# Can Large Language Models Approximate Human Editorial Judgment?

A Novel Benchmark for Back-of-the-Book Indexing

Paul Estrada[1]

[1]IndexerLabs

May 2026

**Abstract**

Professional subject indexing has long seemed like a natural target for automation. At first glance, it appears to be mainly a problem of identifying the important topics in a book and organizing them for retrieval. Yet high-quality back-of-the-book indexing has resisted full automation because the task depends on editorial judgment rather than term frequency alone. A useful index begins with a curated selection of the concepts, people, places, events, and themes that readers are likely to seek out. If the wrong topics are chosen, later improvements to locators, subentries, cross-references, or formatting cannot make the index genuinely useful.

This paper isolates that first problem: whether language models can determine what deserves to be indexed. Rather than evaluating or creating a complete AI-generated index all at once, we test topical selection directly through a controlled pruning benchmark. Each model begins with an intentionally overgenerated pool of 7,400 possible top-level index entries for William Doyle's *The Oxford History of the French Revolution* and reduces it to a human-scale target close to the size of the original printed human index. Performance is measured by the overlap between the AI-pruned entry list and the published human index.

Across 150 runs, five models were tested with 30 independent runs each. At the human-scale checkpoint, IndexLM-1.0, a purpose-trained indexing model fine-tuned on more than 1,000 real-world back-of-the-book indexes, retained approximately 57% of the human index's top-level entries. Claude-Opus-4.6 retained 29%, Gemini-3.1-Pro retained 18%, Mistral-3-Large retained 16%, and GPT-5.4-xhigh retained 15%. These results suggest that LLMs can preserve the topical core of a professional index to a measurable degree, while domain-specific fine-tuning substantially improves performance under realistic length constraints.

## 1 Introduction

Back-of-the-book indexing is often mistaken for a mechanical task. A book contains names, topics, and concepts; an index lists them alphabetically for readers. In practice, however, a usable index is not a concordance but a curated representation of the book's conceptual structure. The indexer must decide which topics are central, which mentions are incidental, which terms readers are likely to seek out, and how much detail can fit within the practical limits of a printed index.

This selection problem is the starting point of indexing quality. Locators, subentries, cross-references, and formatting matter, but they can only operate on the topics that have already been selected. An index that omits the book's central people, places, institutions, events, and concepts cannot become useful merely by improving its page references or presentation.

Recent evaluations of AI-generated book indexes, especially those by the AI Committee of the American Society for Indexing, have focused on whether general-purpose chatbots can produce complete professional-quality indexes directly. Those evaluations are important, but they combine several distinct problems into one task: topic selection, locator accuracy, subentry structure, cross-reference generation, and formatting. This paper isolates the first of those problems: knowing what to index.

We study this question using a 150-run pruning benchmark on William Doyle's *The Oxford History of the French Revolution*. The benchmark begins with 7,400 candidate top-level entries and asks five large language models to prune the list to a human-scale target close to the size of the printed human index. The paper investigates two questions:

1. Can large language models preserve the human-selected topical core of a professional index to a measurable degree?

2. Does a purpose-trained indexing model outperform both general-purpose LLMs and its own non-fine-tuned base model on this topical selection task?

The results suggest that both answers are yes. At the human-scale checkpoint, IndexLM-1.0 retains approximately 57% of the human index's top-level entries, compared with 29% for Claude-Opus-4.6, 18% for Gemini-3.1-Pro, 16% for Mistral-3-Large, and 15% for GPT-5.4-xhigh. The comparison with Mistral-3-Large is especially important because IndexLM-1.0 is fine-tuned from that base model; the gap between 57% and 16% suggests that indexing-specific fine-tuning substantially changes the model's selection behavior.

## 2 Related Work and Prior Evaluations

The most direct recent evaluations of LLM-generated book indexes have been conducted by the AI Committee of the American Society for Indexing [1, 2]. However, their methodology evaluates one particular workflow: a general-purpose chatbot is asked to generate a complete professional-quality index directly in one attempt. In the committee's tests, the models were prompted to create an index for an attached book or chapter, including cross-references and subheadings; later tests also supplied indexing criteria and prior AI-indexing analysis as guidance [1, 2].

This is a valid stress test of one-shot chatbot indexing, but it is also an extremely weak version of the technology. In that workflow, the model must decide what to index, locate and read relevant passages, generate subentries, create cross-references, format the result, and check its own work in a single pass. A more capable LLM system would instead use an agentic harness: a surrounding workflow that decomposes the task into smaller stages, maintains state across steps, calls external tools or APIs, validates intermediate outputs, retries failed stages, and routes uncertain cases to further review. Prior work on language-model systems supports this thesis. ReAct shows that interleaving model reasoning and action can improve multi-step decision-making [13]. Self-Refine and Reflexion show that feedback and revision can improve outputs beyond single-pass generation [14, 15]. Toolformer and SWE-agent show that tool use and interface design can substantially affect model performance [16, 17]. More recent work on agent harnesses and reusable tool-composition skills likewise treats the harness itself as part of the system being evaluated [18, 19, 20].

For this reason, the present paper does not treat one-shot chatbot indexing as a decisive test of LLM capability. The ASI results are important evidence that direct, one-shot chatbot generation does not yet produce professional-quality indexes. They do not evaluate staged, tool-supported, purpose-trained, or agentic indexing systems. This paper therefore treats indexing as a set of

separable components. We divide the task into two broad stages: first, knowing what to index; second, indexing those selected topics through locators, subentries, cross-references, and formatting. This paper evaluates only the first stage by asking whether language models can sift through a deliberately noisy candidate pool and preserve the same top-level topics selected by a professional human indexer.

# 3 Models

## 3.1 Purpose-trained model: IndexLM-1.0

IndexLM-1.0 is the purpose-trained indexing language model tested in this benchmark. It was developed by IndexerLabs for back-of-the-book subject indexing. IndexLM-1.0 is a variant of Mistral-3-Large, fine-tuned on more than 1,000 real-world indexes from the open web, totaling roughly 250 million tokens of training data [3]. To avoid training-data leakage, the fine-tuning corpus excluded William Doyle's *The Oxford History of the French Revolution* and its published index. The experiment therefore evaluates generalization to an unseen indexing task rather than memorization of the reference index.

## 3.2 Base-model comparison: Mistral-3-Large

Mistral-3-Large is included as the non-fine-tuned base-model comparison for IndexLM-1.0.[10] Since IndexLM-1.0 is fine-tuned from Mistral-3-Large, the performance gap between the two models provides a direct test of whether fine-tuning on professional indexes improves topical selection.

## 3.3 Other general-purpose comparison models

The benchmark also includes three additional general-purpose comparison models: Claude-Opus-4.6, Gemini-3.1-Pro, and GPT-5.4-xhigh. Claude-Opus-4.6 is Anthropic's frontier Opus model in the Claude 4 series [7]; Gemini-3.1-Pro is Google's Pro model in the Gemini 3.1 series [9]; and GPT-5.4-xhigh refers to OpenAI's GPT-5.4 model used with the highest available reasoning-effort setting [5, 6]. These models are capable of general text reasoning, but they were not purpose-trained or otherwise modified for the specific task of subject indexing.

# 4 Experimental Design

## 4.1 Benchmark text

The benchmark uses William Doyle's *The Oxford History of the French Revolution* as its test text. The book is a useful case study because it is a large academic history with many indexable people, institutions, places, events, political concepts, and historical elements. It therefore presents a realistic subject-indexing problem rather than a simple name-extraction or keyword-extraction task. Because many potentially indexable terms appear throughout the text, the central challenge is not merely detecting candidates but deciding which candidates are important enough to survive compression into a usable index.

The published human index for the book contains 1,066 top-level entries. In this study, that printed index is treated as the professional reference index. This does not mean that the printed index is the only possible correct index. Subject indexing contains an unavoidable subjective element: different indexers may make different but defensible decisions about phrasing, granularity, structure, and omission.

However, subjectivity does not make the task arbitrary. Indexer Ullstrom argues that although two indexers may produce different indexes for the same book, there should still be significant overlap because the underlying text remains the same [11]. The same point motivates the present benchmark. If a book's central people, institutions, places, events, and concepts are genuinely important to its structure, then a competent index should preserve many of them, even if the final phrasing or organization differs. The printed index therefore provides a concrete human benchmark against which model-pruned entry lists can be compared.

### 4.2 Candidate pool

The experiment begins with an intentionally overgenerated pool of 7,400 possible top-level index entries for the book. This candidate pool was created in a high-recall manner. Each page of the book was processed with Gemini-3-Flash and prompted to identify all potentially indexable terms [12]. The goal at this stage was not to produce a polished index, but to capture as many plausible candidates as possible.

Importantly, the cleaned candidate pool contained every top-level entry from the published human index. This means the pruning task was not limited by missing reference entries: in principle, a model could have retained 100% of the human index if it selected perfectly. The benchmark therefore evaluates pruning judgment rather than candidate-generation recall.

After this page-level generation step, the candidate pool was manually reviewed against the source text. Each page and proposed entry was checked by hand to remove entries that were not supported by the book, including hallucinated terms or entries derived from model inference rather than textual evidence. This manual review was important because the benchmark is designed to evaluate pruning judgment, not hallucination detection. The candidate pool should therefore be noisy in the sense of containing weak, redundant, overly minor, or passing-mention entries, but it should not contain entries with no basis in the book.

### 4.3 Pruning task

Each model begins with the entire book and the 7,400-entry candidate pool and is asked to reduce it to approximately 1,058 entries. This target was set by the first completed IndexLM-1.0 pruning run, which produced a final list of 1,058 top-level entries. Because the published human index contains 1,066 top-level entries, this target is effectively human-scale: it differs from the printed index size by fewer than ten entries.

To keep the comparison fair, the same target range was then used for the other models. All other models converged to the same final size. This means that the benchmark compares models under roughly equal length constraints rather than allowing one model to preserve more human-index overlap simply by keeping a longer list.

The pruning process is iterative. At each step, the model recommends entries to remove from the current candidate pool. Those removals are applied, the shortened list becomes the input to the next pruning step, and the process continues until the list reaches the target range.

### 4.4 Runs and model comparisons

The benchmark contains 150 total pruning runs. Five models are tested, with 30 independent runs per model.

The repeated-run design is important because LLM outputs can vary across runs. A single run may overstate or understate a model's typical performance. Running each model 30 times allows

the study to compare average behavior across repeated trials rather than relying on one output from each model.

### 4.5 Matching procedure

The evaluation compares each AI-pruned list with the published human index. A match is counted when a retained AI entry corresponds to a top-level entry in the human index.

Some matches are exact string matches. Others require light normalization or adjudication because professional indexes can phrase the same topic in slightly different ways. For example, minor differences in punctuation, word order, pluralization, or naming or phrasing convention should not necessarily be treated as different topics if the entry clearly refers to the same person, place, institution, event, or concept.

### 4.6 Evaluation metric

Let $H$ be the set of top-level entries in the published human index. Let $A_{m,r,k}$ be the set of entries retained by model $m$ in run $r$ when the candidate pool has been pruned to size $k$. The primary metric is human-index retention:

$$R_{m,r,k} = \frac{|H \cap A_{m,r,k}|}{|H|} \tag{1}$$

This metric measures the percentage of the human index's top-level entries that remain in the AI-pruned list at a given checkpoint.

The human-scale checkpoint is the main point of comparison because it represents the moment when the candidate pool has been compressed to approximately the size of the published index. At larger candidate-pool sizes, high overlap is less informative because many entries remain. At the human-scale checkpoint, the model has been forced to make substantial selection decisions.

## 5 Data Description

### 5.1 Unit of observation

The unit of observation in this study is a completed pruning run. Each run corresponds to one model reducing the same 7,400-entry candidate pool to the same final size of 1,058 entries. The dataset contains 150 observations: five models with 30 independent runs per model.

Each observation records the model name, the run identifier, the final number of retained entries, and the number of retained entries that match the published human index. Since every run terminates at 1,058 entries, differences in performance cannot be explained by one model retaining a longer index than another. The comparison therefore focuses on which entries each model chose to preserve under the same length constraint.

### 5.2 Variables

The main variables in the dataset are shown in Table 1.

The primary outcome variable is human-index retention. Since the published human index contains 1,066 top-level entries, retention is calculated as

$$\text{retention} = \frac{\text{human\_matches}}{1066}$$

Table 1: Variables recorded for each pruning run.

| Variable | Description |
|---|---|
| `model` | Model used for the pruning run |
| `run_id` | Independent run identifier, from 1 to 30 for each model |
| `final_size` | Number of entries retained after pruning |
| `human_matches` | Number of retained entries matching the published human index |
| `retention` | Human-index retention, defined as `human_matches`/1066 |

This measures the proportion of the published human index's top-level entries retained by the model-pruned list.

### 5.3 Descriptive statistics

Table 2 summarizes the distribution of human-index matches and retention across the 30 runs for each model.

Table 2: Human-index retention at the human-scale checkpoint.

| Model | Mean matches | SD | Min–Max | Mean retention | 95% CI |
|---|---|---|---|---|---|
| IndexLM-1.0 | 608.0 | 38.2 | 561–655 | 57.04% | 55.70–58.37% |
| Claude-Opus-4.6 | 309.0 | 19.9 | 286–332 | 28.99% | 28.29–29.68% |
| Gemini-3.1-Pro | 192.0 | 13.0 | 177–207 | 18.01% | 17.56–18.47% |
| Mistral-3-Large | 171.0 | 11.4 | 157–185 | 16.04% | 15.64–16.44% |
| GPT-5.4-xhigh | 160.0 | 10.0 | 148–172 | 15.01% | 14.66–15.36% |

These statistics show a clear separation between IndexLM-1.0 and the four non-specialized models. IndexLM-1.0 retained an average of 608 human-index entries, corresponding to 57.04% retention. Claude-Opus-4.6 was the strongest general-purpose comparison model, retaining an average of 309 entries, or 28.99%. Gemini-3.1-Pro, Mistral-3-Large, and GPT-5.4-xhigh retained substantially fewer human-index entries.

The base-model comparison is especially important. Mistral-3-Large retained an average of 171 human-index entries, while IndexLM-1.0 retained 608. Because IndexLM-1.0 is fine-tuned from Mistral-3-Large, this 41.0 percentage-point gap provides direct evidence that indexing-specific fine-tuning changed the model's selection behavior.

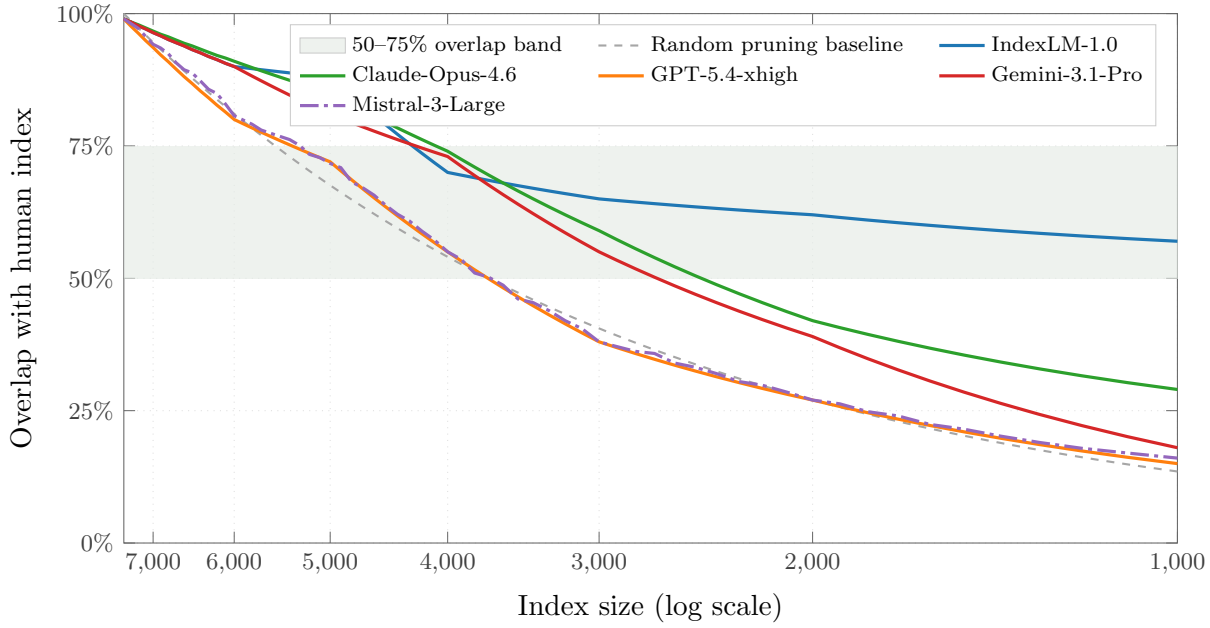
## 6 Results

### 6.1 Human-scale checkpoint

The main comparison is the human-scale checkpoint, where every model has reduced the original 7,400-entry candidate pool to 1,058 entries. Table 2 shows that IndexLM-1.0 retained substantially more of the published human index than any comparison model. Across 30 independent runs, IndexLM-1.0 retained an average of 608.0 human-index entries, corresponding to 57.04% retention. The strongest general-purpose comparison model, Claude-Opus-4.6, retained 309.0 entries on average, or 28.99%. Gemini-3.1-Pro, Mistral-3-Large, and GPT-5.4-xhigh retained 18.01%, 16.04%, and 15.01%, respectively.

The comparison with Mistral-3-Large is the most direct test of the effect of fine-tuning. IndexLM-1.0 is fine-tuned from Mistral-3-Large, so the gap between the two models cannot be explained by a different base model family. At the human-scale checkpoint, IndexLM-1.0 retained 57.04% of the human index, while Mistral-3-Large retained 16.04%. This is a 41.0 percentage-point improvement in human-index retention.

Figure 1: Mean overlap with the published human index as the candidate pool is pruned toward human-scale length, using trajectory data from the earlier 120-run IndexerLabs write-up and the Mistral-3-Large base-model comparison [4].




The shaded band in Figure 1 should not be read as a hard target, but it helps interpret the direction of the result. Very low overlap would suggest that a model is failing to preserve the human-selected topical core. However, overlap much above roughly 80% would also be suspect in this setting: it could indicate that the model was not independently making indexing judgments, but was instead reproducing or memorizing the book's published index.

# 7 Statistical Analysis

## 7.1 Confidence intervals for model means

To compare model performance, we use human-index retention at the human-scale checkpoint. For model $m$ and run $r$, retention is defined as

$$R_{m,r} = \frac{\text{human\_matches}_{m,r}}{1066}$$

where 1,066 is the number of top-level entries in the published human index. For each model, the sample mean retention is

$$\bar{R}_m = \frac{1}{30} \sum_{r=1}^{30} R_{m,r}$$

and the standard error is

$$SE_m = \frac{s_m}{\sqrt{30}}$$

where $s_m$ is the sample standard deviation of the model's 30 retention values. Since each model has 30 runs, we use a $t$ critical value with 29 degrees of freedom:

$$t^*_{29} \approx 2.045$$

Thus, the 95% confidence interval for each model's mean retention is

$$\bar{R}_m \pm t^*_{29} SE_m$$

Table 3 shows the resulting confidence intervals.

Table 3: 95% confidence intervals for mean human-index retention.

| Model | Mean retention | SD | SE | 95% CI |
|---|---|---|---|---|
| IndexLM-1.0 | 57.04% | 3.58% | 0.65% | 55.70–58.37% |
| Claude-Opus-4.6 | 28.99% | 1.86% | 0.34% | 28.29–29.68% |
| Gemini-3.1-Pro | 18.01% | 1.22% | 0.22% | 17.56–18.47% |
| Mistral-3-Large | 16.04% | 1.07% | 0.19% | 15.64–16.44% |
| GPT-5.4-xhigh | 15.01% | 0.94% | 0.17% | 14.66–15.36% |

The intervals show a clear separation between IndexLM-1.0 and every comparison model. IndexLM-1.0's 95% confidence interval is 55.70–58.37%, while the strongest comparison model, Claude-Opus-4.6, has a 95% confidence interval of 28.29–29.68%. These intervals are far apart, so the difference is much larger than the run-to-run variation observed within either model.

### 7.2 Two-sample $t$-tests

To test whether IndexLM-1.0's higher retention could plausibly be due to run-to-run variation, we use two-sample $t$-tests for the two most important comparisons. The first compares IndexLM-1.0 with Mistral-3-Large, its non-fine-tuned base model. The second compares IndexLM-1.0 with Claude-Opus-4.6, the strongest general-purpose comparison model.
For a two-sample $t$-test, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$p = P(T_{df} \geq t)$$

$$df = \min(n_1 - 1, n_2 - 1) = 29$$

**IndexLM-1.0 vs. Mistral-3-Large**

Let $\mu_I$ be the true mean retention rate for IndexLM-1.0 and let $\mu_M$ be the true mean retention rate for Mistral-3-Large. The hypotheses are

$$H_0 : \mu_I - \mu_M = 0$$

$$H_a : \mu_I - \mu_M > 0$$

Using the 30 runs for each model,

$$\bar{x}_I = 0.5704, \quad s_I = 0.0358, \quad n_I = 30$$

$$\bar{x}_M = 0.1604, \quad s_M = 0.0107, \quad n_M = 30$$

Substituting into the test statistic gives

$$t = \frac{0.5704 - 0.1604}{\sqrt{\frac{0.0358^2}{30} + \frac{0.0107^2}{30}}}$$

$$t = \frac{0.4100}{\sqrt{0.0000427 + 0.0000038}}$$

$$t = \frac{0.4100}{0.00683} \approx 60.04$$

$$p = P(T_{29} \geq 60.04)$$

$$p \approx 2.8 \times 10^{-32}$$

This is far smaller than any usual significance level, such as $\alpha = 0.05$ or $\alpha = 0.01$. Therefore, we reject the null hypothesis. IndexLM-1.0 retained a significantly larger share of the human index than Mistral-3-Large. Since Mistral-3-Large is the non-fine-tuned base model for IndexLM-1.0, this supports the claim that indexing-specific fine-tuning improved topical selection.

**IndexLM-1.0 vs. Claude-Opus-4.6**

The second test compares IndexLM-1.0 with Claude-Opus-4.6. Let $\mu_C$ be the true mean retention rate for Claude-Opus-4.6. The hypotheses are

$$H_0 : \mu_I - \mu_C = 0$$

$$H_a : \mu_I - \mu_C > 0$$

Using the 30 runs for Claude-Opus-4.6,

$$\bar{x}_C = 0.2899, \quad s_C = 0.0187, \quad n_C = 30$$

Substituting into the test statistic gives

$$t = \frac{0.5704 - 0.2899}{\sqrt{\frac{0.0358^2}{30} + \frac{0.0187^2}{30}}}$$

$$t = \frac{0.2805}{\sqrt{0.0000427 + 0.0000117}}$$

$$t = \frac{0.2805}{0.00738} \approx 38.02$$

$$p = P(T_{29} \geq 38.02)$$

$$p \approx 1.3 \times 10^{-26}$$

Again, this is far smaller than $\alpha = 0.05$ or $\alpha = 0.01$, so we reject the null hypothesis. IndexLM-1.0 retained a significantly larger share of the human index than Claude-Opus-4.6, the strongest general-purpose comparison model.

### 7.3 Random-pruning baseline

A random-pruning baseline helps interpret whether the models are doing better than chance. If a model randomly selected 1,058 entries from the 7,400-entry candidate pool, the expected number of retained human-index entries would be approximately

$$1058 \cdot \frac{1055}{7400} \approx 150.84$$

As a retention rate, this is

$$\frac{150.84}{1066} \times 100 \approx 14.15\%$$

This gives a rough chance baseline of about 14.15% human-index retention.
The comparison models range from slightly above this baseline to clearly above it. GPT-5.4-xhigh retained 15.01%, and Mistral-3-Large retained 16.04%, only slightly above random pruning. Gemini-3.1-Pro retained 18.01%, somewhat above random. Claude-Opus-4.6 retained 28.99%, clearly above random. IndexLM-1.0 retained 57.04%, far above random.
This supports the first research question: some general-purpose LLMs preserve human-selected topics above chance, but the purpose-trained model preserves far more of the human index than either random pruning or the general-purpose comparison models.

Figure 2: Mean human-index retention by model at the human-scale checkpoint, scaled to the highest mean.

| Model | Mean retention | Visual comparison |
|---|---|---|
| IndexLM-1.0 | 57.04% | |
| Claude-Opus-4.6 | 28.99% | |
| Gemini-3.1-Pro | 18.01% | |
| Mistral-3-Large | 16.04% | |
| GPT-5.4-xhigh | 15.01% | |

Figure 2 gives a visual summary of the main result. IndexLM-1.0 is not just slightly ahead of the other models, as its mean retention is almost twice Claude-Opus-4.6's and more than three times the Mistral-3-Large base model's retention rate.

## 8 Interpretation

The results answer the original question in two parts. First, large language models can approximate human editorial judgment in this narrow indexing task better than chance. Claude-Opus-4.6 in particular retained almost 29% of the human top-level index at the same length constraint, which is well above the random-pruning baseline. This suggests that general-purpose models do learn some signals of topical importance from language alone.

Second, the larger result is that purpose-specific training matters. IndexLM-1.0 retained 57.04% of the human index, while its Mistral-3-Large base model retained only 16.04%. Because the two models share the same base model family, the difference is evidence that fine-tuning on real indexes changed the model's behavior toward the editorial choices made by professional indexers.

This benchmark does not show that IndexLM-1.0 can produce a complete professional index by itself. This study evaluates only top-level topic selection from a cleaned candidate pool. It does not evaluate page locator accuracy, subentry structure, cross-references, formatting, or the ability to generate candidates without human review. Those are major parts of professional indexing. A model that preserves many of the right top-level topics could still produce a weak index if it attaches the wrong pages, uses confusing subheadings, or fails to distinguish major discussion from passing mentions.

The study also depends on one book. Doyle's *The Oxford History of the French Revolution* is a useful test case because it is a dense academic history, but results might differ for textbooks, biographies, technical manuals, trade books, or books with very different naming conventions. Future research should repeat the same benchmark across multiple genres and compare agreement with more than one human indexer when possible.

## 9 Conclusion

This paper studied whether large language models can approximate one part of human editorial judgment: deciding which top-level topics deserve to remain in a back-of-the-book index. The experiment used a 7,400-entry candidate pool for William Doyle's *The Oxford History of the French Revolution* and compared five models across 150 pruning runs. Each model was required to prune the list to the same human-scale size of 1,058 entries, making the comparison about selection quality rather than index length.

The main result is that IndexLM-1.0, a model fine-tuned specifically for indexing, retained far more of the published human index than any general-purpose comparison model. Its mean retention was 57.04%, compared with 28.99% for Claude-Opus-4.6, 18.01% for Gemini-3.1-Pro, 16.04% for Mistral-3-Large, and 15.01% for GPT-5.4-xhigh. The two-sample $t$-tests showed that the gaps between IndexLM-1.0 and both Mistral-3-Large and Claude-Opus-4.6 were statistically significant.

In the larger context of prior evaluations, these findings suggest that one-shot chatbot indexing is not the only meaningful way to evaluate AI indexing systems. The ASI evaluations show that direct chatbot generation remains weak, but this experiment shows that a narrower, staged, and purpose-trained system can perform substantially better on the specific task of topical selection. The practical implication is that future AI indexing systems should probably be evaluated component by component: candidate generation, topic selection, locator assignment, subentry organization, cross-reference creation, and final human review.

Further research should extend this benchmark to more books, more human reference indexes, and later stages of the indexing process. The next most important question is whether strong topical selection can be combined with accurate locators and useful subentries. Until that is shown,

the results support a limited conclusion: LLMs can capture part of the human editorial signal in indexing, and domain-specific fine-tuning appears to improve that ability substantially.

## References


[1] Bartmess, Elizabeth, and Michele R. Combs. *Can the Current Generation of LLMs Produce an Adequate Index?* Prepared by the AI Committee of the American Society for Indexing, November 2025. https://asindexing.org/ai-news/supplement-to-white-paper-ai-and-indexing/

[2] Bartmess, Elizabeth, and Michele R. Combs. *AI and Book Indexing: Trajectory Data*. Prepared by the AI Committee of the American Society for Indexing, March 2026.

[3] IndexerLabs. "Introducing IndexLM-1.0: Democratizing Access to Affordable, High Quality Indexes." *IndexerLabs Blog*, March 1, 2026. https://indexerlabs.com/blog/introducing-indexlm-1

[4] IndexerLabs. "What We Learned Indexing the Same Book 120 Times." *IndexerLabs Blog*, April 12, 2026. https://indexerlabs.com/blog/what-we-learned-indexing-the-same-book-120-times

[5] OpenAI. "GPT-5.4 Model." *OpenAI API Documentation*, 2026. https://developers.openai.com/api/docs/models/gpt-5.4

[6] OpenAI. "GPT-5.4 Thinking System Card." *Deployment Safety Hub*, 2026. https://deploymentsafety.openai.com/gpt-5-4-thinking

[7] Anthropic. "Introducing Claude Opus 4.6." *Anthropic News*, February 5, 2026. https://www.anthropic.com/news/claude-opus-4-6

[8] Anthropic. *Claude Opus 4.6 System Card*. Anthropic, February 2026. https://www-cdn.anthropic.com/c788cbc0a3da9135112f97cdf6dcd06f2c16cee2.pdf

[9] Google. "Gemini 3.1 Pro Preview." *Google AI for Developers*, 2026. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview

[10] Mistral AI. "Mistral Large 3." *Mistral AI Model Card*, December 2025. https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12

[11] Ullstrom, Stephen. "Embracing the Subjective Side of Indexing." *Stephen Ullstrom Indexing & Writing*, February 13, 2026. https://stephenullstrom.com/embracing-the-subjective-side-of-indexing/

[12] Google DeepMind. *Gemini 3 Flash Model Card*. December 2025. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf

[13] Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. "ReAct: Synergizing Reasoning and Acting in Language Models." *International Conference on Learning Representations*, 2023. https://arxiv.org/abs/2210.03629

[14] Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. "Self-Refine: Iterative Refinement with Self-Feedback." *Advances in Neural Information Processing Systems*, 2023. https://arxiv.org/abs/2303.17651

[15] Shinn, Noah, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. "Reflexion: Language Agents with Verbal Reinforcement Learning." *Advances in Neural Information Processing Systems*, 2023. https://arxiv.org/abs/2303.11366

[16] Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. "Toolformer: Language Models Can Teach Themselves to Use Tools." *Advances in Neural Information Processing Systems*, 2023. https://arxiv.org/abs/2302.04761

[17] Yang, John, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. "SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering." *Advances in Neural Information Processing Systems*, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html

[18] Pan, Linyue, Lexiao Zou, Shuo Guo, Jingchen Ni, and Hai-Tao Zheng. "Natural-Language Agent Harnesses." *arXiv preprint arXiv:2603.25723*, 2026. https://arxiv.org/abs/2603.25723

[19] Chen, Shiqi, Jingze Gai, Ruochen Zhou, Jinghan Zhang, Tongyao Zhu, Junlong Li, Kangrui Wang, Zihan Wang, Zhengyu Chen, Klara Kaleb, Ning Miao, Siyang Gao, Cong Lu, Manling Li, Junxian He, and Yee Whye Teh. "SkillCraft: Can LLM Agents Learn to Use Tools Skillfully?" *arXiv preprint arXiv:2603.00718*, 2026. https://arxiv.org/abs/2603.00718

[20] Abuzakuk, Sami, Anne-Marie Kermarrec, Rishi Sharma, Rasmus Moorits Veski, and Martijn de Vos. "Optimizing Agentic Workflows using Meta-tools." *arXiv preprint arXiv:2601.22037*, 2026. https://arxiv.org/abs/2601.22037

[21] Hu, Jun. "Agentic Tool Use in Large Language Models." *arXiv preprint arXiv:2604.00835*, 2026. https://arxiv.org/abs/2604.00835