# Hidden Signals in *Dream of the Red Chamber*: A Computational Text Analysis of Edition Layers, Writers' Habits and Plot Turns

**Weixiang Gan[1], Mengfei Xiao[1,*], Jialin Liu[2], Tenghao Chen[2]**
[1]Lincoln University College, Petaling Jaya 47301, Malaysia
[2]Taylor's Business school, Taylor's university, Subang Jaya 47500, Malaysia
*Corresponding email: gabriel1995@qq.com

## Abstract

This article tests whether the stylistic break between the first 80 and the last 40 chapters of *Dream of the Red Chamber* can be told apart from the splicing of editions. Diagnostic variants and 8-gram overlap show that Project Gutenberg e-text No. 24264 carries Zhiyan-recension text in chapters 1 to 80 except 67. They also show that it carries Cheng-recension text in chapter 67 and chapters 81 to 120. Therefore, every test was repeated on Wikisource transcriptions of the Cheng A and Cheng B prints. Thirty-four masked function characters enter leave-one-chapter-out classification. The analysis also includes an energy-statistic change point, block permutation, a sentence-block bootstrap, and a trend-versus-step comparison. Character networks and an audited emotion lexicon serve as plot-level checks. In all three texts, the single change point falls after chapter 80 (block permutation $p \leqslant 0.004$). The leave-one-chapter-out area under the Receiver Operating Characteristic (ROC) curve is 0.983-0.988. Effect sizes correlate at 0.99 between the e-text and Cheng A. The share of 笑道 ("said with a smile") among quotations falls from about 36.00% to about 7.00%. Seven or eight of chapters 81 to 91 lean toward the earlier style. Sadness words turn after chapter 94 and household figures gain brokerage in the last 40 chapters, neither at chapter 80. In the Cheng prints the break-location interval is wide, and 嗎 and 林黛玉 are edition-sensitive. Authorship notes of annotated editions can cite these edition-specific results, and digital classics should carry base-text and conversion metadata.

## Keywords

*Dream of the Red Chamber*, Authorship attribution, Edition layers, Change-point detection, Function characters, Character network, Emotion lexicon



Weixiang Gan
First author
Lincoln University College



Mengfei Xiao
Corresponding author
Lincoln University College

## Introduction

The annotated *Dream of the Red Chamber (Hongloumeng)* of People's Literature Publishing House (PLPH) first appeared in 1982, based on the Gengchen manuscript for chapters 1 to 80. And on the Cheng A

print for chapters 64, 67 and 81 to 120 (in the Jiemian News), and credited to “Cao Xueqin and Gao E”. The third edition credits “Cao Xueqin, continued by an anonymous author, edited by Cheng Weiyuan and Gao E”, a change Zhang (in the Guangming Daily) traces to Redologists’ findings; a fourth edition followed in 2022 (in the Beijing Youth Daily). Quantitative studies can add checkable figures to such a note, provided they state which edition they count. The proviso reaches many readers. In 2025, 30,395,000 students were enrolled in Chinese senior high schools, according to the Ministry of Education of the People’s Republic of China (MOE). The MOE curriculum standard ranks “whole-book reading and discussion” first of its 18 learning task clusters. Each classroom uses one edition.

Quantitative work on the dispute spans more than forty years. Chen Bingzao’s single-author view and the two-writer views of Chen Dakang and Li Xianping (as cited in Chinese Journal of Applied Probability and Statistics) coexist with the no-difference finding of Yu et al [1]. Hu et al. located the break at chapter 80 by predicting held-out chapters 61 to 90. Three problems remain [2]. The first is edition layers: 120-chapter e-texts are often keyed in from modern annotated editions whose two parts come from the Zhiyan manuscripts and the Cheng prints, so a stylistic break would coincide with the edition seam. And a wholesale substitution, such as replacing every 嗎 with 麼, can overturn judgments resting on single characters. The second is inference: Adjacent chapters share scenes and characters, so permuting chapters as independent units yields p values that are too small. The third is interpretation: The last 40 chapters turn to the family’s decline, and network and emotion changes may come from the plot alone.

By “hidden signals”, this article means traces that only counting reveals. The first is writers’ habits with function characters such as 之，其，了 and 著, whose rates vary little with subject matter. Individual differences in these habits are explained through cognitive entrenchment. The second is the edition layer left by copyists, printers and editors. The third is homophonic names glossed in the Zhiyan commentary, as when a Jiaxu side comment reads the name Shiyin as “hiding the true facts”. Character networks and emotion trajectories belong to the plot level and serve as controls: If their turns do not coincide with the function-character break, that break is not a by-product of plot change.

Five questions follow. Research question (RQ) 1: To which edition systems do the parts of the e-text belong? RQ2: Is there a stylistic break, where, and does it hold in the edition-consistent Cheng A and Cheng B prints? RQ3: Which function characters and speech formats carry it, and which are edition-sensitive? RQ4: Do network and emotion turns coincide with it? RQ5: How are the glossed coded names distributed? The contributions include an edition-layer diagnosis with every test repeated on both Cheng prints. They also include an unsupervised energy-statistic break point tested under serial correlation [3]. Furthermore, they provide masked function-character counts with edition-sensitive markers flagged, along with public scripts and coding records.

## Literature, theory and hypotheses

### *Quantitative studies and edition layers*

As summarized by Wei (in the Chinese Journal of Applied Probability and Statistics), Chen Bingzao inferred a single author from five word classes, whereas Chen Dakang (88 items) and Li Xianping (47 function characters) judged the last 40 chapters to be by others, containing a few of Cao Xueqin’s drafts or completed by relatives and friends of the Cao family.

A study found 嬤嬤 (“nanny”) in 34 of the first 80 chapters and in none of the last 40, and inferred another source for chapters 64 and 67 from words concentrated there. Chapters 61 to 80, except chapter 67, fell in the first part, while all of chapters 81 to 90 fell in the second. Chapters 1 to 80 had no internal break, and chapters 96 to 105 changed gradually. This pattern was explained

either by a continuator who first imitated the earlier text or by Gao E completing drafts. Zhu et al. also found two writers, Yu et al. found no difference in word length, and Zhao et al. reported mixed attributions [4,5]. These studies count different texts and treat chapters or passages as independent observations.

The tools come from stylometry at large: Burrows's Delta and its variants, and rolling stylometry, applied to Henry VIII, The Female Quixote and a Qing novel [6-10]. Change-point methods assume exchangeable observations under the null, so serial correlation between chapters must be handled. Against Da's charge that computational literary studies draw conclusions that fail statistical testing, each conclusion here carries a control that could prove it wrong [11].

Copying, printing and modern editing also altered characters. In the Introductory Remarks of the Cheng B print, Cheng Weiyuan and Gao E say that they corrected the first 80 chapters but "did not presume to alter the original text" of the last 40. Modern annotated editions pair the Gengchen manuscript with Cheng A and revise single variants back and forth. 花魂 ("the flower's soul") in chapter 76 became 詩魂 ("the poet's soul") in the second and third PLPH editions and returned in the fourth. E-texts keyed in from them mix writers, recensions and modern editing. Optical Character Recognition (OCR) noise changes attribution results. Conversion and splicing noise can be tested the same way, by recomputing on another source [12]. If the break after chapter 80 is only an edition seam, it should weaken or move in the edition-consistent Cheng prints. If it comes from the writers, it should recur at the same place with a similar magnitude.

## *Plot-level checks and coded names*

Character networks use co-occurrence windows, and language models extract relation triples. Stage comparisons must handle text volume, which alters density. Emotional arcs use general lexicons, but classical vernacular emotion characters, such as 唬 ("startle"), call for a lexicon built from the text and audited item by item. Foretold fates traced by Chan and Zhao supply plot reference points, and Seng reads chapters 1 to 5 homophonic names as foreshadowing [13-15].

## *Hypotheses and pre-analysis disclosure*

The hypotheses were finalized after exploratory analysis of the e-text. Its design file records a leave-one-chapter-out area under the ROC curve (AUC) of 0.992. It records an energy-statistic maximum at k=80, counts of 3/72 for 嗎 and 2215/219 for 笑道 in two parts, and 15 occurrences of Zhan Guang in last 40 chapters. Hypothesis (H1b, H2a, H2b) and Prediction (P5) count as "confirmed after exploration". The Cheng prints were unused, so tests on them are independent. H3a and H3b remained unrevised.

(1) H1

H1a: The leave-one-chapter-out AUC for chapters 1 to 80 (A) against 81 to 120 (B) exceeds those placebo and other boundaries. H1b: $\hat{k}$ lies in chapters 78 to 82, with block-permutation $p<0.01$ (5- and 10-chapter blocks). A 90.00% sentence-block interval lies within chapters 78 to 82. This revises the original criterion; both are reported. H1c: Chapters 64 and 67 rank in the top 10.00% by later-style probability.

(2) H2

H2a: The Classical characters 之, 其, 亦, 乃, 皆, 或 and 故 decline and 么 and 嗎 rise ($q<0.05$). H2b: The proportions of 笑道 and of full forms of reference change, also in the narration layer.

(3) H3

H3a: Active characters and density peak in chapters 41 to 80. H3b: Centrality shifts in chapters 81 to 120 from young characters (Baoyu, Daiyu, Baochai, Xiangyun) to household figures (Jia Zheng, Lady Wang, Jia Lian). H3c: Modularity is higher in 81 to 120 than in 41 to 80.

(4) H4

H4a (main emotion test): Sadness words occur more in chapters 81 to 120. H4b: Chapters with death words in titles are sadder. H4c: Net valence changes away from the stylistic break, with chapter 80 outside the interval.

(5) P5 (descriptive)

The distribution of the glossed coded names; whether the continuation coined new ones cannot be tested, as the commentary covers only chapters 1 to 80.

# Data and methods

## *Texts and edition diagnosis*

The main material is Project Gutenberg eBook No. 24264, updated December 22nd, 2021, and downloaded October 16th, 2026. It names no base text and bears traces of one-to-many simplified-to-traditional conversion, with 兒 almost everywhere turned into 儿. Cleaningyielded 120 chapters with 722,113 Han characters, deleting 126 conversion artefacts and keeping 392 missing-glyph marks. The main specification uses the prose layer (verse removed) with the chapter as unit ($n_A$=80, $n_B$=40); narration, dialogue and full-text layers serve as checks.

The comparison texts are Wikisource transcriptions (accessed October 17th, 2026) of the 1791 Cheng A print, with 1,327 of its 1,635 scanned pages proofread, and of the 1792 Cheng B print, whose typeset source is not named; a collated edition and a Zhiyan composite of the Gengchen, Jimao and Jiaxu manuscripts with commentary serve for diagnosis only. Glyph normalization replaces 147 pairs of variant and residual simplified forms; 嗎 and 么 are kept apart.

The diagnosis uses diagnostic variants, the share of each chapter's 8-grams found in the same chapter of a comparison text, and glyph markers (in the Table 1).

Chapters 1 to 80 of the e-text except 67 are Zhiyan text: All three Zhiyan variants are present, and each of the 78 chapters other than 64 and 67 overlaps more with the Zhiyan composite than with Cheng A. Chapter 67 and chapters 81 to 120 are Cheng text, and 37 of the last 40 chapters are closer to Cheng A than to Cheng B (in the Figure 1a). Chapter 64 is also Zhiyan text, so the e-text does not reproduce the 1982 PLPH edition, and its printing cannot be determined. The edition seam thus coincides with the hypothesized change of writer. Cheng B colloquializes its own first 80 chapters (77 instances of 嗎 against 12 in Cheng A), which would narrow the A/B difference, so Cheng A is the preferred check text.

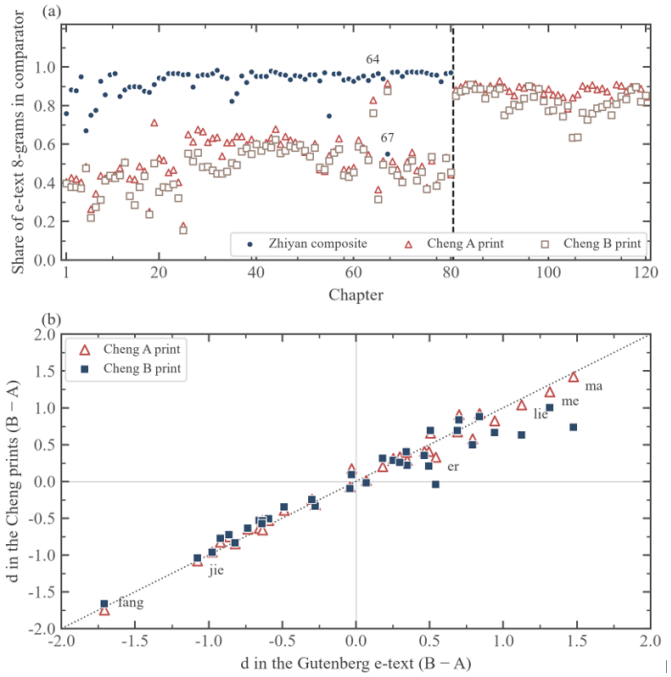



Figure 1. Edition-layer diagnosis and A/B effect sizes in the three texts.

Note: (a) Share of each e-text chapter's 8-grams found in the same chapter of the comparison text. (b) Cohen's d of the 34 function characters, e-text against the Cheng prints; the dotted line is y=x. Source: Project Gutenberg eBook #24264; Wikisource Cheng A, Cheng B and Zhiyan texts.

### *Features*

FW34 comprises 34 function characters, including the Classical 之, 其, 亦, 乃, 皆, 或 and 故 and the vernacular 了, 的, 著, 么 and 嗎; extended and merged sets serve as checks.

Before counting, the aliases of the 81 network characters, 54 "X'er" names such as 鳳姐儿, and content-word compounds of nine polysemous characters (e.g., 方 in 地方, "place") are masked, removing 8.29% and 9.89% of the prose layer in the two parts.

Rates are per 1,000 Han characters. The speech formats are kept out of the classifiers. These formats include the share of 笑道: among all quotations opened by a character plus 道: (X 道:). They also include the shares of four full forms of reference. These shares are modelled per quotation or mention by generalized estimating equations (GEE) clustered by chapter.

Table 1. Edition-layer diagnosis of the Gutenberg e-text.

| Evidence | E-text | Cheng A | Cheng B | Collated | Zhiyan composite |
|---|---|---|---|---|---|
| A Diagnostic variant | / | / | / | / | / |
| Ch. 1: Jiaxu reading of the monk–Taoist passage | Yes | No | No | Yes | Yes |
| Ch. 3: 罥煙眉 (Cheng: 籠煙眉) | Yes | No | No | Yes | Yes |
| Ch. 76: 葬花魂 (Cheng: 葬詩魂) | Yes | No | No | Yes | Yes |
| Ch. 1: opening statement as main text | Yes | Yes | Yes | No | No |
| B Share of e-text 8-grams found | / | / | / | / | / |
| Chs. 1 to 80 except 64 and 67, mean | / | 0.516 | 0.463 | 0.971 | 0.928 |
| Ch. 64 | / | 0.830 | 0.761 | 0.979 | 0.956 |
| Ch. 67 | / | 0.916 | 0.876 | 0.980 | 0.550 |
| Chs. 81 to 120, mean | / | 0.882 | 0.818 | 0.989 | / |
| C Markers (chs. 1 to 80 / 81 to 120) | / | / | / | / | / |
| Count of 嗎 | 3 / 72 | 12 / 75 | 77 / 80 | 3 / 75 | / |
| 笑道: as share of X 道: (%) | 36.72 / 7.34 | 35.55 / 7.31 | 35.51 / 7.30 | 36.95 / 7.41 | / |

Note: Group B is computed after glyph normalization. Source: Project Gutenberg eBook #24264; Wikisource Cheng A, Cheng B, collated and Zhiyan texts.

### *Authorship and break-point tests*

Adjacent chapters are serially correlated in a statistically meaningful way (lag-1 autocorrelations of the first two principal-component scores, PC1 and PC2, within chapters 1 to 80 are 0.56 and 0.38), so all permutations and bootstraps use circular 5-chapter blocks (10 in checks). Cohen's d carries a moving-block bootstrap 95% CI and a block-permutation p converted to q [16]. Delta T is the mean Burrows's Delta between chapters of different parts minus the average of the two within-parts mean across all sampled segments.

L2-regularized logistic regression is evaluated by leave-one-chapter-out (LOO) prediction, with the regularization parameter C chosen by inner cross-validation; buffered LOO also drops the two chapters on either side of the target, and a linear support vector machine (SVM) is the second classifier. The same pipeline is applied to placebo breaks after chapters 20, 40, 60 and 100 and to every boundary k=20, ..., 100.

Rolling classification follows Eder: with chapters 1 to 60 and 101 to 120 as anchors, chapters 61 to 100 receive a probability $P(B)$ of the later style, smoothed with a Gaussian kernel (h=3), and the crossover is where the smoothed curve passes 0.5 for good [17]. A three-group

comparison applies a classifier trained on chapters 1 to 80 versus 92 to 120 to chapters 81 to 91.

$$\widehat{Q}(k)=\frac{k(n-k)}{n}\left[\frac{2}{k(n-k)}\sum_{i\le k<j}\left\|z_i-z_j\right\|-\binom{k}{2}^{-1}\sum_{i<i'\le k}\left\|z_i-z_{i'}\right\|-\binom{n-k}{2}^{-1}\sum_{k<j<j'}\left\|z_j-z_{j'}\right\|\right] \quad (1)$$

Where $z_i$ is the standardized feature vector of chapter $i$, $n$ the number of chapters, and $\hat{k}=\operatorname{argmax}_{10\le k\le 110}\widehat{Q}(k)$ marks a break after chapter $\hat{k}$. Null distributions come from chapter-wise permutation, 5- and 10-chapter circular block permutation (999 each) and all 119 circular shifts, and the uncertainty of $\hat{k}$ from a sentence-block bootstrap (1,000 replicates) and leave-one-chapter-out deletion. Trend models with and without a step are fitted to PC1 to PC5 and compared by the Bayesian information criterion (BIC) and by an AR(1) parametric bootstrap; binary segmentation accepts a secondary break only if block permutation gives $p<0.01$. Every test is run on all three texts. The edition shift of a character is its mean rate difference between the same chapters 1 to 80 of two texts over the pooled standard deviation, with chapters 81 to 120, Cheng A text in both, as the control.

### *Plot-level checks and coded names*

The network nodes are the 81 characters mentioned at least 20 times, with ambiguous terms of address such as 老爺 (“the Master”) and 太太 (“Madam”) excluded; edges count co-occurrence in five-sentence windows ($w\ge 2$), and communities come from greedy modularity optimization [18]. As the stages differ in length, inference uses equal-size thinning to 1,761 windows and chapter-label permutation in 5-chapter blocks.

The candidate emotion lexicon has 44 items in four categories (sadness, joy, anger and fear), counted after masking 54 non-emotion strings like 笑道. One coder judged up to 50 random hit sentences per item (1,748 total); items with precision of at least 0.70 were kept. Net valence contrasts joy with the three negative categories; events are coded from chapter titles (16 death, 7 calamity and 16 festivity chapters). H4b uses Ordinary Least Squares (OLS) with Newey-West and HC3 standard errors. Emotion change points use the above method with a joint bootstrap over sentences and lexicon items.

The single change point maximizes the energy statistic of Equation (1):

The 29 homophone comments in chapters 1, 2, 3, 7, 8 and 24 of the Zhiyan composite are merged into 22 units (17 persons and 5 places), and a keyword search over all 4,048 commentary blocks returned 43 blocks, coded by a single coder. Python 3.13 scripts (random seed 20261016) produce every number reported.

## Results

### *Function characters and speech formats*

Of the 53 prose-layer features, 24 have $q<0.05$ and 16 have $|d|\ge 0.8$ (in the Figure 2). All seven Classical function characters decline in chapters 81 to 120 (in the Table A1 in Appendix), five with $q<0.05$ (乃 and 其: $q=0.075$) and six of seven in narration layer. 方, which after masking survives as 方才 (“just now”), falls most, replaced by 剛才. 嗎 occurs three times in chapters 1 to 80, in chapter 67, and 72 times in last 40; combined with 么 it gives $d=1.43$ [0.97, 2.02].

Speech introduction changes more than any single character: the chapter-level d for the 笑道: share is -2.83 [-3.56, -2.24]. The share depends on scene type, rising from 31.82% in chapters 1 to 40 to 41.65% in chapters 41 to 80 before collapsing. Yet the narration-layer rate of 道: is unchanged: the continuation quotes as often but changes the verb, with fewer 笑道 and more 問道. Full forms of reference fall in three of four groups, and 薛寶釵 occurs 25 times in chapters 1 to 80 and never later; for 史湘云 the estimate is too imprecise to show a difference (OR 1.79 [0.62, 5.16]). H2a holds for 5/7 Classical characters and the interrogatives, and H2b for 笑道 and three forms of reference. The parts separate on PC2 ($d=2.15$ [1.55, 3.45]) rather than the Classical-vernacular PC1, and Ward clustering does not split the parts (two-cluster adjusted Rand index=0.11); its three-cluster cut isolates chapters 1 to 5, 17, 18 and 78.

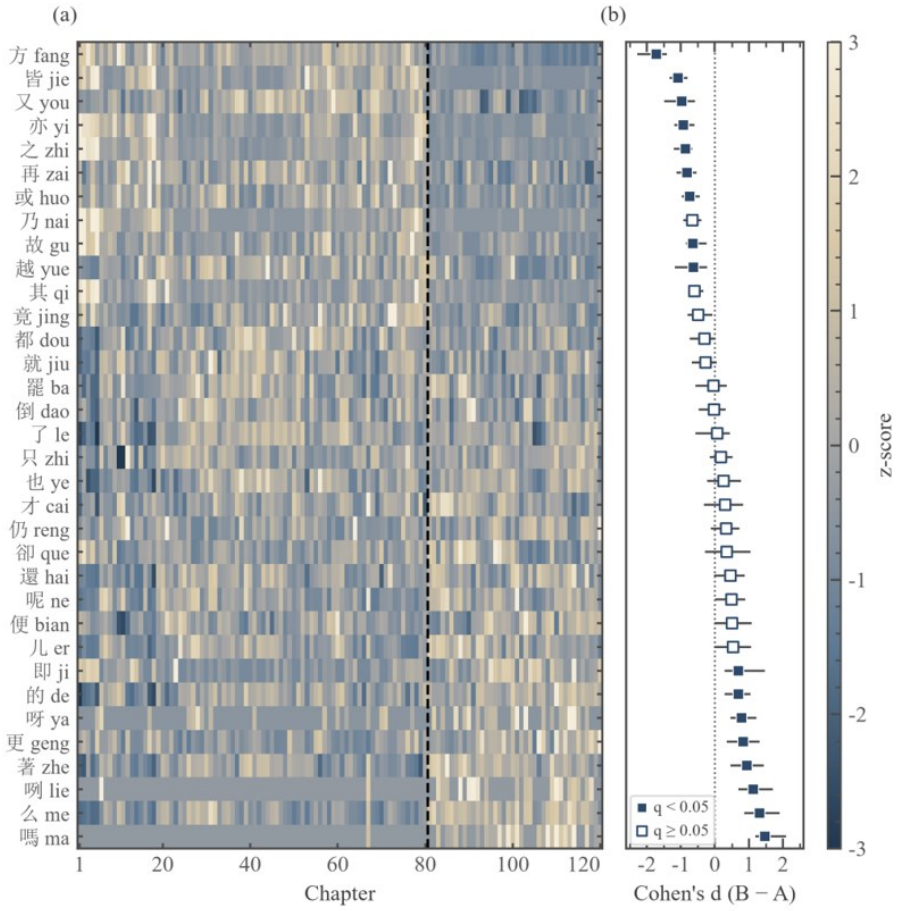



Figure 2. Chapter-level z-scores of the 34 function characters and the differences between the two parts.

Note: (a) z-scores truncated at ±3, rows ordered by d; the dashed line lies between chapters 80 and 81. (b) d with 95% moving-block bootstrap intervals; solid squares mark q<0.05. Source: Author's calculations from Project Gutenberg eBook #24264.

### *Classification, placebo boundaries and the break point*

The LOO AUC on the e-text is 0.988, with chapters 23 and 67 misclassified (in the Table A2 in Appendix). The 95th percentile of the null AUC is 0.742 under block permutation (0.617 chapter-wise), and p reaches its minimum of 0.001 under both. Separability does not locate the break. Placebo splits after chapters 20, 40 and 100 give 0.876, 0.739 and 0.783 (after chapter 60, 0.483, a reverse bias of LOO in small samples). Delta T is 0.014 for chapters 1 to 40 against 41 to 80 (p=0.328) and 0.171 for chapters 41 to 80 against 81 to 120 (p=0.0005).

Over successive boundaries the AUC is flat between k=80 and 92: 11 of the 81 boundaries reach 0.98, the maximum is at k=91, and k=80 ranks seventh (in the Figure 3c). H1a is partly supported.

The energy statistic puts the change point after chapter 80 in all three texts. In the e-text, $\widehat{Q}$ peaks at 47.45 on a flat top (Q(79)=45.49, Q(81)=46.24); the 95th percentiles of the four null distributions range from 7.09 (chapter-wise) to 43.66 (circular shifts), and p from 0.001 to 0.008 (in the Figure 3b). In all, 984 of 1,000 sentence-block replicates and 119 of 120 deletions keep the break at chapter 80. A quadratic trend plus a step at chapter 80 is BIC-best for PC1 to PC5 (AR(1) bootstrap p=0.010; 0.189 in PC2). H1b holds on the e-text for both criteria (feature-bootstrap interval [18, 90]).

The Cheng prints give the same point estimate with wider intervals. Only 55.40% and 19.10% of their sentence-block replicates fall at chapter 80, against 43.00% and 76.20% at chapter 18. The revised H1b criterion fails on both prints, while the original one holds in all three texts

(feature-bootstrap intervals [18, 91] in both prints). Chapter 18 competes through the opening frame and the homecoming poems, and removing chapters 1 to 5, 17 and 18 narrows the intervals (in the Table A2). The break after chapter 18 weakens with 10-chapter blocks in the e-text (p=0.144), and all three texts break again after chapter 90. On 5,000-character segments, and in 27 of the 28 main sensitivity cells, the break falls exactly at the A/B boundary. These cells share one text and show insensitivity to specification, not independent replication.

### *Chapters 81 to 91 and chapter 67*

In the e-text chapters 61 to 80 have a mean P*(B)* of 0.077, none above 0.5, while 7 of chapters 81 to 91 lie below 0.5 (in the Figure 3a). These chapters split rather than sit in between: Chapters 83 and 86 score 0.75 and 0.73 and chapters 81, 82, 85 and 90 at most 0.24 (C=0.1), and weaker regularization pushes the two groups apart. The crossover moves within chapters 85 to 91 with the anchors and the regularization, and intermediate probabilities also occur in a placebo inside chapters 1 to 80 (24 of 40 chapters between 0.2 and 0.8). In the three-group comparison chapters 81 to 91 sit at 0.644 on the axis from the centroid of chapters 1 to 80 (0) to that of chapters 92 to 120 (1). And Delta places 10 of the 11 closer to chapters 92 to 120, though less close than chapters 92 to 120 themselves (d=-1.17, p=0.003).

Chapter 67 is close to chapters 81 to 120 on 4 of 6 indicators in all three texts: Its LOO P*(B)* of 0.998 is the highest in chapters 1 to 80, its SVM score is positive, it is nearer the later Delta centroid (0.737 vs. 1.028), and its PC1 score is -4.33, whereas PC2 and rolling classification point to the first part. Chapter 64 scores 0/6 (LOO P*(B)*=0.004). H1c holds only for chapter 67, and, as the Cheng prints are Cheng text throughout, its anomaly is not only a difference between recensions.

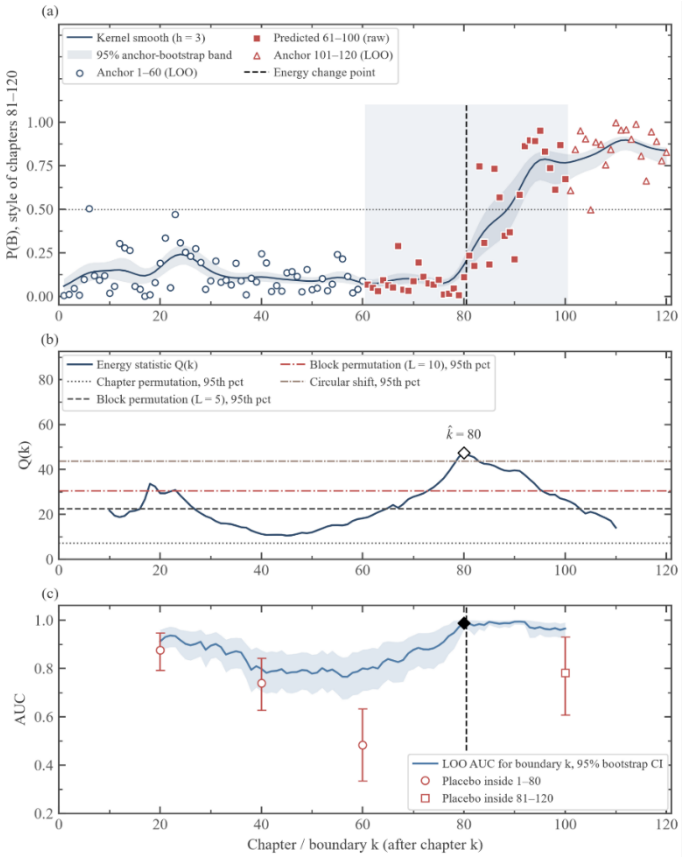



Figure 3. Rolling classification probability, energy statistic and boundary AUC.

Note: (a) P*(B)* by chapter, smoothed curve (h=3) and 95% bootstrap band. (b) Q(k) with the 95th percentiles of four null distributions. (c) LOO AUC of successive and placebo boundaries. Source: Author's calculations from Project Gutenberg eBook #24264.

### *Edition effects*

The d values of the 34 characters correlate at 0.992 between the e-text and Cheng A (same sign for 33) and at 0.965 between the e-text and Cheng B (in the Figure 1b). Over chapters 1 to 80 the edition shift between the e-text and Cheng A averages 0.094 standard deviations (0.015 in the control chapters 81 to 120) and raises the mean |d| of the e-text by 0.02, about 3.00%; the mean |d| of Cheng B is about 13.00% lower than that of Cheng A, consistent with the revisions its Introductory Remarks describe.

These statistics confirm that the edition shift is not uniform across texts. Cheng A preserves the original character habits more faithfully than Cheng B. The control chapters show minimal variation, validating the stability of the core text. Such quantifiable differences provide a reliable baseline for distinguishing editorial interventions from authorial fingerprints.

Two markers are edition-sensitive. With 77 instances of 嗎 in chapters 1 to 80 of Cheng B, the d of 嗎 falls from 1.48 to 0.74 (么 and 嗎: 1.43, 1.32 and 1.04 in the three texts), and the drop in 林黛玉, from 20.07% to 1.86% in Cheng A, disappears in Cheng B (1.30% vs. 1.40%). The 笑道: share and the 鳳姐儿 proportion hold in both prints.

### *Character networks*

The three stages have 69, 71 and 73 active characters. After thinning to 1,761 windows, density differs by -0.003 [-0.027, 0.024] between chapters 81 to 120 and 41 to 80 (p=0.564); the higher whole-stage density of chapters 41 to 80 reflects that stage's length. So H3a is not supported. Brokerage changes (in the Figure 4): the betweenness sum of Jia Zheng, Lady Wang and Jia Lian rises from 0.11 in chapters 41 to 80 to 0.43, by 0.29 [0.13, 0.43] after thinning (p=0.036).

Jia Zheng's betweenness in chapters 81 to 120 is 2.52 standard deviations above what his weighted degree predicts; dropping single chapters leaves the trio's sum between 0.354 and 0.480, and assigning 老爺 and 太太 to Jia Zheng and Lady Wang keeps the direction (0.53). The four young characters, whose sum fell from 0.88 to 0.52 between the first two stages, do not lose brokerage in chapters 81 to 120 (p=0.542), and degree centralization is unchanged (p=0.619). So H3b is partly supported; modularity (0.16, 0.23 and 0.19; p=0.141) does not support H3c.

### *Emotion intensity and change points*

The audit removed 7 of the 44 items, and the 37 retained items have a pooled precision of 0.927. Sadness words rise from 1.89 to 3.50 per 1,000 characters (d=0.87 [0.32, 1.45], p=0.006), supporting H4a; fear rises (d=0.57, p=0.022), joy and anger show no detectable change, and net valence moves from -0.40 to -0.53 (d=-0.37 [-0.87, 0.08], p=0.193; -0.74 with an unaudited 42-item lexicon and fewer masked strings).

The death-chapter coefficient is 1.24 [0.12, 2.37] (p=0.031 with Newey-West and 0.059 with HC3 standard errors), weak evidence for H4b. Sadness changes after chapter 94 (p=0.001-0.019 across the three permutation schemes; joint-bootstrap 95% interval [62, 105]; in the Figure 4) and peaks in chapter 98, where Daiyu dies. The net-valence change point after chapter 99 is significant only under chapter-wise permutation (block p=0.116) and moves to chapter 56 when 笑 ("laugh") is deleted. Within chapters 81 to 120 the sadness change point stays at 94 (block p=0.084), and with each part standardized separately none remains (p=0.582). H4c is not supported.

### *Coded names*

Of the 22 units, 13 appear only in chapters 1 to 80, mostly in chapters 1 to 8, and 9 recur later, among them Zhen Shiyin, Jia Yucun and the four Chun sisters; the hanger-on Zhan Guang appears 15 times in the last 40 chapters against 3 times in the first 80. Of the 16 name-glossing comments found, 14 lie in chapters 1 to 9, so the clustering of coded names in the opening chapters reflects where the commentary glosses names and is not tested here.

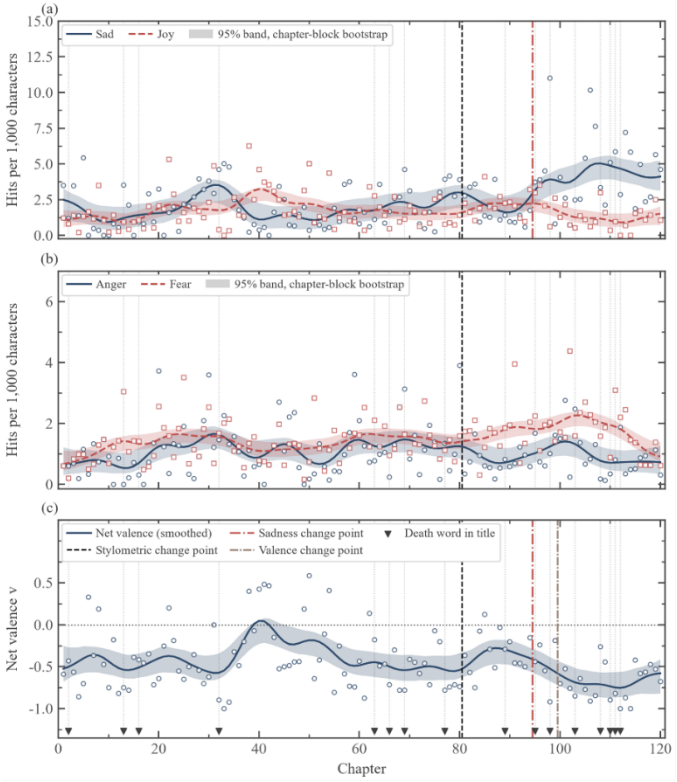



Figure 4. Chapter-level trajectories of emotion intensity.

Note: (a) Sadness and joy; (b) anger and fear; (c) net valence. Smoothed values (h=3) with 95% moving-block bootstrap bands; triangles mark death chapters, and dashed and dash-dotted lines mark the stylometric and emotion change points.

Source: Author's calculations from Project Gutenberg eBook #24264.

# Discussion

## *The break point, edition layers and earlier findings*

Were the break only the Zhiyan-Cheng seam of the e-text, it should vanish or move in the Cheng prints. However, it does not; the effect sizes correlate at 0.99 with Cheng A, and the Zhiyan layer adds only about 0.02 to the mean |d|. Splicing affects the precision of the location, not the bulk of the difference.

The result contradicts Chen Bingzao's single-author view, whose correlations may reflect genre. It also contradicts the no-difference finding of Yu et al.: word length is an aggregate measure, like mean sentence length, which hardly differs here (20.85 vs. 20.67 characters).

It agrees with Hu et al. and Zhu et al. on different writers, adding an unsupervised break in three texts, p values that allow for serial correlation and sentence-level intervals; function characters cannot identify the writer or separate the editing by Cheng Weiyuan and Gao E from the continuation itself.

Two results depart from Hu et al. They assigned all of chapters 81 to 90 to the later part, whereas 7 of chapters 81 to 91 have $P(B)<0.5$ here; with their anchors the crossover moves to chapter 85, and their wider feature set may add to the gap.

A secondary break also appears clearly after chapter 18, following the frame and verse chapters in the narrative sequence. This pattern strongly fits the earlier attribution of differences within chapters 1 to 20 to underlying narrative plot development and structural shifts.

## *Chapters 81 to 91, edition-sensitive markers and chapter 67*

Chapters 81 to 91 are not a stretch. As a group they sit at 0.64 (e-text) and 0.60 (Cheng A) on the A-B axis. The classifier puts chapters 83, 86, 87 and 91 at $P(B)\geqslant 0.5$

and other seven below. With a break after chapter 90 in all three texts, the last 40 chapters are better described as two segments, the first containing several chapters close to chapters 1 to 80. Chen Dakang's surviving drafts, Li Xianping's completion by the Cao family's relatives and friends and the imitating continuator of Hu et al. all predict this shape. Deciding among them requires matching the plot of these chapters against the foreshadowing in chapters 1 to 80 and their Cheng variants. In Cheng B the group sits at 0.52, so claims about a "transitional stretch" must name their edition.

The three texts bear out the warning: The d of 嗎 halves across editions, and the drop in 林黛玉 vanishes in Cheng B. The Classical characters, the 笑道: share and the 鳳姐儿 proportion hold in all three texts and can serve as evidence that does not depend on a single glyph; 嗎 and 林黛玉 should be reported only as edition markers. The concentration of 裡 cited for chapter 67 cannot be tested, because all three texts here normalize 裡 to 里. The Introductory Remarks call chapter 67 "present in some copies and absent in others, with the same title but different text"; that it leans later even in the Cheng prints fits a writer other than that of chapters 1 to 80, although PC2 and rolling classification point earlier.

### *Plot-level checks*

The network does not "thin out" in chapters 81 to 120: At equal text volume, its density is unchanged. In chapters 41 to 80 the broker is Jia Mu (betweenness 0.58); in chapters 81 to 120 Jia Zheng's betweenness rises to 0.25, above what his weighted degree predicts, in episodes where the head of the household handles confiscation, examinations, weddings, funerals and lawsuits; in Moretti's terms, a shift of brokerage from the grandmother toward the father's generation could serve as a stage-level indicator for family novels [19]. The sadness turn after chapter 94 lies next to Daiyu's death, one of the foretold nodes; whether the continuation keeps faith with the foreshadowing is disputed among Redologists, and these data make no judgment on it. Two discounts apply: The sadness interval includes chapter 80 and the change point vanishes under separate standardization, and a single character, 笑, moves the net-valence change point, so emotion words may carry writer habits. The function-character break at chapter 80, the sadness turn at chapter 94 and the stage-level rise in household brokerage do not coincide, so the break is not a by-product of plot.

### *Actionable measures*

Future PLPH reprints could add to the authorship note a paragraph that names the texts and methods and reports the break between chapters 80 and 81 in three texts, the small share of the edition layer, and the positions of chapters 81 to 91 and 67. Teachers could set a counting exercise: count 笑道 and 問道 in five chapters of each part in one edition, then repeat the count in another. Article 12 of the *Opinions on Promoting Ancient Books Work in the New Era* calls for structured ancient texts whose digital resources connect to the national cultural big-data system (in the General Office of the CPC Central Committee and General Office of the State Council). The e-text used here names no base text, splices two recensions and keeps 126 conversion artefacts and 392 missing glyphs. Digital classics could carry four items of metadata (base text and printing, conversion path, missing-glyph list and collation log), with the diagnostic scripts of this article as a template.

## Limitations and further research

The Cheng A transcription has 308 unproofread scanned pages (18.80%), the Cheng B transcription does not name its typeset edition, and the printing behind the e-text cannot be determined; the Cheng prints rule out the Zhiyan-Cheng seam but cannot separate the editing by Cheng and Gao from the continuator's usage. Masking covers only the 81 network characters, 54 "X'er" names and 9 polysemous characters, and characters such as 了 and 著 are not sense-disambiguated. Secondary breaks depend on block length, so only the main break is

confirmatory. The lexicon audit, the layer check and the commentary coding were each done by a single coder, so no Cohen's κ can be reported, and the recall of the keyword search is unknown.

Co-occurrence stands in for interaction, ambiguous terms of address received only rule-based substitution, emotion is measured lexically, events are coded from titles, and phonetic relations follow modern Mandarin. The e-text hypotheses were written after exploratory analysis; the Cheng checks are independent but use the same pipeline. Further work could apply the pipeline to hand-collated Zhiyan manuscripts and Cheng A facsimiles, add syntactic and large-language-model features, and have two coders judge homophonic readings across all 120 chapters [20].

# Conclusion

In the Gutenberg e-text and the Cheng A and Cheng B prints, the single change point of 34 function characters falls after chapter 80. The e-text's edition seam lies at the same place but accounts for little of the difference. The break consists of fewer Classical characters, more interrogative 么 and 嗎 and fewer 笑道: and 鳳姐儿, with 嗎 and 林黛玉 edition-sensitive. Several of chapters 81 to 91 resemble the first 80, the last 40 divide again around chapter 90, and chapter 67 leans later in all three texts. Network density is unchanged while household figures gain brokerage, and sadness turns after chapter 94. Authorship notes should name their edition, and digital classics should carry base-text and conversion metadata.

# Funding

This work was not supported by any funds.

# Acknowledgements

The authors would like to show sincere thanks to those technicians who have contributed to this research.

# Conflicts of Interest

The authors declare no conflict of interest.

# References

[1] Yu, Y., Liu, W., Feng, Y. (2022) A quantitative study on *Dream of the Red Chamber*: word-length distribution and authorship attribution. *Complexity*, 2022(1), 9077360.

[2] Hu, X., Wang, Y., Wu, Q. (2014) Multiple authors detection: a quantitative analysis of *Dream of the Red Chamber*. *Advances in Adaptive Data Analysis*, 6(4), 1450012.

[3] Matteson, D. S., James, N. A. (2014) A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505), 334-345.

[4] Zhu, H., Lei, L., Craig, H. (2021) Prose, verse and authorship in *Dream of the Red Chamber*: a stylometric analysis. *Journal of Quantitative Linguistics*, 28(4), 289-305.

[5] Zhao, L., Shi, J., Zhang, C., Liu, Z. (2025) Authorship detection on classical Chinese text using deep learning. *Applied Sciences*, 15(4), 1677.

[6] Burrows, J. (2002) "Delta": a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267-287.

[7] Evert, S., Jannidis, F., Proisl, T., Reger, I., Pielström, S., Schöch, C., Vitt, T. (2017) Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32(Suppl. 2), ii4-ii16.

[8] Plecháč, P. (2021) Relative contributions of Shakespeare and Fletcher in Henry VIII: an analysis based on most frequent words and most frequent rhythmic patterns. *Digital Scholarship in the Humanities*, 36(2), 430-438.

[9] Peksoy, E. (2026) Who wrote the penultimate chapter of *The Female Quixote*: a stylometric

authorship attribution study. *Digital Scholarship in the Humanities*, 41(3), 1579-1593.

[10] Yang, Y., Lyu, G. (2025) A stylometric analysis on authorship attribution of Marriage Bonds to Awaken the World. *Humanities and Social Sciences Communications*, 12(1), 1822.

[11] Da, N. Z. (2019) The computational case against computational literary studies. *Critical Inquiry*, 45(3), 601-639.

[12] Hill, M. J., Hengchen, S. (2019) Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4), 825-843.

[13] Chan, K. Y. (2020) Love in dreams and illusions: fate and prognostication in *Hongloumeng*. *Comparative Literature: East and West*, 4(1), 28-44.

[14] Zhao, E. Y. (2023) The known ending: narrative closure and reading experience in *Dream of the Red Chamber*. *International Communication of Chinese Culture*, 10(2-4), 171-179.

[15] Seng, Y. C. (2016) A study on homophonic and puns in *Dream of the Red Chamber*. *Journal of Chinese Literature and Culture*, 4(2), 85-95.

[16] Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289-300.

[17] Eder, M. (2016) Rolling stylometry. *Digital Scholarship in the Humanities*, 31(3), 457-469.

[18] Clauset, A., Newman, M. E. J., Moore, C. (2004) Finding community structure in very large networks. *Physical Review E*, 70(6), 066111.

[19] Moretti, F. (2011) Network theory, plot analysis. *New Left Review*, 68, 80-102.

[20] Huang, B., Chen, C., Shu, K. (2025) Authorship attribution in the era of LLMs: problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter*, 26(2), 21-43.

*Weixiang Gan*, male, Ph.D., is a postdoctoral researcher and serves as Chief International Business Expert, Chief Corporate Academic Advisor, and Senior Corporate Training Mentor at Zhenguigu Technology Co., Ltd. He is also Associate Editor of the Asia-Pacific Economic and Management Review. His research interests include consumer behaviour, marketing psychology, AI-enabled enterprise digital transformation, AI-driven big data forecasting, ethics, human intelligence studies, corporate internal control, SME internationalisation, higher education curriculum theory, regional economics, industrial economics, sixth-sector industrial integration, and green innovation.

*Mengfei Xiao*, male, Ph.D., is a postdoctoral researcher and entrepreneur. His research interests include tourism management, human resource management, entrepreneurship, international trade, public relations management, and intellectual property management.

# Appendix

Table A1. Function characters and speech formats with large differences between the two parts (Gutenberg e-text).

| Feature | Chs.1 to 80 | Chs.81 to 120 | d[95%CI] | q | d, ChengA |
|---|---|---|---|---|---|
| A Function characters (prose layer, per 1,000) | / | / | / | / | / |
| 方 | 1.65 | 0.48 | -1.71 [-2.28, -1.40] | ≤0.002 | -1.75 |
| 皆 | 0.76 | 0.06 | -1.08 [-1.33, -0.80] | 0.007 | -1.08 |
| 亦 | 0.72 | 0.14 | -0.92 [-1.19, -0.60] | 0.014 | -0.83 |
| 之 | 3.53 | 1.16 | -0.86 [-1.20, -0.67] | 0.029 | -0.75 |
| 其 | 0.70 | 0.24 | -0.59 [-0.76, -0.33] | 0.075 | -0.53 |
| 么 | 4.13 | 7.05 | 1.32 [0.86, 1.90] | 0.003 | 1.22 |
| 嗎 | 0.00 | 0.31 | 1.48 [1.19, 2.09] | ≤0.002 | 1.43 |
| 么 and 嗎 | 4.13 | 7.37 | 1.43 [0.97, 2.02] | ≤0.0002[a] | 1.32 |
| B Speech tags (narration layer, per 1,000) | / | / | / | / | / |
| 笑道 | 9.79 | 1.97 | -1.51 [-1.99, -1.24] | ≤0.002 | / |
| 問道 | 0.66 | 1.27 | 0.80 [0.18, 1.42] | 0.019 | / |
| 道: | 26.39 | 27.14 | 0.06 [-0.54, 0.56] | 0.862 | / |
| C Proportions (%) | / | / | GEE OR [95% CI] | p | / |
| 笑道: / X 道: | 36.72 | 7.34 | 0.13 [0.10, 0.17] | <0.001 | / |
| 鳳姐儿 / 鳳姐 | 36.50 | 6.79 | 0.12 [0.07, 0.22] | <0.001 | / |
| 林黛玉 / 黛玉 | 28.78 | 1.66 | 0.13 [0.04, 0.43] | <0.001 | / |

Note: CI: moving-block bootstrap; q: block permutation with Benjamini-Hochberg correction; OR: odds ratio; "≤" marks the smallest attainable value. [a]Uncorrected block-permutation p. Source: Author's calculations from Project Gutenberg eBook #24264.

Table A2. Authorship and change-point tests in the three texts.

| Test | E-text | Cheng A | Cheng B |
|---|---|---|---|
| LOO AUC [95% CI] | 0.988 [0.963, 1.000] | 0.987 [0.964, 1.000] | 0.983 [0.954, 1.000] |
| Misclassified chapters | 23, 67 | 67, 85, 90 | 67, 72, 85, 90 |
| Buffered LOO AUC | 0.975 | 0.969 | 0.973 |
| Delta T (block p) | 0.140 (≤0.0002) | 0.129 (≤0.0002) | 0.102 (≤0.0002) |
| Change point $\hat{k}$ (maximum Q) | 80 (47.449) | 80 (44.529) | 80 (35.552) |
| p: chapter-wise / block 5 / block 10 / shift | ≤0.001 / ≤0.001 / ≤0.001 / 0.008 | ≤0.001 / 0.002 / ≤ 0.001 / 0.008 | ≤0.001 / 0.004 / 0.004 / 0.008 |
| Sentence-block 90% interval (share at 80) | [80, 80] (0.984) | [18, 80] (0.554) | [18, 80] (0.191) |
| Without chs. 1 to 5, 17, 18: 90% interval | [80, 80] | [80, 80] | [80, 91] |
| Step vs. quadratic trend, AR(1) bootstrap p | 0.010 | 0.035 | 0.078 |
| $\hat{k}$ within chs. 1 to 80; 81 to 120 (block p) | 18 (0.033); 90 (0.006) | 18 (0.004); 90 (0.004) | 18 (0.012); 90 (0.007) |
| Rolling crossover [95% interval] | 89 [85, 91] | 87 [84, 91] | 90 [88, 92] |
| Chs. 81 to 91 with P*(B)* < 0.5 | 7 | 7 | 8 |
| Chs. 81 to 91 on the 1 to 80→92 to 120 axis | 0.644 [0.531, 0.730] | 0.595 [0.486, 0.682] | 0.520 [0.389, 0.616] |
| Ch. 67: indicators near chs. 81 to 120 (of 6) | 4 | 4 | 4 |

Note. Block length 5 unless stated; "≤" marks the smallest attainable value; the AR(1) test uses PC1 to PC5. Source: Project Gutenberg eBook #24264; Wikisource Cheng A and Cheng B prints.