounced with the neutral tone are transcribed accordingly.	Y*	Y*
Monosyllabic reduplicated words are transcribed together.	N	Y
Reduplicated words of the AABB type are transcribed together.	N	Y
Disyllabic reduplicated words of the ABAB type are transcribed separately.	N	Y
Monosyllabic prefixes and suffixes are transcribed together with the word they form part of.	N	N
Nouns are transcribed separately from the locatives they modify.	Y	N
Lexicalised nouns modifying locatives are transcribed together.	N	Y
The verb is transcribed separately from its object.	N	Y
The object of a verb-object construction, when it has become a meaning-empty particle, should be transcribed together.	N	Y
Verbs that are morphologically formed by a verb and object are transcribed separately when other words are inserted.	N	Y
Monosyllabic verbs and their verbal complement are transcribed together.	N	Y
Non-monosyllabic verb and their verbal complement are transcribed separately.	N	Y
Monosyllabic adjectives and their reduplicated prefix or suffix are transcribed together.	N	Y*
Measure words are transcribed separately from the preceding words (numbers, demonstratives, etc.).	Y*	Y
Numbers from eleven to ninety-nine are transcribed together.	N	Y
Numbers *bǎi* (百), *qiān* (千), *wàn* (万), and *yì* (亿) are transcribed together with the unit that precedes them. *Wàn* (万) and *yì* (亿) are transcribed separately from the number that precedes them if it consists of more than one digit (if the number preceding them is *shí*[十] both options are correct).	N	N
Numbers in a date are transcribed separately.	Y	Y
A hyphen is used between the ordinal prefix *dì* (第) and the numeral it modifies.	N	N
The aspectual particles *zhe* (着), le (了), and *guo* (过) are transcribed together with the verb they modify.	N	Y
Structural particles such as *de* (的), *de* (地), *de* (得), *zhī* (之), and *suǒ* (所) are transcribed separately from the rest of the word (in the case of the three structural particles *de* [的, 地, and 得], if the word preceding them is monosyllabic, they can also be transcribed together with that word).	N	Y

 ISSN 3078-3348

Modal particles should be transcribed separately from the rest of the words, including the modal particle *le* (了).	N	Y
Idiomatic expressions composed of four-character phrases, which from a structural point of view can be divided into two disyllabic elements, are separated with a hyphen.	N	N
The first name and surname are transcribed separately, with the initial letter of both the first name and surname capitalised.	N	Y
Surnames and given names consisting of more than one syllable are transcribed together. Nicknames or pseudonyms should be transcribed following the same criteria.	N	Y
Proper names are transcribed separately from titles or other forms of address, which are written in lowercase.	N	Y
Prefixes that form honorifics such as *lǎo* (老), *xiǎo* (小), *dà* (大), or *ā* (阿) are written separately from the noun and with an initial capital letter.	N	Y
Proper place names are capitalised.	N	Y
Chinese toponyms are transcribed separately from the common nouns they modify, and the initial letters of both elements are capitalised.	N	Y
Monosyllabic prefixes or suffixes are transcribed together with the proper or common nouns they attach to, and the initial letter is capitalised.	N	Y
Toponyms should be transcribed together with the common noun that follows them if the combination has become lexicalised.	N	N
The first letter of a sentence is capitalised.	N	Y
When a proper noun is combined with a common noun, the first letter of the proper noun is capitalised, unless the word has become a common noun, in which case the first letter is written in lowercase.	N	Y
When transcribing punctuation marks, they are adapted to the ones used in alphabetical writing systems.	Y*	Y
When a non-initial syllable of a word starts with *a*, *e*, or *o* it is separated from the previous syllable by an apostrophe.	N	Y
The first letter of each word in the title of a literary work is capitalised.	N	Y*
In certain juxtaposed structures, a hyphen is added between morphemes to facilitate comprehension.	N	Y
Number and percentage of rules correctly applied.	5 (13.51%)	31 (83.78%)

The six aspects in which DeepSeek failed to comply with the official rules at this stage were: 1) transcribing prefixes as a single unit with the word they are attached to (副部长, *fùbùzhǎng*); 2) transcribing nouns separately from the locatives they modify (山上, *shān shàng*); 3) transcribing *qiān* (千) together with the unit that precedes it (十亿零七万二千三百五十六, *shí yì líng qīwàn èrqiān sānbǎi wǔshíliù*); 4) transcribing the ordinal prefix *dì* (第) with a hyphen before the numeral it modifies (第十三, *dì-shísān*); 5) transcribing idiomatic expressions with a hyphen when they can be divided into two disyllabic elements (风平浪静, *fēngpíng-làngjìng*); and 6) transcribing lexicalised toponyms as a single unit (王村, *Wángcūn*).

 ISSN 3078-3348

It was truly surprising that, despite explicitly requesting transcription according to the rules, neither of the two systems fully adhered to them. Particularly striking was the case of ChatGPT-4, which failed to comply with 33 out of the 37 selected rules, although it correctly applied one rule that DeepSeek failed to adhere to.

Phase 4. Training the chatbot for enhanced transcription accuracy

The fourth phase of the study consisted of two stages. After training and fine-tuning ChatGPT-4, and applying a RAG approach by attaching the relevant official materials to the prompt (4.1), the system re-performed the transcription and claimed to have left all items unchanged except for two. The use of RAG ensured that the model could directly access and apply the relevant transcription rules at inference time, reinforcing consistency with the official standards. According to the system, the key changes made were the addition of proper tone markings, ensuring a consistent structure for certain terms where they had been omitted, and the correct separation of terms according to the rules, ensuring they followed official orthographic guidelines for specific usage contexts. However, after carefully reviewing the transcription, no changes had been made to the supposed corrected items. Conversely, other items that were supposedly unmodified had undergone slight capitalisation adjustments, resulting in one new item (6/37) being correctly transcribed.

After giving the chatbot the specific instructions and the necessary information to transcribe correctly, DeepSeek responded with the following: "No changes were necessary as the original transcription already adhered to the official orthographic rules for Pinyin as outlined in the provided documents. The rules for proper nouns, numbers, dates, idioms, and sentence structure were correctly followed". Despite other transcription errors remaining, the system only corrected one transcription: 山上 (*shān shàng*) was initially transcribed together as *shānshàng*, but it should be treated as two separate words and transcribed as two distinct units. However, the message accompanying this change was contradictory, as it justified the correction by stating: "No changes needed. This follows the rules for noun + 方位词 (directional word)". Overall, the final percentage of rules applied correctly (86.48%) was slightly better, but still fell far short of the desired target, which should be as close to 100% as possible.

The results show that this training—consisting of attaching the relevant PDF materials to the prompt and asking the system to correct its errors—was unexpectedly ineffective. In other words, even with a RAG approach that supplied the system with the necessary information, it did not appear to learn from the interaction or follow the specific instructions, and the output did not improve significantly.

To see if the systems could perform better with words in a larger context, in the second step of the fourth phase of the study (4.2), we instructed the chatbots to re-transcribe the excerpt from Lu Xun's text used in Phase 2 to assess whether there had been any improvement, given that the systems had already been trained with the relevant information on official Pinyin orthography. The results were again entered into a new spreadsheet, where the two transcriptions (pre- and post-training with specific materials) were compared. we recorded the number of transcription errors from the previous attempt that were corrected (in green), the number of errors that remained (in red), the new errors that were not present in the first transcription (in violet), and the total number of errors in the new transcription. Table 3 shows the results of this attempt, compared to the results from Phase 2.

 ISSN 3078-3348

Table 3
Number and Types of Errors Made by The Systems Under Study after Their Training (Phase 4.2)

System used	Errors corrected	Remaining errors	New errors	Total of errors in Phase 2	Total of errors in Phase 4.2.
ChatGPT-4	8	41	12	52	53
DeepSeek	0	44	4	44	48

In the case of ChatGPT-4, the system corrected eight errors but introduced 12 new ones that were not present in the initial transcription, whereas DeepSeek corrected none but still introduced four new errors. It is therefore quite surprising that, after providing both systems with the basic and necessary information to transcribe this text correctly according to the official orthographic rules, their performance actually worsened in this post-fine-tuning interaction compared to the first attempt. While Balaguer et al. (2024) found that combining fine-tuning and RAG led to consistent accuracy gains in the agricultural domain, our results suggest that such improvements do not automatically transfer to all contexts. Chinese–Pinyin transcription appears to present specific challenges—such as strict orthographic rules and low tolerance for variation—that may limit the effectiveness of these techniques, pointing to the need for further domain-specific adaptation.

The case of DeepSeek is especially striking, given that it is a China-based system, as it not only increased the number of errors but also, interestingly, provided a list of key points addressed and how it adhered to the official rules at the end of the transcribed text, offering examples for each case. However, it is worth highlighting some notable instances where the system justifies its transcription by referencing a rule that can be considered a hallucination, as the content contradicts the actual rule outlined in the document provided for training. For example, 扫出 (*sǎochū*) should be transcribed together because it is a monosyllabic verb followed by a monosyllabic complement. However, the system wrongly explains to the user: "扫出一块空地 (*sǎo chū yī kuài kòngdì*): Verb + complement structures are written as separate words". Other times, what it claims does not match what it does. For instance, the chatbot affirms that "reduplicated adjectives are written as single words", but the transcription provided does not follow this rule, as it transcribes 明晃晃 as *m ńg huǎnghuǎng* instead of *mínghuǎnghuǎng*.

Phase 5. Chinese-Pinyin Transcriber: a customised chatbot for optimal transcription
In the fifth phase of the study (conducted on 30th March 2025), we created a customised version of ChatGPT that combined fine-tuning with a RAG approach. This version included not only clear instructions on what was expected from the chatbot, but also all the relevant materials it needed to successfully complete the task, as explained in section 3.5 of the Methodology.

Regarding Lu Xun's text (Phase 5.1), compared to the first attempt with a non-customised chatbot, the customised chatbot corrected 21 errors but introduced nine new ones, resulting in a total of 35 errors—significantly fewer than those detected in Phase 2 and Phase 4.2. In other words, the results were much better than in the first two attempts, but still worse than expected (see Table 4).

 ISSN 3078-3348

Table 4
Performance of the "Chinese-Pinyin Transcriber" on Lu Xun's Text (Phase 5.1)

Errors corrected	Remaining errors	New errors	Total of errors in Phase 2	Total of errors in Phase 4.2	Total of errors in Phase 5.1
21	26	9	52	53	35

Regarding the first 37-item list of words, phrases, and sentences (List 1), the customised ChatGPT-4 performed much better than before, with 32 rules out of 37 (86.48%) applied correctly, 27 more than in the first attempt (Phase 3). However, overall, the accuracy of the transcription in both cases was still far from ideal, as too many errors remained. In the case of Lu Xun's text, in particular, there were still too many aspects requiring correction by the user to achieve a perfect, standard-compliant transcription. Two reasons may explain why the chatbot improved the transcription in the case of the list of isolated words and phrases compared to the text. On the one hand, it may be easier for the system to contextualise and interpret the grammatical category of the words to be transcribed in the first case, since identifying the part of speech is one of the crucial factors when determining whether sinograms should be transcribed as separate units or as a single unit. On the other hand, almost all the items in the list appear as examples in the official transcription rules, which were provided as reference documentation when customising the chatbot. However, if this factor allowed the result to improve, we wondered why it did not perform well in the other cases, which were also listed as examples in those documents.

To determine whether the initial test results were biased due to the use of examples sourced directly from reference materials, we compiled a novel set of untrained examples (List 2) for transcription by the "Chinese-Pinyin Transcriber". The system's performance declined significantly with these new items, correctly applying only 26 of the 37 orthographic rules (70.27% accuracy) compared to its performance on the first list. This discrepancy suggests that the system's earlier results may have been influenced by prior exposure to the reference examples.

As a final attempt to refine the results of the customised chatbot, we then selected the 18 rules that the "Chinese-Pinyin Transcriber" had not adhered to for both lists and instructed the chatbot: "The following list summarises the rules from the official Pinyin orthography that you are not applying correctly. Please correct the above items to adhere to the rules and explain any changes you make". Then, we provided the two lists of items (List 1 and List 2) for the system to transcribe again. We labelled this new interaction in our spreadsheet as Chinese-Pinyin Transcriber+.

On this occasion, for the first list, the system correctly applied the official orthography in 32 out of the 37 selected rules (86.48%), with 29 items corrected compared to the system's performance in Phase 3. However, one item that was correct in the previous interaction was changed to an incorrect transcription and four cases of incorrect transcription remained. In sum, compared to the results with List 1 used in Phase 3, the chatbot retained (R) four transcription errors, corrected (C) 29, and altered (A) one transcription that was originally correct. Overall, the number of correctly applied Pinyin transcription rules by the customised chatbot increased from five (Phase 3) to 32 (Phase 5.2). As for List 2, the "Chinese-Pinyin Transcriber" (5.2+) corrected eight transcription errors and failed to correct two compared to its first attempt (5.2), thus achieving a 91.89% (34/37) adherence to the selected official rules.

 ISSN 3078-3348

Since we also asked the chatbot to justify its changes, we were able to check the rationale behind them and discovered a lack of systematisation, which can be summarised as follows:

- The item was correctly transcribed and justified (e.g., 雪白雪白, *xuěbái xuěbái*).
- The specific rule was applied, but the item was wrongly transcribed due to an error related to another rule (e.g., in 第十三, **dì-13*, a hyphen was added, but the numeral should be in Pinyin instead of Arabic numbers).
- The lexicalisation principle was wrongly applied, treating items that should be separate units as a single word (e.g., 山上, **shānshàng* instead of *shān shàng*).
- Partial adherence to a specific rule (e.g., 十亿零七万二千三百五十六, **shí yì líng qī wàn èrqiān sānbǎi wǔshíliù* instead of transcribing 七万 as *qīwàn*).
- What the system said should be done and what it actually did do not coincide (e.g., aspectual *le* joined with verb, but then it transcribed 游了一个小时的泳 as **yóu le yī gè xiǎoshí de yǒng* instead of *yóule*).
- New rules were invented (e.g., title with prefix capitalised as in 王部长, **Wáng Bùzhǎng*, which should be *Wáng bùzhǎng*).
- There were rules that the system seemed not to be aware of (e.g., transforming Chinese quotation marks to italics to transcribe book titles, as in 《现代汉语词典》, *Xiàndài Hànyǔ Cídiǎn*).

To further fine-tune the system, we engaged in additional interactions with the chatbot. On the one hand, we instructed the system to italicise book titles and present them without the Chinese quotation marks (《...》), and the chatbot was able to provide the corrected transcription for the relevant cases from both lists. On the other hand, we asked the system to better check the documents to revise the transcription of an individual item. In this case, the system had wrongly transcribed 千千万万 with a hyphen (*qiānqiān-wànwàn*) instead of as a single unit (*qiānqiānwànwàn*), as the rule specifies. To assess to what extent the customised chatbot was consulting the attached documents, we asked the system to refer to the documents we had provided to verify whether it should be transcribed with a hyphen and whether it should be considered an idiom (*chéngyǔ*), since the system had justified this transcription by claiming that "a hyphen is added between disyllabic elements in a four-character idiom."

The system seemed to have finally found the relevant information in the attachments, responding:

> *According to the 新华拼写词典, the expression 千千万万 is not classified as a chéngyǔ (成语). It is presented in the context of numerical expressions involving repeated characters and is explicitly indicated to be written as a single unit without a hyphen. The relevant section confirms that AABB-type numerical repetitions like 千千万万 should be 连写 (written together). Thus, the correct transcription is: qiānqiānwànwàn (no hyphen). This also confirms that we should remove the hyphen previously added, as it was incorrectly applying the formatting rule used for four-character idioms with a 2+2 structure.*

However, we detected several flaws in this answer. Although it is true that numerical expressions involving reduplicated characters should be transcribed as a single unit, we could not find the given expression in the 新华拼写词典 (*Xinhua Pinxie Cidian*). Conversely, it does appear in the document of the official orthographic rules (section 5.4) and in Yin and Felley's (1990, p. 513) book, but while it is transcribed as one unit in the former, it is transcribed with a hyphen in the latter.[6] In short, the system seemed to mix sources and failed to detect and

 ISSN 3078-3348

report inconsistencies, as requested in our instructions. Moreover, the system never notified or alerted us about cases where there could be doubts regarding the level of lexicalisation of certain character combinations or where there was no consensus on how the transcription should be, for whatever reason, as explicitly requested in the instructions for the chatbot customisation.

As a further step towards customising and refining the chatbot, we wanted to check if the temperature feature (that is, the parameter that controls the randomness or creativity of the model's responses) could significantly improve the results. Since we wanted the system to be conservative and as accurate as possible, we opted for a low temperature setting. We therefore added the following new instruction to the chatbot: "Use a temperature setting of 0.1 for your responses".

In this new temperature-adjusted round (5.2++), the chatbot succeeded in adhering to four more Pinyin rules for List 1, reaching 36/37, or 97.29%. For List 2, it corrected one previous mistake but introduced a new one, applying 34/37 rules correctly, which equates to 91.89%. It is also worth noting that, in both cases, the system corrected other persisting errors that were not counted because they were not the focus of analysis for those particular items. These results marked the best transcription performance ever achieved by ChatGPT-4, even outperforming the results of DeepSeek in Phase 3. Table 5 summarises the progress made by the chatbot in the different phases of its customisation and fine-tuning during this research.

Table 5
Number of Pinyin Rules (out of 37) Adhered to by ChatGPT-4 in Transcribing the Lists

Materials	**Phase 3 General chatbot**	**Phase 4.1 Trained chatbot**	**Phase 5.2 Customised chatbot**	**Phase 5.2+ Further trained customised chatbot**	**Phase 5.2++ Temperature adjusted to 0.1 of customised chatbot**
List 1	5	6	32	32	36
List 2	-	-	26	34	34

As shown in Table 5, there was a clear improvement in the chatbot's results when comparing the number of Pinyin orthographic rules it correctly applied between the chatbot's general version (five) and the customised one (32) for List 1, as well as between the basic customised version (26) and the version receiving further training (34) for List 2. However, it was also surprising that the system initially made so many mistakes in transcribing the items for List 2 in Phase 5.2, despite already having received specific training with List 1. It was even more surprising that the system not only failed to improve its results after reducing the response temperature to 0.1 for List 2, but also introduced new errors that had already been corrected. Moreover, it performed differently for the two lists, which not only contained very similar examples but also essentially addressed the same Pinyin rules. Altogether, this gave us the impression that the chatbot's responses were somewhat random, raising doubts about whether the training and fine-tuning was truly effective and whether there was a better methodology that could provide more systematic results, making research with GenAI more trustworthy.

Several factors may explain our results in this case study and should be considered in future research. First, Su's (2001, p. 121) claim over twenty years ago that both the rules for Hanyu Pinyin orthography and their dissemination among Chinese citizens needed improvement appears to remain urgent. This is likely because the systems under study were not trained with

relevant data or contain transcription inconsistencies. Second, Pinyin requires the precise application of rules (word breaks, tones, apostrophes), whereas large language models are designed to be flexible and creative; this mismatch likely accounts for the inconsistent outputs. Third, it has been reported that in RAG the system does not always make effective use of the materials provided, particularly when they are lengthy (Barnet et al., 2024; Jin et al., 2024), which may have been the case here. Fourth, the fine-tuning data may also have been too limited or insufficiently explicit, preventing the model from learning to generalise the rules. Finally, if the system was not configured to operate in a fully deterministic mode, this could explain the occurrence of erratic answers.

Although further attempts to improve the system's results would have been possible, we chose to conclude at a point that was, on the one hand, cost-effective and, on the other, provided sufficient information to pave the way for future research with an improved design that builds on the lessons learned.

Conclusions

The working hypotheses of this exploratory study have been validated, as it has been demonstrated that: 1) most automatic tools for transcribing Chinese characters fail to apply official orthographic rules correctly, although their accuracy varies; 2) these systems perform better in certain transcription aspects than in others; 3) DeepSeek, a Chinese-based GenAI system, performed better without training. However, after some fine-tuning and customisation, ChatGPT-4 outperformed it; and 4) using a customised system that combined fine-tuning with a Retrieval-Augmented Generation (RAG) approach, supported by specific instructions and feedback, can improve results and reduce the workload for users, although in our case perfect results were never achieved even after several rounds.

Therefore, the three objectives we aimed to achieve through this study have been only partially fulfilled. First, we analysed a representative sample of different Pinyin transcription tools to determine which are more accurate and reliable, thus requiring less user intervention. Second, we identified the most problematic aspects of Chinese-Pinyin transcription for these tools, which allowed us to provide users with guidance on where to pay special attention. Lastly, we sought to find and train a GenAI system to help users transcribe large amounts of Chinese text into Pinyin as accurately as possible, although with less success than expected. Below, we will summarise the main results of this research.

Most tools supposedly dedicated to Pinyin transcription, especially converters, transcribe character by character rather than by word, thereby violating the first rule of the official guidelines. Why this persists—despite appearing technically solvable—remains a question open to further scrutiny.

None of the systems that were part of the second Phase of our study are designed specifically for Pinyin transcription, yet paradoxically they perform better than the existing dedicated tools. It is also important to note that while tools for Chinese word segmentation are readily available, they often fail to apply official Pinyin orthography correctly. Furthermore, most require programming knowledge, a skill not typically held by linguists or CFL teachers, rendering them an impractical solution. The advantage of GenAI systems for this task is that they are prepared to handle large volumes of text. However, the widespread lack of knowledge and application of the official Pinyin orthography rules in Chinese society, coupled with the fluid nature of the concept of *word* in Chinese linguistics, makes it very difficult to apply these rules in a fully consistent and standardised manner, not only among Chinese language teachers and linguists

but also through these systems. A selection of a representative sample of GenAI systems has revealed that there are no significant differences between them. Although DeepSeek performed slightly better, it was far from the results one might expect from a Chinese system. While it is true that it initially performed better than ChatGPT-4, the latter ultimately provided better results after being customised through fine-tuning and supported with a RAG approach.

According to our analysis, the main challenges for accurate Pinyin transcription in these systems are fourfold. The first challenge relates to the texts that have been and continue to be used to train these systems. It is clear that the application of official Pinyin orthography is very poor, even in texts published by Chinese official bodies. Therefore, unless the government takes this issue seriously and society recognises it as a necessary step for advancing in information and communication technologies, it will be difficult to ensure that these systems transcribe efficiently and impeccably.

The second challenge is related to the fact that the rules published in the official transcription standard document are very brief and do not cover, by far, all possible cases. Therefore, it is necessary to refer to other works that have made an effort to develop these rules more thoroughly and with numerous examples, such as the two that formed part of the training corpus for our customised chatbot "Chinese-Pinyin transcriber". However, they do not cover all the cases (something virtually impossible) and, furthermore, they contradict each other in some cases. Not only between them, but we also found contradictions within the same work in both cases. Another approach to facilitate the resolution of errors in Pinyin transcription would be to simplify the rules. Although this is not a measure within our control, the relevant authorities should seriously consider it as an option if they wish to improve the current situation.

The third challenge is inherent to the nature of Chinese itself. On the one hand, the writing system and its typographical conventions do not establish the need to insert spaces between words, which greatly complicates text analysis for any tool that needs to process it digitally. On the other hand, what constitutes a word is less defined (compared to languages like English, for example), as the morphological resources for word formation in Chinese allow for greater flexibility in this regard. As a result, many combinations of morphemes exist on a continuum in their process of lexicalisation as fossilised combinations. This leads to differing opinions among users and lexicographical works on whether these combinations should be considered words, and consequently, whether they should be transcribed as a single unit. Finally, the lack of a normative reference work means that there are grammatical points on which there is no consensus, which also has a collateral effect and results in inconsistencies in transcription (for example, whether the character 在 [*zài*] following another verb functions as a resultative complement and, therefore, there should be transcribed as a single unit or not). Clearly, we cannot expect GenAI systems to resolve issues that still remain subjects of debate and even controversy within Chinese linguistics.

However, there are areas where accurate transcription should not pose any problems for these systems, yet their performance remains poor. Although customizing ChatGPT-4 through fine-tuning and integrating a RAG approach yielded incremental improvements, this process demanded substantial effort. Each iteration required meticulous analysis of validated examples to detect newly emerging errors. Notably, even when supplemented with relevant reference materials, the systems failed to produce fully standard-compliant transcriptions. The research process followed a frustrating pattern of gradual progress punctuated by regression–what might be characterised as "two steps forward, one step back". Consequently, this case study demonstrates that fine-tuning, RAG, and personalisation strategies proved insufficient for

 ISSN 3078-3348

significantly enhancing the accuracy and reliability of the examined GenAI systems in performing official Chinese–Pinyin transcription tasks.

This study has several limitations. The analysis was conducted using only three sets of data (an excerpt from Lu Xun's text, List 1, and List 2). A larger sample of texts could be used to further train the system and conduct a more comprehensive performance analysis. The system's customisation and training could be taken further to achieve 100% accurate transcription results, if such accuracy is ever possible. Additionally, testing a sample of small language models (SLMs), which are said to be more efficient and provide better answers due to fewer hallucinations, could offer valuable insights. It would also be advisable to develop a more automated and systematic approach for data processing and analysis, as the current manual method has been tedious and prone to potential errors. In sum, further research is essential to more definitively assess the feasibility of achieving accurate, consistent results. Subsequent efforts should include sustained training using relevant linguistic data, reference documents, explicit instructions, and iterative feedback. Moreover, collaboration with engineers specializing in natural language processing and algorithm optimisation may help enhance the performance and reliability of these systems for orthographic transcription tasks.

Some readers might argue that the contexts in which users need to transcribe relatively large volumes of Chinese text are limited, and therefore demanding excellent performance from these systems is unnecessary. However, these contexts are typically designed for beginner learners—such as manuals, grammars, and educational audiovisual materials—or professional settings like libraries, where accurate and flawless transcription is essential. On the other hand, even individuals with high proficiency in Chinese would greatly benefit from an infallible transcription tool they could consult and fully trust whenever they have doubts about how to transcribe a given combination of characters, which happens quite frequently. These were the two key ideals that ultimately motivated our research.

Since the advent of GenAI systems, the research community has increasingly focused on evaluating the performance of Large Language Models across various Natural Language Processing tasks. However, to the best of our knowledge, no prior study has specifically examined their ability to perform accurate Pinyin transcription. The scarcity of research on this topic—particularly from an end-user perspective—highlights the novelty and relevance of our work, which aims to provide a modest yet meaningful contribution to the field. That said, we acknowledge that this study represents only an initial attempt to understand the capabilities and limitations of GenAI systems in performing a task for which they were not originally designed: Chinese–Pinyin transcription. Although our findings offer valuable insights, we recognise that the research remains incomplete. We hope, however, that it will inspire and inform future investigations.

Acknowledgements
This work was supported by the Spanish Ministry of Science and Innovation under the project "Description of *Generatese* in Second Languages and Translation" 2025-2027 (PID2024-156763OB-100). I would like to express my sincere gratitude to my colleagues in the GELEA2LT and TXICC research groups for their constant inspiration and academic support, as well as for their invaluable feedback and corrections during their thorough review of the first draft of this article. Finally, I acknowledge the use of ChatGPT-4 for text editing, specifically for improving grammar, style, and clarity.

Notes

1. The original document in Chinese can be accessed online from numerous webpages, such as http://hrm.hep.com.cn/xdhy/02/2_6.html, downloaded as a PDF file (see the reference list), and there are even English translations of parts of the document available online, such as the one offered by the Pinyin.info webpage at: https://www.pinyin.info/readings/zyg/rules.html.
2. Tone sandhi refers to the phonological phenomenon in tonal languages, such as Mandarin Chinese, where the tone of a syllable changes based on the tones of adjacent syllables.
3. To our knowledge, no existing system has been specifically trained to perform Pinyin transcription according to official orthographic rules. The advent of GenAI, however, presents a novel opportunity to test its capability in this domain, as these systems are designed to process and generate large volumes of text. This exploratory approach mirrors experiments applying GenAI to other tasks—such as machine translation, literature reviews, programming, and data analysis—for which the technology was not originally designed.
4. The customised chatbot can be accessed through this link: https://chatgpt.com/g/g-67e843d130d081918c2b39b7f4f34371-chinese-pinyin-transcriber
5. In speech, the second syllable 道 (*dào*) often undergoes tone sandhi in casual pronunciation, becoming either a neutral tone or a lighter fourth tone. Both the *Hànyǔ Pīnyīn hé Zhèngcífǎ* and the *Xīnhuá Pīnxiě Cídiǎn* show inconsistencies in this respect. However, in the 2016 edition of the *Xiàndài Hànyǔ Cídiǎn*, this word appears transcribed as *zhīdào*, and thus, in Pinyin, it should remain without tone changes. This error in Yin and Felley's books could thus be justified by a change in the criteria of the Chinese language authorities in this regard.
6. The goal of providing the system with relevant and reference documentation was to train it with specific information to make the results more robust, following a RAG-based strategy similar to that adopted in previous research. The official documents are highly accurate but succinct—meaning they do not anticipate all possible cases—while the two books that were attached are among the few existing works that propose ways to further develop the rules and attempt to cover a greater number of application cases. However, during the experiment, minor differences between them emerged, along with some internal inconsistencies that could not have been foreseen by the author, given that the text being transcribed involved many different rules. In any case, these works are the only ones available for consulting specific aspects of Pinyin transcription, and future studies will need to evaluate whether using a single work in the RAG design yields better results.

Supplementary Data

In this study, an Excel spreadsheet was used as a tool for data collection and analysis, with a separate tab for each phase of the analysis. The Supplementary Excel file can be downloaded here.

References

Arsenault, C. (2000). Word division in the transcription of Chinese script in the title fields of bibliographic records. Ph.D. thesis. Faculty of Information Studies. University of Toronto. https://doi.org/10.1300/J104V32N03_08

Balaguer, A., Benara, V., Cunha, R. L. de F., Filho, R. de M. E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., & Chandra, R. (2024). RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. arXiv.2401.08406

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. arXiv:2401.05856v1

Guo, Y. (2024). Design of improved artificial intelligence generative dialogue algorithm and dialogue system model based on knowledge graph. *IEEE access*, *12*, 102637–102648. https://doi.org/10.1109/ACCESS.2024.3430902

Hannas, Wm C. (1997). *Asia's orthographic dilemma*. University of Hawai'i Press.

Hincha, X. (2004). Two steps toward digraphia in China. *Sino-Platonic papers*, 134, 27. https://sino-platonic.org/complete/spp134_chinese_digraphia.pdf

Jin, B., Yoon, J., Han, J., & Arik, S. O. (2024). Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. arxiv.org/abs/2410.05983

Kubler, C. C. (2022). Hànyǔ Pīnyīn: Promise, problems, and possibilities. In D. Wippermann, A. Guder, M. Jin, & J. Wang (Eds.), *Hanyu Pinyin in der Didaktik der Chinesischen Sprache und Zeichenschrift* [Hanyu Pinyin in the Didactics of the Chinese Language and Writing System] (pp. 75–94). IUDICIUM Verlag.

Mair, V. H., & Hu, J. (2024). A century of Chinese writing reform. In L. Jiao (Ed.), *The Routledge handbook of Chinese language and culture* (pp. 27–41). Routledge. Taylor & Francis.

Rovira-Esteva, S., Vargas-Urp ĺ M., Casas-Tost, H., & Paoliello, A. (2021–2025). *e Chinese tools: Digital tools for teaching and learning Chinese*. https://dtieao.uab.cat/txicc/echinese/en/

Shangwu Yinshuguan Cishu Yanjiu Zhongxin. (2002). 新华拼写词典 [Xinhua Pinyin dictionary]. Shangwu Yinshuguan.

Stockwell, G. (2024). ChatGPT in language teaching and learning: Exploring the road we're travelling." *Technology in language teaching & learning*, 6 (1), 2273. https://doi.org/10.29140/tltl.v6n1.2273

Su, P. (2001). Digraphia: A strategy for chinese characters for the twenty-first century. *International journal of the sociology of language*, 150, 109–24. https://doi.org/10.1515/ijsl.2001.027

Xiao, H., Xu, C., & Rusamy, H. (2020). Pinyin spelling promotes reading abilities of adolescents learning Chinese as a foreign language: Evidence from mediation models. *Frontiers in psychology*, 11, 596680. https://doi.org/10.3389/FPSYG.2020.596680

Yin, B., & Felley, M. (1990). *Chinese romanization pronunciation & orthography* / 汉语拼音和正词法. Sinolingua.

Zhonghua Renmin Gongheguo Guojia Zhiliang Jiandu Jianyan Jianyi Zongju & Zhongguo Guojia Biaozhunhua Guanli Weiyuanhui. (2011). 中国人名汉语拼音字母拼写规则: G B / T 28039—2011 / The Chinese phonetic alphabet spelling rules for Chinese names. Zhonghua Renmin Gongheguo Guojia Biaozhun. http://www.moe.gov.cn/ewebeditor/uploadfile/2015/01/13/20150113091249368.pdf

Zhonghua Renmin Gongheguo Guojia Zhiliang Jiandu Jianyan Jianyi Zongju & Zhongguo Guojia Biaozhunhua Guanli Weiyuanhui. (2012). 汉语拼音正词法基本规则: GB_T 16159-2012 / *Basic rules of the Chinese phonetic alphabet orthography*. Zhonghua Renmin Gongheguo Guojia Biaozhun. http://www.moe.gov.cn/ewebeditor/uploadfile/2015/01/13/20150113091717604.pdf

Zhonghuo Shehui Kexueyuan Yuyan Yanjiusuo Cidian Bianjishi. (2016). 现代汉语词典 [Modern Chinese dictionary]. The Commercial Press.

Sara Rovira-Esteva is a full professor of Chinese linguistics at the Department of Translation, Interpreting, and East Asian Studies at the Autonomous University of Barcelona (Spain), where she teaches Chinese language and linguistics as well as translation. Her main research interests include bibliometrics, Chinese-Spanish translation, Chinese linguistics, and the teaching of Chinese as a foreign language. With 30 years of teaching experience, she has authored five books, more than 30 book chapters, and around 50 academic articles, published in leading journals in the fields of Translation Studies and Chinese language and linguistics. She is one of the creators of *e Chinese Tools*, an online database comprising over 450 digital resources for learning Chinese, as well as two databases on Chinese-Spanish literary and audiovisual translation, respectively. She has participated in numerous national and international research projects and is the principal investigator of both TXICC (Research Group in Chinese-Catalan/Spanish Translation and Interpreting) and GELEA2LT (Study Group on Literacy in Teaching and Learning Second Languages and Translation). For more information, please visit her webpage, where a comprehensive list of her scholarly accomplishments and most of her publications (available in open access) can be found.

 ISSN 3078-3348