Shannon's letter statistics hold on old and new English, and his third-order value reads high

A preregistered retest of the 1951 entropy table on books from 1788 to 1925 and on present-day Wikipedia, with a planted Markov text and a shuffled-text control

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F508 in the findings ledger. Signal (tier B).

Abstract

In 1951 Claude Shannon published a table of how many bits each letter of printed English carries when the reader knows 0, 1 or 2 letters before it. We recomputed that table on three pooled corpora: books from 1788 to 1859, fiction from 1892 to 1925, and present-day Wikipedia featured articles, each reduced to 26 letters and a space. The single-letter and letter-pair rows match Shannon's within 0.12 bits on all three corpora. His third-order value of 3.1 bits sits 0.39 bits above ours on the old books and 0.23 above ours on Wikipedia. We also trained a letter-count reader on each corpus and scored it on held-out text. With the corrected split, the reader needs 1.63 to 2.41 bits per letter on its own kind of text and never reaches Shannon's 1.3. Within fiction, a change of era costs the reader 0.10 and 0.25 bits per letter, and a change from fiction to The Federalist in the same era costs 0.44, so the preregistered claim that era costs more than genre failed. A planted Markov text with known entropy was recovered within 0.004 bits.

1 What this means for a reader

Shannon's 1951 table is still quoted as a fact about English. The rows built from single letters and letter pairs hold on text two centuries apart. The row built from letter triples comes out about a third of a bit lower than his, so quote the higher rows with care. A reader that knows only letter counts stays well above his 1.3 bits per letter. His figure describes a human reader who brings knowledge of words and of the world.

2 Introduction

Shannon [1] described two ways to measure how predictable English is. The first counts letters. Let the text be reduced to 26 letters and a space, so 27 symbols. The N-gram entropy is the uncertainty of the next letter when the N minus 1 letters before it are known; F1 uses single-letter frequencies, F2 knows one letter back and F3 two. All are in bits per letter, the average number of yes-or-no answers needed per letter. Blind choice among 27 symbols costs 4.75 bits. His table for 27 symbols reads 4.76, 4.03, 3.32 and 3.1 for F0 to F3, and 2.14 bits per letter from word statistics. The second way asks a person to predict the next letter of a passage. From that experiment he put an upper bound near 1.3 bits per letter on the entropy of English when 100 letters of context are known.

Later work kept both routes. Cover and King [2] replaced guessing with betting, so a subject's stake on each letter gives a probability, and reported estimates near 1.3 bits per character. Brown and colleagues [3] built a word trigram model, measured its cross-entropy on a balanced sample of English and gave an upper bound of 1.75 bits per character. Takahira, Tanaka-Ishii and Debowski [4] compressed large corpora in six languages with prediction by partial matching [5] and extrapolated the code length to infinitely long text to estimate the entropy rate itself.

Those studies push the estimate of the entropy rate downward. Our question is narrower. Do the low rows of

Shannon's table, which are easy to recompute, still hold on text of other eras and kinds? And when a reader built from letter counts is trained on one kind of text and reads another, which costs more, a change of era or a change of genre?

3 Preregistration and design

A preregistration is a dated plan fixed and hash-locked before any result is seen. This study had two. The first was locked on 1 November 2026 at 08:13 UTC. It named the corpora, the statistics, the reader, the planted text, the control, nine numeric forecasts and the kill condition, a rule that decides against the run with no judgement after it. The decision it informed was which claims from saved reading material survive a changed frame, so which ones we can reuse.

The first run had a fault of ours. The code promised in a comment to shuffle the books before splitting training text from test text and never did. The old corpus's test text was therefore all The Federalist, and the mid corpus's all Fitzgerald, so era was read through one book. We wrote a second preregistration to fix the split and locked it at 08:20 UTC the same day. It was not blind: it was written after the first run had been read. The first run stands as recorded, and both are reported here.

Claims and kill condition, as locked:

- Claim 1, first run: the 1951 table holds on present data within 0.15 bits for F1, 0.2 for F2 and 0.3 for F3.
- Claim 2, first run: the reader reaches 1.3 bits per letter. The forecast said it would not.
- Claim, second run: era costs a reader more than genre does. The forecast said it would not.
- Void if the reader misses the entropy of the planted text by more than 0.1 bits.

4 Data and provenance

The books are public-domain texts from Project Gutenberg [6]: The Federalist (1788), Austen (1813), Bronte (1847), Melville (1851) and Dickens (1859) make the old corpus; Doyle (1892), Joyce (1914 and 1916), Wodehouse (1920) and Fitzgerald (1925) make the mid corpus. The new corpus is a sample of Wikipedia featured articles [7], about 0.77 MB, fetched through the Wikipedia API on 1 November 2026. Wikipedia text is licensed CC BY-SA; we use it for statistics only and release no text from it. The source record names each file and its size; the text files themselves were not hashed at download, so a reproduction may draw slightly different editions.

Each text was cut to the Project Gutenberg start and end markers, lowercased, and every run of characters outside a to z was replaced by a single space. In the first run each corpus takes an equal slice from each of its books, 750,000 characters in all, and the first 600,000 train the reader. In the second run every book gives its first 80 percent to training and its last 20 percent to test, and the sets are kept apart by genre: old fiction (four novels), mid fiction (five works), old nonfiction (The Federalist) and new nonfiction (Wikipedia). Each set has 600,000 training characters, except Wikipedia with 588,712, and 100,000 test characters.

Shannon's own test passages are not openly available, so we compare against his printed table. The table values were recorded in the first preregistration from a scan of the 1951 paper.

5 Methods

The letter statistics

For a block b_i of N minus 1 letters and a next letter j, the N-gram entropy is

$$F_N = -\sum_{i,j} p(b_i, j) \log_2 p(j \mid b_i) \tag{1}$$

with probabilities taken as relative frequencies in the pooled corpus. This is a plug-in estimate: observed frequencies go straight into the formula with no correction for sample size. F0 is $\log_2 27$, 4.755 bits.

The word-based figure divides the entropy of the word frequency distribution, H_w, by the mean number of characters per word including the following space, L:

$$F_w = \frac{H_w}{L} \tag{2}$$

Shannon reached his 2.14 by extending a fitted Zipf's law tail, the rule that a word's frequency falls in inverse proportion to its rank. We counted words directly.

The reader

The reader is a context model: it predicts the next letter from counts of what followed the same preceding letters in training text, in the manner of prediction by partial matching. It blends contexts of length 0 to 6. Starting from a uniform distribution over 27 symbols, each context length k with n_k counts over d_k distinct next letters updates the estimate as

$$p_k(x) = \frac{c_k(x) + d_k p_{k-1}(x)}{n_k + d_k} \tag{3}$$

where $c_k(x)$ is the count of letter x after that context. The reader is scored by its log loss, the average of minus $\log_2$ of the probability it gave to the letter that came next:

$$R = -\frac{1}{n} \sum_i \log_2 q(x_i | c_i) \tag{4}$$

over n letters of held-out text, text the reader was not trained on. The first run scores 40,000 letters of each test text and the second 30,000.

The reader also plays Shannon's guessing game: it ranks the 27 symbols and we record the rank r of the true letter. With q_r the share of letters found at rank r, Shannon's upper and lower bounds are

$$U = -\sum_r q_r \log_2 q_r \tag{5}$$

$$B = \sum_r r(q_r - q_{r+1}) \log_2 r \tag{6}$$

Transfer penalties

A transfer penalty is the extra log loss a reader pays on another kind of text, against a reader trained on that kind, on the same held-out passage. With $R(a, b)$ the log loss of the reader trained on set a reading the test text of set b,

$$D(a, b) = R(a, b) - R(b, b) \tag{7}$$

The era penalty in fiction uses a and b from old and mid fiction. The genre penalty uses old fiction and The Federalist, which share an era.

6 Controls and planted checks

A planted check is data built so the right answer is known. Ours is a random order-2 Markov chain over 27 symbols, 2,000,000 symbols long, with transition probabilities drawn from a Dirichlet distribution. Its entropy per symbol is 2.673 bits by eq. 1 with N of 3. A reader of context length 2 was trained on the first half of the chain and scored on 20,000 later symbols. It missed the true entropy by 0.004 bits, against a kill threshold of 0.1. The void condition did not fire.

A control is a version of the data with the structure under test removed and scored the same way. Here the held-out test text has its characters shuffled 20 times, 4,000 letters each, and the trained reader reads each shuffle. The real text should cost far less than any shuffle. In the first run the shuffles cost a mean of 8.51 bits per letter against 2.73 for the real Wikipedia text; in the second, 9.26 against 1.93 for old fiction. The shuffles cost more than 4.75 bits because the reader is confident and wrong on scrambled text. The real text beat all 20 shuffles in both runs. The p value from draws, $p = (k + 1)/(n + 1)$ with k shuffles at or below the real loss out of n, is 0.048, the smallest value 20 draws allow.

7 Results

The table

Table 1 and Figure 1 set Shannon's table against ours. The first run computed the letter statistics on each pooled corpus.

Table 1. Shannon's 1951 table for 27 symbols against the three corpora, in bits per letter. The gap is to Shannon's value, for the old and new corpora.

Statistic	Shannon 1951	Old	Mid	New	Gap old, new
F1	4.03	4.085	4.067	4.113	0.055, 0.083
F2	3.32	3.345	3.324	3.440	0.025, 0.120
F3	3.1	2.712	2.698	2.872	0.388, 0.228
Word-based	2.14	1.801	1.853	1.747	0.339, 0.393



Figure 1. Shannon's 1951 values against the three corpora. The first two groups agree within 0.12 bits; the third and fourth sit lower than his. Drawn from the first run's result file and the preregistered table values.

The single-letter and letter-pair rows match Shannon's within 0.12 bits on all three corpora. The largest of those gaps is F2 on Wikipedia, 0.120. His third-order value of 3.1 bits sits 0.39 bits above ours on the old books and 0.23 above ours on Wikipedia. The first-run verdict on the old table failed on that row alone, and the verdict on the new table passed, since 0.228 is inside the 0.3 bound. The word-based figure is 0.34 to 0.39 bits below his. Words in the Wikipedia sample are longer, 5.92 characters with the space against 5.41 in the old books, and their frequency distribution carries 10.34 bits per word against 9.74.

The reader

In the first run the reader trained on the old corpus needed 2.13 bits per letter on the old held-out text, inside the forecast band of 1.85 to 2.35. Because of the faulty split, that held-out text was all The Federalist. The guessing game gave bounds of 1.34 and 2.35 by eqs. 6 and 5, and the reader's first guess was right on 58 percent of letters. The claim that this reader reaches 1.3 bits failed, as forecast.

The second run, with the split fixed, gives the matrix in Figure 2. With the corrected split, the reader needs 1.63 to 2.41 bits per letter on its own kind of text and never reaches Shannon's 1.3. The Federalist is the easiest text for its own reader, 1.63, and Wikipedia the hardest, 2.41.

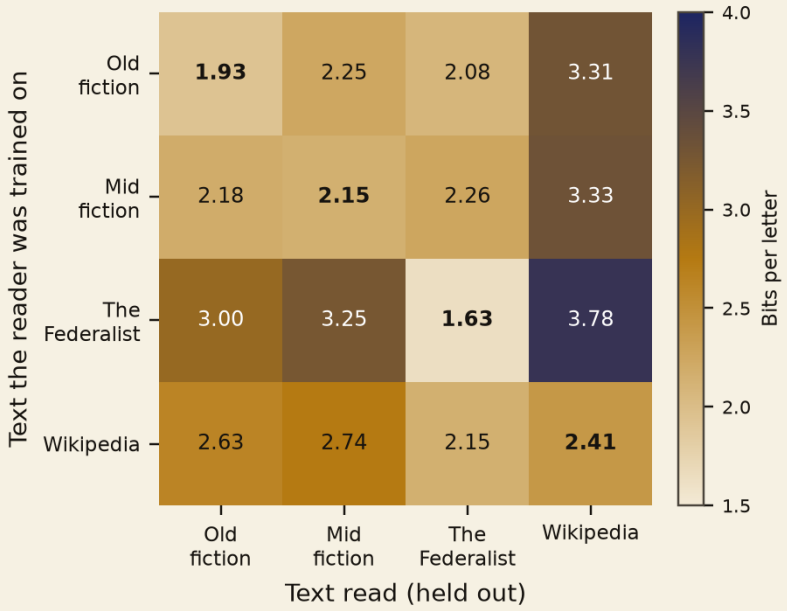


Figure 2. Log loss of each reader on each held-out text, second run. The diagonal is each reader on its own kind of text. Drawn from the second run's result file.

Table 2 gives the transfer penalties of the second run against their forecasts.

Table 2. Transfer penalties in bits per letter, second run, by eq. 7, against the preregistered forecasts.

Change	Forecast	Band	Measured
Era, old fiction reader on mid fiction	0.1	0.02 to 0.3	0.100
Era, mid fiction reader on old fiction	0.1	0.02 to 0.3	0.253
Genre, old fiction reader on The Federalist	0.5	0.2 to 0.9	0.444
Era and topic, Federalist reader on Wikipedia	1	0.4 to 1.6	1.365

Within fiction, a change of era costs the reader 0.10 and 0.25 bits per letter, and a change from fiction to The Federalist in the same era costs 0.44, so the preregistered claim that era costs more than genre failed. We forecast that failure. A mid-fiction reader on Wikipedia pays 0.92, where era, genre and topic all change at once.

Of the numeric forecasts, 4 of 9 held in the first run and 4 of 5 held after the fix. Appendix A lists each one.

8 Discussion

The low rows of Shannon's table describe English letters across two centuries of text. Single-letter frequencies

and letter-pair statistics move by a few hundredths of a bit between Austen and Wikipedia. That is the part of the table a reader can quote.

F3 is different. Our value is lower than his on every corpus, by more than our forecast allowed on the old books. We have not found the cause. Two are open, and neither was tested. A plug-in estimate falls below the true value as the number of contexts grows against the sample, and 27 squared, 729, two-letter contexts is many for 750,000 characters if rare pairs matter. Shannon's F3 also rests on letter counts from other text than ours, and a difference in the source counts would move it. Either way the third row should be quoted with its source and method.

The word-based gap fits a difference of method. Counting words directly misses the long tail of rare words that a fitted Zipf curve adds, and a missing tail lowers the entropy. We did not refit his tail to confirm it.

The reader result agrees with the line of work after Shannon [2] [3]: letter counts alone do not reach a human reader's figure. Our context model sits at 1.63 to 2.41 bits per letter on held-out text. Shannon's 1.3 is a bound on a reader who knows words, grammar and the world.

The transfer result was the frame we most wanted to change. A reader from one era of fiction reads another era at a cost of 0.10 to 0.25 bits, and the same reader reads The Federalist at 0.44. What a text is about and how it is argued changes letter statistics more than when it was written.

9 Limitations

- Wikipedia is the only new text, so era, genre and topic stay tangled in every new-text row.
- No large language model was used as a reader. That was the change of frame we most wanted, and it stays open.
- The word-based gap may come from Shannon's fitted Zipf tail against our direct count. We did not test that.
- The cause of the F3 gap is unknown. Sampling bias in the plug-in estimate and a difference in his source counts are both candidates, and neither was measured.
- Shannon's own test passages are not openly available, so the comparison is with his printed table.
- The first run's split was faulty, and the second preregistration was written after reading the first run, so it is not blind.
- Scoring was done by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 every F1, F2, F3 and word-based value in Table 1, the gaps, the planted entropy and its gap of 0.004, the reader's log losses, bounds and first-guess share, the penalties in Table 2, the shuffle control means and p values, and every verdict in Appendix A were re-read from the two runs' result files. They match the findings ledger. The forecast counts 4 of 9 and 4 of 5 were counted from the verdicts in those files. Both preregistrations' SHA-256 sums were recomputed and match the hashes recorded in the result files. Shannon's table values are as recorded in the first preregistration from the paper scan; we did not re-read the scan for this paper. Both figures were drawn by a script that reads only the result files and the first preregistration. The study was not re-run. The literature in the introduction was found by searches run on 9 November 2026 and read at the level of abstracts and publisher pages; the full texts of the cited papers were not read for this paper.

11 Corrections

- 1 November 2026: the first run's test text for the old and mid corpora was a single book each, because a promised shuffle was never done. A second preregistration fixed the split and was run the same day. The

first run stands as recorded and both are reported.

- 9 November 2026: first published version.

12 Reproduction

- Data: the Project Gutenberg books listed in the data section, and a sample of Wikipedia featured articles of about 0.77 MB.
- Strip the Project Gutenberg header and footer, lowercase, and map every run of characters outside a to z to a single space.
- Compute eq. 1 for N of 1 to 3 and eq. 2 on each pooled corpus, and compare with Shannon's table.
- Train the context model of eq. 3 with contexts up to 6 letters on 600,000 characters of each set, and score eq. 4 on held-out text of every set. For the second run, take the first 80 percent of each book for training and the last 20 percent for test.
- Check the reader on a planted order-2 Markov chain, and against 20 character shuffles of the held-out text.
- First run: preregistration SHA-256 551f636e8364984be905e2266229c4b94f031c3f6f7a4262c9934f377e2e3344. Plan SHA-256 0da1855b4a3831bcf4cca09a1493645a2c81601f2d3a83a4a4722b752bdd3a97. Run 20261101T081322Z, seed 20261113.
- Second run: preregistration SHA-256 70d161373422827c33dd0682d33e31a36f9ee03ec974b228dd0c7d41afdb90f8. Plan SHA-256 41242bb227893e5d7b39c5cd863bc0d4729d132a335f15a58fcd3f9514ff5903. Run 20261101T082050Z, seed 20261119.
- Cost: on one desktop CPU the analyses took 30 and 21 seconds, and each set of controls under 2 seconds.

13 Appendix A. The preregistered verdicts

Each verdict is the pass or fail of rules written before the run, read off by code. A verdict passes only when every rule holds. Values are read from the result files.

Table 3. First run, locked 1 November 2026 at 08:13 UTC.

Verdict	Deciding rule	Value	Holds
Instrument works	planted gap at most 0.1	0.004	yes
Table holds, old	gaps F1 below 0.15, F2 below 0.2, F3 below 0.3	0.055, 0.025, 0.388	no
Table holds, new	gaps F1 below 0.15, F2 below 0.2, F3 below 0.3	0.083, 0.120, 0.228	yes
Reader reaches 1.3	old reader on old text at most 1.3	2.130	no
Forecast, own-era loss	1.85 to 2.35	2.130	yes
Forecast, guessing upper bound	1.9 to 2.5	2.350	yes
Forecast, word-based gap, new	0.1 to 0.5	0.393	yes
Forecast, era penalty, old reader on new	0.1 to 0.5	0.735	no
Forecast, era penalty, new reader on old	0.1 to 0.6	0.084	no
Forecast, genre penalty	0.1 to 0.45	0.661	no
Forecast, new against old own-text loss	-0.1 to 0.25	0.599	no
Forecast, word-order information	0.4 to 1.2	0.498	yes
Forecast, gain from four times the training text	0.1 to 0.4	0.450	no
Shuffle control	real loss below every shuffle's 95th percentile	2.73 against 8.44	yes

The first run's era and genre values were read through one test book per corpus and are superseded by the second run for the era and genre question. Word-order information is the extra loss when the words of the Wikipedia test text are shuffled. The training gain is the extra loss of a Wikipedia reader trained on a quarter of the text.

Table 4. Second run, locked 1 November 2026 at 08:20 UTC, not blind.

Verdict	Deciding rule	Value	Holds
Era costs more than genre	era penalty, old to mid fiction, above genre penalty, old fiction to The Federalist	0.100 against 0.444	no
Forecast, own-text fiction loss	2.0 to 2.5	1.931	no
Forecast, era, old to mid	0.02 to 0.3	0.100	yes
Forecast, era, mid to old	0.02 to 0.3	0.253	yes
Forecast, genre	0.2 to 0.9	0.444	yes
Forecast, Federalist reader on Wikipedia	0.4 to 1.6	1.365	yes
Shuffle control	real loss below every shuffle's 95th percentile	1.93 against 9.20	yes

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **N-gram entropy**. The uncertainty of the next letter when the N minus 1 letters before it are known, averaged over the text; Shannon wrote it F with a subscript N, so F1 uses single-letter counts, F2 one letter of context and F3 two. Field: information theory, after Shannon 1951. Excludes the entropy of the language as a whole, which is the limit as N grows.
- **bits per letter**. The average number of binary digits needed to write down each letter of a text under a given model of the text; one bit halves the number of equally likely choices, so 4.75 bits per letter is the cost of picking blindly among 27 symbols. Field: information theory. Excludes a count of bytes in a file, and any statement about meaning.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **kill condition**. A rule written in the preregistration that, when met, decides against the claim with no room for judgement after the run. Field: research method (our usage, after Popper's falsification). Excludes reasons found after the run for setting a result aside.
- **plug-in estimate**. An entropy computed by putting the observed relative frequencies of a sample straight into the entropy formula, with no correction for the size of the sample. Field: statistics. Excludes corrected estimators that add a term for unseen or rare events.
- **Zipf's law**. The observation that the frequency of a word falls roughly in inverse proportion to its rank in the frequency list. Field: quantitative linguistics. Excludes an exact law that holds for every word or every text.
- **context model**. A model that predicts the next symbol from counts of what followed the same few preceding symbols in training text, blending long and short contexts; prediction by partial matching (PPM) is the classic form. Field: data compression and statistical language modelling. Excludes models that use word meaning, grammar or knowledge outside the letter counts.
- **log loss**. The average of minus the base-2 logarithm of the probability a model gave to the letter that came next, in bits per letter; it is the length of the code an ideal coder driven by the model would write. Field:

information theory and machine learning; also called cross-entropy. Excludes the share of letters guessed right, and the entropy of the text itself, which log loss can only bound from above.

- **held-out text**. Text kept apart from the text a model was trained on and used only to score it. Field: statistics and machine learning. Excludes any text whose counts entered the model.
- **guessing game**. A reader guesses the next letter in order of preference until right; the share of letters found at the first, second and later guesses gives an upper and a lower bound on the entropy. Field: information theory, after Shannon 1951. Excludes a direct measure of the probabilities a reader holds.
- **transfer penalty**. The extra log loss a reader pays on another kind of text compared with a reader trained on that kind of text, both scored on the same held-out passage. Field: our usage, from machine learning practice on domain shift. Excludes the difference in difficulty between two texts read by their own readers.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **order-2 Markov chain**. A random text in which each symbol depends only on the two symbols before it, through a fixed table of probabilities, so its entropy per symbol can be computed exactly. Field: probability. Excludes real text, whose dependence reaches further back.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental design. Excludes a comparison with a different dataset or with a published number.
- **p value (from draws)**. The share of control draws scoring at or above the real score, with one added to the count and to the number of draws, so it is never zero. Field: statistics. Excludes a p value from a textbook formula that assumes a distribution.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field: research method (our usage). Excludes an interpretation written after reading the numbers.

15 References

[1] Shannon, C. E. (1951). Prediction and entropy of printed English. Bell System Technical Journal 30(1), 50-64. doi:10.1002/j.1538-7305.1951.tb01366.x

[2] Cover, T. M. and King, R. C. (1978). A convergent gambling estimate of the entropy of English. IEEE Transactions on Information Theory 24(4), 413-421. doi:10.1109/TIT.1978.1055912

[3] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C. and Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English. Computational Linguistics 18(1), 31-40. ACL Anthology J92-1002.

[4] Takahira, R., Tanaka-Ishii, K. and Debowski, L. (2016). Entropy rate estimates for natural language: a new extrapolation of compressed large-scale corpora. Entropy 18(10), 364. doi:10.3390/e18100364

[5] Cleary, J. G. and Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. IEEE Transactions on Communications 32(4), 396-402. doi:10.1109/TCOM.1984.1096090

[6] Project Gutenberg. Public-domain texts in the United States. https://www.gutenberg.org. Retrieved 1 November 2026.

[7] Wikipedia. Featured articles, English Wikipedia, licensed CC BY-SA. https://en.wikipedia.org. Retrieved 1 November 2026 through the Wikipedia API.