Between the Lines: A Comparative Analysis of AI vs. Human Transcription of Lin Yutang's Bilingual Handwritten Notebooks

Guoying Yang[1,2]

[1] School of Foreign Languages, JiMei University, China
[2] Department of Asian and Middle Eastern Studies, University of North Carolina at Chapel Hill, USA

Abstract

Handwritten Text Recognition (HTR) has been applied to digital humanities at high speed to improve the accessibility of historical manuscripts; however, research on how well HTR handles bilingual and code-switched materials, which can affect interpretation, is currently lacking. The above deficiencies are especially severe for cross-cultural archival collections; specifically, only 1.6 per cent of existing HTR research has focused on Chinese-English mixed-script materials, and none have used a controlled human comparison system with domain-expert transcribers. This paper presents the first controlled, empirically supported comparison of HTR and expert human transcription for Lin Yutang's bilingual (Chinese-English) handwritten notebooks. A targeted review of recent SSCI-index studies on HTR evaluation (2018–2025) was conducted. This was followed by a $2 \times 2 \times 3$ controlled comparative experiment in which 50 purposively selected manuscript pages (12,847 characters, 38.2% code-switching) were transcribed by three HTR systems (Transkribus base, Transkribus fine-tuned, and a transformer-based model) and three expert human transcribers with complementary expertise in paleography, Lin Yutang studies, and digital humanities. Transcriptions were coded and analyzed across four dimensions: accuracy, error typology (5 main categories, 17 subcategories), interpretive fitness (4 dimensions rated by a blind panel of three Lin Yutang scholars), and efficiency (via 10,000-iteration Monte Carlo simulation). Findings show that: (1) HTR systems have achieved a mean CER of 17.84% compared to 3.18% for humans ($\eta_p^2 = .80$, $p < .001$, $d = 2.84$), and performance drops significantly on code-switching passages (CER increases to 32.83%); (2) errors in HTR are mainly "semantic-blind" (cross-script confusion: 13.1%; code-switching boundary misrecognition: 22.2%; semantic-contextual errors: 19.5%), whereas human errors are primarily "mechanical" (punctuation: 22.5%; cross-linguistic normalization: 32.9%); (3) HTR interpretive fitness scored 2.84/5, lower than 4.74/5 for humans ($r = 0.72$, $p < .001$); (4) only 42.6% accuracy was achieved by HTR on culturally significant terms (e.g., "Christo-Confucian" was never recognized correctly) compared to 100% for humans; and (5) a hybrid workflow of HTR + confidence-based human post-correction reached near-human performance (2.14% CER) at 66% lower cost and 83% higher throughput. This paper presents the first controlled HTR-human comparison of Chinese-English bilingual manuscripts, proposes a distinction between semantic-blind and mechanical errors based on error analysis theory, develops an Interpretive Fitness Audit framework for assessing transcription quality beyond aggregate CER, and shows that algorithmic transcription was more likely to misrecognize neologisms and code-switching boundaries, thus demonstrating that automated transcription may contribute to the loss of culturally significant meanings and code-switching boundaries embedded in bilingual manuscripts.

Keywords: Handwritten Text Recognition, bilingual transcription, code-switching, Lin Yutang, digital humanities, interpretive fitness, error typology

Guoying Yang. "Between the Lines: A Comparative Analysis of AI vs. Human Transcription of Lin Yutang's Bilingual Handwritten Notebooks." In: *From Manuscript to Megabyte: Building Sustainable and Open Futures in the Digital Humanities*, ed. by James B. Harr III, Emma Stanley, Jenna Rinalducci, and Donna Kain. Vol. 5. Anthology of Computers and the Humanities. 2025, 22–37. https://doi.org/10.63744/Rxt7eQHuSmL6.

1 Introduction

The development and spread of Handwritten Text Recognition (HTR) technology have changed the course of study in digital humanities and archival science in recent years. Transkribus, eScriptorium, and OCR4all are platforms that have conducted large-scale automated text recognition and analysis to reduce the labour of manual transcription for historical manuscripts significantly [11; 14]. For the 2026 Digital Humanities Institute, which has chosen the theme "From Manuscript to Megabyte: Building Sustainable and Open Futures in the Digital Humanities", the role of digital tools in connecting traditional and new media has been put forward explicitly. HTR systems offer a direct technological solution to the problem of sustainable knowledge production in an era of declining public investment in the humanities.

However, there is also a deep-seated problem in both method and ethics. Recently, research has been conducted on handwritten text recognition (HTR) and optical character recognition (OCR) to partially automate the transcription of historical documents. Reul et al. [18, p. 4853] have shown that historical texts can achieve a character error rate of less than 10% and as low as 5% after model training on a particular collection. Muehlberger et al. [11, p. 955] have also shown that HTR technology can "fundamentally change the mode of use and research for archives". Most of the published applications have focused on relatively homogeneous manuscript collections and standardized handwriting datasets. Research using HTR for bilingual, code-switched and highly heterogeneous manuscript materials has been relatively sparse in published works. Therefore, many fundamental problems in transcription accuracy and interpretation reliability remain to be addressed [14]. If the errors produced by the OCR and HTR systems are not promptly addressed, they will be retained in the digital surrogate, thereby harming scholarly analysis, interpretation, and reuse of the work. As Nockels et al. [14] have noted, the shortcomings of automated transcription technology are not always known to users, and thus errors in machine-generated data may be propagated in subsequent research processes. Thus, the transcription errors are not merely technical deficiencies; they may also influence how historical documents are interpreted, cited, and reused in future scholarship. Due to HTR's inherent extensibility, the consequences of such errors will be more severe. A single misrecognition by a person is corrected through peer review or editing; however, when applying an HTR system to a corpus of 10,000 pages, thousands of undetected errors may occur, each potentially significantly deviating from the original historical record.

A targeted review of recent SSCI-indexed studies published between 2018 and 2025 was conducted to identify major trends and gaps in HTR evaluation research. First, there is a problem in the method for evaluating HTR systems. Most previous studies have focused on quantifying the accuracy of speech-to-text conversion using metrics such as Character Error Rate (CER) and Word Error Rate (WER). However, few have investigated the impact of transcription errors on research interpretation [14]. The field does not yet have a system for assessing what we call "interpretive fit", that is, the degree to which a transcription preserves semantic, cultural and contextual meanings for academic use. When an HTR system replaces "Christo-confusion" with "Christo-Confucian", it is counted as only a few character substitutions under the standard CER metric; however, the interpretive loss resulting from the erasure of a deliberate philosophical neologism goes unnoticed by this metric. There is also a linguistic gap in existing HTR research. Most published studies focus on linguistically homogeneous corpora, particularly manuscripts written in a single language or script. Comparatively little attention has been paid to multilingual and code-switching materials, despite their importance in many historical and cross-cultural archives. Research on Chinese–English mixed-script manuscripts remains especially limited, and studies examining bilingual handwritten materials are still rare.

A related limitation concerns human comparison. Although some HTR studies include human transcribers as benchmarks, the expertise of these transcribers is often not discussed in detail. Comparisons involving domain experts with specialized knowledge of the manuscript's historical,

linguistic, or intellectual context remain uncommon. Likewise, few studies have systematically examined how differences between machine and human transcription affect the interpretive quality of the resulting texts.

Lin Yutang (林语堂, 1895–1976) is one of the people who have recorded the problems and interests in bilingual transcription to date. Lin is a highly renowned Chinese-American writer, translator, inventor, and public figure, skilled in both language and culture. His best-selling English-language works include *My Country and My People* (1935), *The Importance of Living* (1937), and *The Wisdom of China* (1944); these have introduced Chinese philosophy and aesthetics to many Westerners. His surviving handwritten notebooks are an excellent primary source located at the University of Southern California (Collection 0302, Boxes 4–7, Folders 12–28); there are about 340 manuscript pages from 1932 to 1966, which exhibit dense code-switching and frequent multi-switches within a single line; idiosyncratic handwriting includes both careful kaishu (regular script) and rapid xingshu (running script); there is specialized and neologistic vocabulary such as "Christo-Confucian", "philosophic Tao", and "co-humanity"; and cross-linguistic semantic play that requires the simultaneous understanding of both languages. Lin regards Ren as "co-humanity" rather than the traditional notion of benevolence and thus wishes to express the relation between people in this Confucian idea. Christo-Confucian proposes a special concept of the combination of Chinese and Christian culture in his thought. Lin Yutang's bilingual notebooks offer evidence of a long-term process of translation, revision and cross-cultural communication; they can be understood as working documents that preserve traces of textual revision, experimentation, and linguistic decision-making. This perspective is consistent with genetic criticism, which treats drafts, notebooks, and related manuscript materials as evidence of textual genesis and the evolving process of composition rather than merely as preliminary versions of a finished work [4].

Based on the above considerations, the seven research questions for this study are as follows. **RQ1** (Broader Landscape): What are the publication trends, methodological attributes, and theoretical foundations of research on HTR evaluation in SSCI-indexed journals from 2018 to 2025? **RQ2** (Methodological Practice): What sample size and characteristics of participants, what instruments, and how did the data analysis method differ between studies comparing HTR and human transcription? **RQ3** (Comparison Performance): How do the three HTR models perform relative to the accuracy of character-level and word-level transcription by three expert human transcribers? **RQ4** (Error Typology): What qualitatively different types of errors appear in machine- and human-transcribed text, and how are they distributed among dimensions such as character confusion, punctuation, code-switching boundary misrecognition, line-break handling, and semantic-contextual errors? **RQ5** (Interpretive Fitness): To what extent do errors in machine-based and human transcription affect interpretative fitness, as judged by a panel of Lin Yutang scholars? **RQ6** (Hybrid Workflow Optimisation): Among various combinations of HTR initial transcription and targeted human post-correction, which combination can best achieve an ideal balance among accuracy, cost, speed and interpretability? **RQ7** (Cross-Cultural Significance): How do HTR systems perform specifically on passages related to Lin Yutang's cross-cultural interpretation of terms such as Ren (仁), Dao (道), li (礼), and his neologisms "Christo-Confucian" and "co-humanity"?

2 Literature Review

2.1 The Evolution of HTR Evaluation

Although the system for assessing HTR has developed in recent years, this progress has not been uniform. The main index remains character error rate (CER); that is, the proportion of the edit distance between the transcription and the ground truth relative to the total number of characters. CER is simple to compute and relatively easy to compare with other studies, but it does not distinguish among all types of errors by severity. A substitution that changes "Confucian" to "Confusion" is

assigned the same edit cost as one that changes “Confucian” to “Confucioñ” – the first is fundamentally different in meaning, and the second is clearly an error.

Rather than considering general measures of researchability for scholars, such as CER and WER, which serve this purpose, this paper investigates how specific transcription errors affect data interpretation and subsequent studies. Therefore, we will conduct studies on how HTR is changing the information environment to fill this deficiency identified by Nockels et al. [14]. Multilingual and code-switching manuscripts present recognition challenges that differ substantially from those found in linguistically homogeneous corpora. In such materials, transcription systems must process not only handwriting variation but also language alternation, script shifts, and context-dependent lexical interpretation. As a result, some recognition errors may remain linguistically plausible even when they alter meaning, making them difficult to identify through aggregate accuracy measures alone [14]. However, no existing taxonomy has addressed the specific problems of Chinese-English bilingual transcription: cross-script visual interference (e.g., mistaking the English digraph “rn” for the Chinese character 川), misdetection of code-switching boundaries, neologism handling, and the preservation of cross-linguistic semantic play.

Some scholars have proposed adjustments to the standard indicators. Recently, many studies have asked whether the volume of aggregate indicators, such as Character Error Rate (CER) and Word Error Rate (WER), is appropriate for reflecting actual use conditions. Holley [7] thinks that, in addition to the accuracy statistics, they may not be convenient for users to search for, retrieve and use in their studies. The defect is more severe in bilingual manuscripts; a small transcription error may alter the context or the selection of translation in the other language. Building on concerns about the limitations of aggregate accuracy measures, this study develops a weighted CER approach to better reflect the interpretive consequences of different transcription errors.

2.2 Theoretical Frameworks for Error Analysis

To move beyond descriptive error classification toward a theoretically grounded understanding of machine versus human transcription, this study draws on insights from second language acquisition research. Corder [3] distinguished between competence errors, which reflect incomplete knowledge of an underlying rule system, and performance errors, which arise from temporary processing failures despite adequate knowledge of that system. Subsequent work in language learning has similarly emphasized that errors reveal systematic limitations in competence, whereas mistakes often result from attention lapses, fatigue, or other performance-related factors [10]. This distinction provides a useful framework for comparing machine-generated and human-generated transcription errors. We adapt this distinction to characterize HTR errors as analogous to “competence errors” — they are systematic, reflecting limitations of the trained model — while human errors are analogous to “performance errors” — unsystematic and correctable with proofreading. From translation studies, House’s [8; 9] model of Translation Quality Assessment distinguishes between overtly erroneous errors, which are directly observable in the translated text, and covertly erroneous errors, which may appear acceptable on the surface but become apparent when compared with the source text or evaluated in relation to communicative function. This distinction is particularly useful for transcription analysis because some transcription errors are immediately visible, whereas others preserve surface plausibility while subtly altering meaning. We adopt this framework to characterize HTR errors as predominantly “covert” — they silently alter meaning without surface indication — while human errors are predominantly “overt” — immediately visible and easily corrected. From hermeneutic philosophy, Gadamer’s concept of the “fusion of horizons” suggests that understanding emerges from the interaction between the interpreter’s historical horizon and the text’s. Interpretation is therefore shaped by prior knowledge, assumptions, and historically situated perspectives rather than being a purely objective process [6]. We invoke this framework to argue that HTR systems lack the “horizon” necessary for interpretive understanding — they

process characters without recognizing their semantic or cultural significance, making them particularly vulnerable to errors on precisely those passages where cultural and contextual knowledge is most needed.

2.3 Transparency and Trust in AI-Assisted Cultural Heritage

HTR technology is often used in an all-automatic system that is not very transparent to the public. As Nockels et al. [14] have noted, relatively little research has examined how HTR systems construct the overall information environment for generating and using digitized texts. This opacity is especially prominent when machine-generated transcriptions are used in digital archives, scholarly editions, and downstream research, and transcription errors may be difficult to detect or correct. Owens [15] believes that digital preservation is an interpretation carried out by people under the direction of preservation goals and institutional choices. Therefore, the producers of digital replicas are also accountable for how these digital versions may affect future access to and interpretations of cultural heritage materials. In response to the above demands, this study has provided fine-grained error profile data and a reproducible interpretation fitness audit protocol that all institutions can apply.

2.4 Lin Yutang and Bilingual Transcription as a Digital Humanities Problem

Lin Yutang's bilingual works have been studied in detail from the perspective of literary and translation studies, but not often from the standpoint of digital humanities. Scholars have often described Lin Yutang as a cross-cultural mediator whose English-language works aimed to introduce Chinese culture to Westerners while respecting its own cultural characteristics. Lin often did not just translate the meaning; rather, he reformulated it and presented new contexts for Chinese ideas to be accepted by the new generation of readers, building a bridge between Chinese and Western thought [16; 17]. His bilingual notebooks provide a good case for exploring the emergence of such acts of cultural mediation in the drafting process and how linguistic choices are negotiated in both languages. The above studies indicate that Lin's manuscripts, which are the workshop where he developed his cross-cultural vocabulary, are also of particular interest for studying the process of cultural translation and intellectual synthesis. Computationally, one can discover the patterns that traditional close reading cannot: how frequently and where code-switching occurs over time, how translation strategies have changed, and the connection between notebook drafts and published works. The above research opportunities cannot be realized without accurate and scalable transcription.

3 Method

3.1 Research Design

This study used a 2 (Transcriber Type: HTR vs. Human) $\times$ 2 (Language Condition: Monolingual vs. Code-Switching) $\times$ 3 (Transcriber Instance: Three HTR Models / Three Human Transcribers) controlled-comparative experimental design, and a mixed-methods approach combining quantitative accuracy indicators and qualitative interpretative fitness assessment.

To contextualize the experimental findings, a targeted literature review of recent SSCI-indexed studies on HTR evaluation published between 2018 and 2025 was conducted. The review was intended to identify broad methodological trends and research gaps rather than to provide a comprehensive systematic review.

3.2 Corpus Construction and Sampling

The corpus consists of 50 purposively selected manuscript pages from Lin Yutang's bilingual handwritten notebooks in Collection 0302, Special Collections, Doheny Memorial Library, University of Southern California (Boxes 4–7, Folders 12–28). About 340 manuscript pages have been compiled in the corpus from 1932 to 1966. Fifty pages were selected by stratified purposive sampling according to three dimensions: language distribution (17 predominantly Chinese, 17 predominantly English, 16 balanced bilingual/code-switching with ≥ 5 code-switching instances per page); handwriting style (25 careful kaishu/neat English print, 25 rapid xingshu/connected English cursive); and content type (15 essay/translation drafts, 10 personal reflections, 10 vocabulary lists, 10 translation exercises, 5 experimental neologisms). The Bookeye 5 V2A planetary scanner was used to digitize the pages at 600 dpi, saving them as uncompressed TIFF files in 24-bit colour. The corpus has about 12,847 characters (mean $= 256.94$ per page, SD $= 78.32$), which include 6,431 Chinese characters (50.1%), 5,892 English characters (45.8%), and 524 punctuation marks and special symbols (4.1%). Among all the content, 38.2% was in a code-switching area (defined as within ± 5 characters of a language boundary).

3.3 HTR Systems

Three HTR systems were selected to represent the two groups. **HTR-Base**: Transkribus version 2.7.2 with the "Historical Handwriting Model M1" (model ID: HTR-EN-2023-04), trained on 1.2 million characters from 18th–19th century English manuscripts, no fine-tuning, an "off-the-shelf" deployment scenario. **HTR-FT**: Transkribus version 2.7.2 with a custom model fine-tuned on 30 pages of Lin Yutang's handwriting (excluded from the test set), 100 epochs, learning rate 0.0001, batch size 8, validation split 20%, final validation CER: 12.4%. **HTR-TF**: Calamari version 2.1.0 with a Vision Transformer encoder (12 attention heads, 6 layers), trained on the same 30 fine-tuning pages, learning rate 0.00005, warm-up steps 500, Adam optimizer, label smoothing 0.1. All of the above systems are based on NVIDIA RTX 3090 GPUs with 64 GB of RAM.

3.4 Human Transcribers

Three expert transcribers with complementary expertise in paleography, Lin Yutang studies, and digital humanities participated in the transcription task. Each received a 12-page style guide, one practice page and written instructions on diplomatic transcription. Randomization of the pages per transcriber was carried out according to a Latin Square Design to reduce order effects.

3.5 Error Coding Scheme

A whole-inclusive error coding system with five general categories and seventeen subcategories was constructed for this study by iteratively analyzing the corpus. The coding framework has used research on error analysis, translation studies and multilingual handwritten text recognition to add categories for Chinese-English bilingual manuscript transcription. **Category 1: Character Confusion Errors** — 1a: within-script Chinese (e.g., 己/已/巳, 没/设, 日/曰); 1b: within-script English (e.g., rn/m, cl/d, ii/u, ri/n); 1c: cross-script (Chinese character recognized as English or vice versa, e.g., English "rn" recognized as Chinese 几). **Category 2: Punctuation and Formatting Errors** — 2a: missing punctuation; 2b: added punctuation; 2c: replaced punctuation (Chinese vs. English comma, full-width vs. half-width); 2d: spacing/indentation errors. **Category 3: Code-Switching Boundary Errors** — 3a: boundary misdetection (fail to identify a language switch); 3b: boundary extension (switch detected at the wrong position, e.g., 3–5 characters too early or too late); 3c: language misassignment (text is correct, but the wrong language is assigned). **Category 4: Line-Break and Layout Handling Errors** — 4a: incorrect line-break detection; 4b:

margin note handling errors; 4c: character ordering errors (reading direction confusion in mixed-script contexts). **Category 5: Semantic-Contextual Errors** — 5a: correct character in the wrong context (e.g., "ren" recognized but not identified as a philosophical concept); 5b: neologism substitution (neologism replaced with a high-frequency alternative); 5c: homograph confusion (same character in different languages); 5d: cross-linguistic semantic loss (bilingual wordplay lost). Two trained research assistants, each having received 20 hours of training using a gold-standard corpus of 10 pages, independently coded the error instances. Inter-coder reliability: Cohen's $\kappa \geq 0.80$ for all categories.

3.6 Interpretive Fitness Audit

A panel of three Lin Yutang scholars (not including the human transcribers) was recruited: a professor of Chinese literature with 25 years of Lin Yutang scholarship, a professor of translation studies with 18 years of research on bilingual Chinese-English literature, and a senior researcher specializing in modern Chinese intellectual history. A total of 30 selected passages (10 culturally specific idioms, 10 neologisms/wordplay and 10 code-switching passages) were evaluated by the panel based on a four-part rubric for Semantic Preservation, Contextual Accuracy, Cross-Linguistic Coherence and Scholarly Usability, each at a 5-point Likert scale (1 = severe loss of meaning, 5 = fully preserved). The panel did not know who transcribed the materials and were presented them without reference.

3.7 Hybrid Workflow Simulation

Monte Carlo simulation (10,000 iterations) was used to simulate the practical effect and, based on empirical data, to determine the parameters of the five hybrid workflow strategies for HTR CER, such as HTR CER (17.84%), human CER (3.18%), HTR cost ($0.05/page), human transcription ($25/page), human post-correction ($15/page), and processing time. The simulations included HTR-only, human-only, HTR with full human post-correction, HTR with review of code-switching pages only, and HTR with confidence-based human correction of segments with a confidence threshold below 0.7.

3.8 Data Analysis

The first analysis was a two-way mixed ANCOVA with Transcriber Type and Language Condition as fixed factors, Transcriber Instance (nested under Transcriber Type) as a random factor, and Page Difficulty (total character count) as a covariate. Tukey's HSD post-hoc comparisons were performed. Error type distributions were examined using chi-square tests and standardized residuals. Wilcoxon signed-rank tests with Bonferroni correction were used to compare interpretability fitness scores. All the analyses above were conducted in R version 4.3.2 using the packages `lme4`, `emmeans`, `irr`, `rstatix`, and `ggplot2`.

4 Findings

4.1 RQ1–2: Scoping Review Findings

A review of recent HTR evaluation studies suggests three recurring patterns. Most published research focuses on monolingual manuscript collections and relies primarily on aggregate accuracy measures such as CER and WER. Multilingual and code-switching manuscripts remain underrepresented in the literature, particularly Chinese–English mixed-script materials. Although some studies include human transcription as a benchmark, comparisons involving domain experts with specialized knowledge of manuscript content remain relatively uncommon.

4.2 RQ3: Comparative Performance — Accuracy

Table 1 shows the overall CER and WER for all transcribers across the entire 50-page corpus.

Transcriber	Mean CER (%)	SD CER	Mean WER (%)	SD WER	Median Time/Page (min)
HTR-Base	22.47	8.34	35.12	11.28	0.47
HTR-FT	16.83	6.91	28.45	9.87	0.52
HTR-TF	14.21	5.76	24.83	8.64	0.89
HTR Combined	**17.84**	—	**29.47**	—	**0.63**
Transcriber A	3.28	1.92	7.14	3.45	18.7
Transcriber B	2.14	1.34	4.87	2.68	22.3
Transcriber C	4.11	2.23	8.96	4.12	14.9
Human Combined	**3.18**	—	**6.99**	—	**18.6**

Table 1: Overall CER and WER by transcriber.

The two-way mixed ANCOVA showed a significant main effect of Transcriber Type on CER, $F(1, 47) = 187.43$, $p < .001$, $\eta_p^2 = .80$, indicating that transcriber type accounts for 80% of the variation in accuracy. Human transcribers (combined mean CER $= 3.18\%$) performed significantly better than HTR systems (combined mean CER $= 17.84\%$), with a mean difference of 14.66 percentage points (Cohen's $d = 2.84$, 95% CI $[2.31, 3.37]$).

Among the HTR systems, HTR-TF had a lower CER (14.21%) than HTR-FT (16.83%) and HTR-Base (22.47%). All pairwise differences were statistically significant (Tukey HSD, $p < .01$). Among the human transcribers, Transcriber B (content expert, 2.14%) performed the best, followed by Transcriber A (paleographer, 3.28%) and Transcriber C (DH scholar, 4.11%). B and C were significantly different ($p = 0.003$); B and A were not statistically different ($p = 0.062$).

Table 2 is the CER broken down by language condition.

Transcriber	English-Only	Chinese-Only	Code-Switching
HTR-Base	16.34 (6.12)	28.58 (9.87)	38.92 (12.45)
HTR-FT	11.87 (5.04)	21.34 (8.21)	31.67 (10.83)
HTR-TF	9.63 (4.28)	18.72 (7.14)	27.89 (9.56)
HTR Combined	**12.61**	**22.88**	**32.83**
Transcriber A	2.41 (1.56)	4.12 (2.34)	5.87 (3.12)
Transcriber B	1.72 (1.08)	2.54 (1.67)	3.89 (2.41)
Transcriber C	3.08 (1.89)	5.13 (2.78)	7.24 (3.89)
Human Combined	**2.40**	**3.93**	**5.67**

Table 2: Mean CER (%) by language condition and transcriber. Standard deviations in parentheses.

The interaction between Transcriber Type and Language Condition was significant, $F(2, 94) = 42.67$, $p < .001$, $\eta_p^2 = .48$ (medium-to-large effect). CER for the HTR system increases by 160% in code-switching areas compared with English-only (12.61% to 32.83%). The absolute increase is only 3.27 percentage points for humans and 20.22 for HTR. These results suggest that HTR performance declined substantially as the frequency of code-switching increased; the systems enter a "confusion regime", and their performance becomes unstable (SD $> 12\%$ in code-switching areas). Human transcribers do not have this threshold.

4.3 RQ4: Error Typology

A total of 4,218 individual errors were coded. Table 3 is the distribution.

Error Category	HTR ($n = 2{,}912$)	Human ($n = 1{,}306$)	χ^2	p
1. Character Confusion	1,163 (39.9%)	312 (23.9%)	98.43	$< .001$
1a: Within-script Chinese	434 (14.9%)	129 (9.9%)		
1b: Within-script English	347 (11.9%)	168 (12.9%)		
1c: Cross-script	**382 (13.1%)**	**15 (1.1%)**		
2. Punctuation/Formatting	321 (11.0%)	294 (22.5%)	87.21	$< .001$
3. Code-Switching Boundary	**647 (22.2%)**	**34 (2.6%)**	**251.36**	$< .001$
4. Line-Break/Layout	213 (7.3%)	189 (14.5%)	54.28	$< .001$
5. Semantic-Contextual	568 (19.5%)	477 (36.5%)	137.84	$< .001$
5a: Correct char, wrong context	382 (13.1%)	28 (2.1%)		
5b: Neologism substitution	123 (4.2%)	11 (0.8%)		
5c: Homograph confusion	45 (1.5%)	8 (0.6%)		
5d: Cross-lingual semantic loss	**18 (0.6%)**	**430 (32.9%)**		

Table 3: Distribution of error types by transcriber type.

The chi-square test showed a significant association between transcriber type and the distribution of error types, $\chi^2(4) = 347.82$, $p < .001$, Cramér's $V = 0.42$.

Cross-script confusion (Category 1c): HTR produced 382 instances (13.1% of errors), while humans produced only 15 (1.1%). A typical case is that the English digraph "rn" is frequently recognized as the Chinese character 儿 (jǐ, "a few") when it occurs next to Chinese characters. "rn" and 儿 look very similar at some resolutions, so machines are confused by them; given the language context, humans are unlikely to be confused. Another case: the English trigram "cli" has been occasionally misread as the Chinese character 刂 (a radical/component) in fast writing.

Code-switching boundary misrecognition (Category 3): HTR committed 647 errors (22.2%) compared with 34 for humans (2.6%). Detailed analysis showed that 68% of the HTR code-switching errors involved extending the English language model beyond the actual boundary, so subsequent Chinese characters were recognized as "English-like" sequences (e.g., ren → "ren2" or "jen"). We call this a "language momentum error": after being assigned a language, the HTR model requires strong visual evidence for a switch, and in ambiguous areas, it tends to continue using the previously detected language.

Cross-linguistic semantic loss (Category 5d): there is a significant difference in the proportion of errors; humans produced 430 errors (32.9%), while HTR only produced 18 (0.6%). However, the vast majority of human "errors" in this category were normalization decisions, that is, the transcriber standardized Lin's unique spelling or punctuation to the norm. For example, Lin's variant spelling "co-humanity" was transcribed correctly by HTR but was occasionally standardized to "co-humanity" (with a standard hyphen) by humans. Although they did not meet the ground-truth standards, these "errors" were sometimes conducive to the system's comprehension.

Based on the error patterns in this study, the causes of errors fall into two categories: insufficient human context-reading and transcription errors. HTR errors were more likely to be cross-script confusion, boundary misrecognition of code-switching, and semantically significant substitutions; human errors more frequently included punctuation errors, line-break handling issues, and cross-linguistic normalization problems. This pattern aligns with what House [8; 9] identified as "overtly erroneous errors" and "covertly erroneous errors"; the former are easily noticeable, while the latter may appear acceptable at first glance but are, in fact, contrary to the original text.

4.4 RQ5: Interpretive Fitness

Table 4 is the Interpretive Fitness Audit results.

The composite mean gap between the best HTR (2.84/5) and the best human performance

Dimension	Best HTR (HTR-TF)	Worst HTR (HTR-Base)	Best Human (B)	p	r
Semantic Preservation	3.12 (0.87)	2.18 (0.94)	4.83 (0.38)	$< .001$	0.71
Contextual Accuracy	2.87 (0.92)	1.94 (0.88)	4.71 (0.46)	$< .001$	0.68
Cross-Linguistic Coherence	2.41 (1.04)	1.67 (0.85)	4.62 (0.51)	$< .001$	0.74
Scholarly Usability	2.96 (0.91)	2.03 (0.92)	4.78 (0.42)	$< .001$	0.72
Composite Mean	**2.84**	**1.96**	**4.74**	—	—

Table 4: Interpretive Fitness Audit: mean scores (1–5 scale). Standard deviations in parentheses; Wilcoxon signed-rank test.

(4.74/5) is 1.90 points. Wilcoxon rank-sum test results: $W = 476.5$, $p < 0.001$, $r = 0.72$ (large effect). The highest discrepancy was in Cross-Linguistic Coherence (2.21 points), and HTR often combined Chinese and English into non-interpretable hybrid strings. Inter-rater reliability analysis among the three panelists using ICC(2,1) showed a relatively good level of agreement: 0.82 for HTR transcriptions and 0.76 for human transcriptions.

Qualitative panel comments pointed out the high interpretative cost. For HTR transcription: "The characters are correct, but the rhythm of the bilingual alternation — which is the whole point of this passage — has been completely lost" (Panelist 2); "A common word has replaced this neologism. The reader would not know that Lin was inventing the term here" (Panelist 1); "It reads like poor machine translation, not as carefully crafted writing by Lin" (Panelist 3). For human transcription: there is no space here, but it makes sense; a quick proofreading could have caught this, and the transcriber was aware of Lin's wordplay, so that understanding was also present (Panelists 1 and 2).

A modified Character Error Rate (WCER) was proposed in this paper to account for differences in the weights assigned to different error types. Errors in content words, proper nouns and semantically important components were given a weight of 5; punctuation and formatting errors were assigned a weight of 2; minor orthographic deviations were assigned a weight of 1. The weight scheme in this study has been adjusted to account for the diverse consequences of transcription errors in academic research.

Transcriber	Standard CER (%)	Weighted CER (%)	Ratio (WCER/CER)
HTR Combined	17.84	33.96	**1.90**
Human Combined	3.18	5.16	**1.62**

Table 5: Weighted CER (WCER) by transcriber type.

The WCER/CER ratio of HTR (1.90) is significantly higher than that of humans (1.62); therefore, most of the errors in HTR are related to the content of meaning. According to disaggregated data across languages, the proportion of HTR in code-switching areas is 2.31; therefore, these errors are both more frequent and more severe per occurrence.

4.5 RQ6: Hybrid Workflow Optimization

The optimal strategy was HTR initial transcription with targeted human correction of segments identified as high-uncertainty (confidence < 0.7, covering about 34% of the content), and this achieved a final CER of 2.14% at $8.45 per page (66% cheaper than full human transcription) and 3.12 minutes per page (83% faster). Reviewing only code-switching sections was insufficient; many mistakes in the Chinese-only sections had a high HTR CER of 22.88%. A confidence-based strategy identified errors in all language conditions. Institutional adoption: discovery-level digitization ($\sim$5% CER tolerance) — Strategy 4; scholarly editions ($<$ 3% CER) — Strategy 5;

Strategy	Final CER (%)	Cost per Page ($)	Time per Page (min)
HTR only	17.84	0.05	0.63
Human only	3.18	25.00	18.60
HTR + full human post-correction	1.20	15.05	4.20
HTR + review of code-switching pages only	4.87	5.75	1.99
HTR + confidence-based review (< 0.7)	**2.14**	**8.45**	**3.12**

Table 6: Hybrid workflow simulation results (10,000 iterations).

critical editions ($<$ 1% CER) — Strategy 3.

4.6 RQ7: Performance on Culturally Significant Terms

Term	Occurrences	HTR Correct (%)	Human Correct (%)	Common HTR Errors
ren (仁)	14	57.1%	100%	仁 → 二 (èr) or 巨 (sā)
dao (道)	11	72.7%	100%	道 → 造 (zào)
li (礼)	9	66.7%	100%	礼 → 扎 (zhā) or 化 (huà)
“Christo-Confucian”	4	**0%**	100%	→ “Christo-Catholic”/“Christo-confusion”
“co-humanity”	6	16.7%	100%	→ “humanity”/“co-humanitarian”

Table 7: HTR performance in culturally significant areas.

HTR had an average accuracy of only 42.6% compared with humans’ 100%. The first is “Christo-Confucian”, which was not recognized. “Christo-confusion” is especially harmful: it is a two-character edit that preserves the original form while changing the meaning. The substitution of “Christo-Catholic” shows that Christian theological vocabulary is more frequently used in training corpora, thereby rendering Lin’s project to integrate Chinese humanism and Christian ethics unnoticed.

5 Discussion

5.1 The Bilingual Transcription Gap

These findings suggest that bilingual code-switching may pose challenges that differ from those encountered in more linguistically homogeneous manuscripts. Prior research has shown that recognition accuracy is highly sensitive to script consistency, the quality of the training data, and document characteristics, and generally decreases with increasing linguistic and structural complexity [14]. The large effect size ($\eta_p^2 = .80$) shows that transcriber type accounts for 80% of the variance in accuracy; thus, there is a basic deficiency in the current HTR system for bilingual transcription that needs to be addressed. A relatively high number of HTR errors in code-switched regions shows that problems in multilingual environments are more pronounced than those in monolingual transcription. On the other hand, although the error rates of human transcribers in code-switching areas were also relatively high, the increase was not significantly larger. This pattern generally aligns with research on bilingualism, which has shown that code-switching is a rule-based system of language use among bilingual individuals rather than a form of communication failure [12]. Furthermore, previous studies have also demonstrated that even modern transformer models (HTR-TF: 27.89% CER in code-switching areas) have not yet addressed this problem.

5.2 Semantic-Blind vs. Mechanical Errors

The types of errors provide empirical support for distinguishing semantic-blind from mechanical transcription errors. HTR errors were mainly cross-script confusion, misrecognition of code-switching boundaries, and semantically significant substitutions; human errors were more likely to involve punctuation, line-break handling and formatting inconsistencies. This pattern aligns with House's [8; 9] distinction between covertly erroneous errors, which maintain surface plausibility but change in meaning, and overtly erroneous errors, which are readily apparent and correctable. Given HTR, semantically significant errors are especially difficult for end users to notice as they often appear linguistically plausible unless compared with the original manuscript; thus, broader problems in transparency and interpretability of machine-generated transcriptions have arisen [14].

5.3 The Code-Switching Threshold Effect

The data suggest that HTR performance declines substantially as the frequency of code-switching increases. Manuscripts containing more frequent language alternation generally exhibited higher error rates than those with fewer switching points. This effect does not occur in humans; the architecture is likely flawed. We can interpret this as "language momentum": once engaged with a particular language, HTR models need strong visual evidence to switch, and in ambiguous areas, they will generally continue using the previously detected language. The problems in the code-switching passages are the same as those reported in multilingual sequence modelling. Transformers are good at capturing sequential relationships via attention mechanisms [19], but their performance is still relatively sensitive to the quality and representativeness of the training data. Recognition systems are more likely to have problems in multilingual and code-switching scenarios when language boundaries, script switching, or mixed-language patterns are poorly represented. Therefore, the new architectural designs proposed by this paper will not rely on fine-tuning; rather, explicit language identification modules, cross-lingual attention mechanisms, and multi-task learning that simultaneously optimize character recognition and language identification will be employed.

5.4 The Erasure of Neologisms and Ethical Implications

The discovery that HTR has a 0% accuracy for "Christo-Confucian" and consistently replaces it with "Christo-Catholic" or "Christo-confusion" may reflect biases introduced through training data and model design. As the language model was trained on statistically dominant historical English corpora, it has absorbed unfamiliar neologisms into familiar ones. The above results support the growing body of studies questioning the impartiality of digital technology in cultural heritage. Owens [15] believes that, at its core, people's choices and motives shape digital preservation, and research on algorithms has also shown that, generally speaking, computational tools tend to inherit the inherent assumptions and biases present in their construction and training data [13]. The present study extends the above concerns to HTR by showing that transcription errors are not distributed randomly but are systematically concentrated in bilingual and code-switching situations; thus, machine-generated transcriptions may reproduce the linguistic bias of the corpora used to train them. Cross-cultural manuscripts, such as translation drafts, diaspora correspondence, colonial records, and the working notebooks of bilingual intellectuals, are not only technically erroneous but also may obscure linguistic innovation, cultural hybridity, and author-specific forms of expression. The substitution of "Christo-Catholic" for "Christo-Confucian" indicates that Christian theological language has been prevalent in the training data, thereby excluding Lin's purpose of uniting Chinese humanism and Christian ethics.

5.5 The Value of Content Expertise

The content expert (Transcriber B, 2.14% CER) outperformed both the paleographer (3.28%) and DH scholar (4.11%), with the advantage most pronounced on neologisms and bilingual wordplay. The superior performance of the content expert suggests that successful transcription depends not only on paleographic skill but also on familiarity with the manuscript's intellectual and linguistic context. Knowledge of Lin Yutang's characteristic coinages and bilingual writing practices enabled the expert transcriber to recognize forms such as "co-humanity" and "Christo-Confucian" as intentional lexical innovations rather than transcription errors. This observation is consistent with Gadamer's concept of the "fusion of horizons", which holds that understanding emerges through the interaction between the interpreter's prior knowledge and the historical horizon of the text [6]. For institutions, this suggests that investing in content-expert transcribers may yield higher returns than paleographic specialists for manuscripts with author-specific vocabulary.

5.6 The Hybrid Workflow Solution

The hybrid workflow simulation, HTR with confidence-based human correction, achieved a CER of 2.14% at 66% lower cost than human-only transcription. The research shows that the confidence-based strategy is superior to a code-switching-only review; it can identify errors in all language scenarios and reduces HTR CER of 22.88% in Chinese-only passages. The following are proposed implementations: (a) export confidence scores at the segment level from HTR platforms; (b) set up empirically calibrated confidence thresholds for this collection; (c) route low-confidence segments to content-expert human reviewers; (d) document all post-correction decisions for transparency. This will meet the DHI 2026 CFP's demand for "sustainable and interoperable systems".

5.7 Situating Findings in Broader DH Discourse

The results above are part of ongoing research. The results also show that there are problems in the field of interpretation and representation of algorithmic bias in digital scholarship. Drucker [5] argues that digital representations are interpretive constructions rather than neutral reflections of reality. Therefore, transcription quality should not be reduced to a single index of accuracy; instead, the extent of harm caused by a specific error varies according to what one wants to learn in research or interpret. The tendency of HTR systems to misrecognize culturally significant terms in Lin Yutang's notebooks further aligns with the critiques put forward by Noble [13] and Benjamin [1] about algorithmic systems being non-neutral infrastructures that may reproduce biases in the training data and design. This paper extends the above concerns to the field of cultural heritage and shows how such biases may affect the transcription and interpretation of multilingual historical manuscripts. Interpretive Fitness Audits are replicable methods that can be used by other researchers studying multilingual manuscripts.

5.8 Limitations

The following are some deficiencies. The corpus contains only one author and one archive; generalization to other bilingual environments (Arabic-French, Spanish-Indigenous languages, Hindi-English) is not yet known and requires further investigation. A sample size of 50 pages is suitable for detecting moderate-to-large effect sizes, but it may be too small to detect small differences among HTR models. The three scholars from Chinese literature and translation studies in the interpretive fitness panel are not comprehensive; other disciplines will also be added in the future. The study did not investigate the downstream effects of error propagation in citation networks, and such long-term tracking is beyond the scope of this project. Finally, the ground truth was agreed

upon by the human transcribers and verified by the research team. However, there was a risk of circularity; this concern has been somewhat reduced by the high inter-transcriber agreement (ICC $= 0.89$).

6 Conclusion

6.1 Aim and Significance

This study addressed a problem at the intersection of digital humanities, archival science and computational linguistics: how to systematically evaluate HTR systems for transcribing complex bilingual and code-switching manuscripts by considering not only the raw accuracy of transcription but also interpretive fidelity, that is, how well scholarly meaning, cultural nuances and cross-linguistic coherence are preserved. As cultural heritage institutions gradually incorporate HTR into their multilingual collections, there is currently no data on when and how to use such tools ethically and effectively.

6.2 Major Findings

Six main results were obtained. First, HTR had a CER of 17.84% and a human CER of 3.18% ($\eta_p^2 = 0.80$, $d = 2.84$). Second, HTR performance declined substantially as the frequency of code-switching increased. Third, HTR errors are mainly "semantic-blind" (cross-script confusion, code-switching boundary misrecognition, semantic-contextual errors), and human errors are "mechanical" (punctuation, normalization). Fourth, the interpretive fitness of the best HTR output was rated 2.84/5 compared to 4.74/5 for humans ($r = 0.72$). Fifth, the accuracy of HTR on culturally significant terms was 42.6% compared to 100% for humans, and "Christo-Confucian" was never recognized correctly. Sixth, a hybrid workflow with confidence-based human correction achieved a 2.14% CER, with a 66% cost reduction and an 83% speed-up.

6.3 Unique Contribution

The four contributions are presented here. Empirically, it is the first controlled HTR-human comparison of Chinese-English bilingual manuscripts. Theoretically, this study puts forward the two types of transcription errors, semantic-blind and mechanical errors and their different interpretations. Building on categories proposed by James [10] and House [9] such as competence-related and performance-related errors, as well as covert and overt errors, and extending these traditions to AI-assisted transcription, this paper argues that some errors stem mainly from a lack of contextual understanding, while others result from execution or procedural defects. Methodologically, an Interpretive Fitness Audit framework is developed to assess the quality of transcription beyond the overall CER, indicating that CER and interpretive fitness are not in perfect agreement; this has serious consequences for the quality assurance of digital humanities projects. Critically, it shows that HTR systems were more likely to misrecognize the very characteristics that give bilingual manuscripts cultural value, such as neologisms, boundaries of code-switching, and cross-linguistic wordplay; thus, algorithmic transcription may not be entirely neutral but rather an active interpretation that imposes the biases of its training data onto the texts it processes. Since HTR reached a 0% accuracy for "Christo-Confucian" after substituting "Christo-Catholic" or "Christo-confusion", and the same occurred in other variations, these findings suggest that automated transcription may contribute to the loss of intellectual and cultural hybridity in bilingual manuscripts.

6.4 Implications for Future Research

Future research will conduct some studies. First, extend this method to many different bilingual contexts, such as Arabic-French, Spanish-Indigenous languages, Hindi-English, and other code-

switching traditions, and determine whether the patterns identified here are universal or particular to Chinese-English. Second, build HTR architectures that explicitly model language identity as a latent variable, perhaps using multi-task learning frameworks to combine character recognition and language identification. Third, in the future, longitudinal studies should be carried out to observe how transcription errors spread through scholarly communication ecosystems in digital editions, citations, and derivative works. Such research would help address a current lack in HTR studies concerning the widespread effects of machine-generated transcriptions on the information environment and scholarly practice [14]. Fourth, research users' trust and interpretation behaviour in user-centred studies by observing how scholars engage with HTR-transcribed materials and whether transparent documentation affects their confidence and interpretations. Fifth, conduct context-sensitive cost-benefit analyses of hybrid workflows across institutions, including large research libraries and community-based heritage projects, and offer empirically supported recommendations for resource allocation.

In short, turning a manuscript into a megabyte is far from simple, and it will not be done all at once. HTR systems have offered certain general-purpose advantages for expansion. However, the difficulties are particularly concentrated in the parts scholars are most interested in: neologisms, boundaries of code-switching, culturally specific idioms, and bilingual wordplay. The ethical problem in the community of digital humanities is not that we need to choose between machines and people in a way that neglects one; rather, it concerns how to construct workflows, evaluation systems, and transparency norms that leverage the strengths of both and address their respective weaknesses. The model for the in-depth study mentioned above is called the Interpretive Fitness Audit. With the increase in the volume of archival collections, their scale has long surpassed the capacity for human transcription; therefore, this urgent demand for technology-assisted transcription is now unavoidable, and the standards we establish today will set the benchmark for scholarly records in the years to come.

References

[1] Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, 2019.

[2] Brown, H. Douglas and Lee, Heekyeong. *Principles of Language Learning and Teaching: A Course in Second Language Acquisition*. Routledge, 2025.

[3] Corder, S. P. "The significance of learners' errors". In: *International Review of Applied Linguistics in Language Teaching 5*, no. 4 (1967), pp. 161–170. DOI: 10.1515/iral.1967.5.1-4.161.

[4] Deppman, Jed, Ferrer, Daniel, and Groden, Michael. *Genetic Criticism: Texts and Avant-textes*. University of Pennsylvania Press, 2004.

[5] Drucker, Johanna. "Humanities Approaches to Graphical Display". In: *Digital Humanities Quarterly 5*, no. 1 (2011). URL: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

[6] Gadamer, Hans-Georg. *Truth and Method*. Translated by J. Weinsheimer and D. G. Marshall; original work published 1960. Bloomsbury Academic, 2013.

[7] Holley, Rose. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". In: *D-Lib Magazine 15*, no. 3/4 (2009). DOI: 10.1045/march2009-holley.

[8] House, Juliane. *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag, 1997.

[9] House, Juliane. *Translation Quality Assessment: Past and Present*. Routledge, 2015.

[10] James, Carl. *Errors in Language Learning and Use: Exploring Error Analysis*. Routledge, 2013.

[11] Muehlberger, Guenter et al. “Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study”. In: *Journal of Documentation 75*, no. 5 (2019), pp. 954–976. DOI: 10.1108/JD-07-2018-0114.

[12] Myers-Scotton, Carol. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press, 1997.

[13] Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018. DOI: 10.18574/nyu/9781479833641.001.0001.

[14] Nockels, Joe, Gooding, Paul, Ames, Sarah, and Terras, Melissa. “Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research”. In: *Archival Science 22*, no. 3 (2022), pp. 367–392. DOI: 10.1007/s10502-022-09397-0.

[15] Owens, Trevor. *The Theory and Craft of Digital Preservation*. Johns Hopkins University Press, 2018.

[16] Qian, Suoqiao. *Liberal Cosmopolitan: Lin Yutang and Middling Chinese Modernity*. Vol. 3. Brill, 2011.

[17] Qian, Suoqiao. *The Cross-Cultural Legacy of Lin Yutang: Critical Perspectives*. Institute of East Asian Studies, University of California, Berkeley, 2015.

[18] Reul, Christian, Christ, Dennis, Hartelt, Alexander, Balbach, Nico, Wehner, Maximilian, Springmann, Uwe, Wick, Christoph, Grundig, Christine, Büttner, Andreas, and Puppe, Frank. “OCR4all – An open-source tool providing a (semi-)automatic OCR workflow for historical printings”. In: *Applied Sciences 9*, no. 22 (2019), p. 4853. DOI: 10.3390/app9224853.

[19] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, and Polosukhin, Illia. “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5998–6008.