Diagnosing Hierarchical Retrieval Failure in Long-Document RAG: A LongBook Verifier Ablation Study

Author: Panagiotis Gkilis

Abstract

Experiment C is a follow-up ablation study for the LongBook Verifier benchmark. Its purpose is to diagnose why the original hierarchical book RAG method underperformed chapter-summary chain in Experiment B. The experiment uses the existing 5-book Experiment B corpus and the existing 80 gold questions. It does not modify the manuscript, the gold questions, or the original Experiment A/B scripts.

The original measured result was that chapter-summary chain achieved higher context recall than hierarchical book RAG. Experiment C separates the hierarchical retrieval pipeline into controlled variants: the current hierarchical method, the same method without neighbor expansion, oracle-chapter routing without neighbors, oracle-chapter routing with neighbors, and the chapter-summary chain baseline. The results support the error-compounding hypothesis but show that it was incomplete: first-stage chapter-selection limits matter, and neighbor-expansion dilution was also a major failure mode.

Background

LongBook Verifier evaluates whether retrieval methods and model outputs are grounded in long narrative documents. Experiment B used a 5-book long-form narrative corpus containing 240,767 words and 80 gold questions. The benchmark compared five retrieval methods:

method	context precision	context recall
naive_first_context	0.1475	0.1458
naive_last_context	0.4150	0.3302
flat_chunk_rag	0.3375	0.4302
chapter_summary_chain	0.4000	0.4771
hierarchical_book_rag	0.3475	0.3365

The original result was measured rather than assumed: chapter-summary chain was the best original retrieval method by context recall, and hierarchical book RAG came third by context recall, not second. It ranked below chapter-summary chain and flat chunk RAG, while only slightly above naive last-context retrieval.

This was counterintuitive because hierarchical retrieval is often expected to help long-document search by first selecting a broader section and then retrieving local evidence inside that section. Experiment C tests where the loss occurred in this specific corpus and protocol.

Research Question

Why did hierarchical book RAG underperform chapter-summary chain in this corpus/protocol?

The starting hypothesis was error compounding: if the first retrieval stage selects a wrong or weak chapter, later chunk retrieval is already constrained to the wrong context. Experiment C tests that hypothesis while also isolating the effect of neighbor expansion.

Methods

Experiment C uses the existing Experiment B 5-book corpus and the existing 80 gold questions. It creates 400 scored rows: 80 questions multiplied by five ablation methods. The evaluation uses context recall and context precision based

on evidence-term overlap, matching the LongBook Verifier scoring style.

The five methods are:

```
| method | description |
|---|---|
| `hier_current` | The original hierarchical book RAG behavior: chapter selection, chunk retrieval, and neighbor expan
| `hier_no_neighbors` | The same hierarchical chapter selection and in-chapter chunk retrieval, but with neighbor expa
| `chapter_summary_chain` | The existing chapter-summary chain baseline from Experiment B. |
| `hier_oracle_chapter` | A diagnostic oracle variant that uses the gold expected chapter and retrieves chunks only in
| `hier_oracle_chapter_neighbors` | A diagnostic oracle variant that uses the gold expected chapter, retrieves chunks
```

The oracle variants are diagnostic only. They use gold expected-chapter labels and therefore are not deployable production retrieval methods. Their purpose is to estimate how much performance is lost before or after the chapter-selection stage.

Failure types were assigned as follows:

```
| failure type | definition |
|---|---|
| `wrong_chapter` | The expected chapter was not selected. |
| `right_chapter_wrong_chunk` | The expected chapter was selected but evidence terms were not retrieved. |
| `neighbor_dilution` | Neighbor expansion lowered precision or recall compared with the corresponding no-neighbor ver
| `ok` | Expected evidence was retrieved. |
```

Results

The ablation produced 400 result rows: 80 questions times five methods. Oracle chapter mapping succeeded for all 80 questions.

```
| method | context recall | context precision | hit@1 | hit@3 | hit@5 |
|---|---:|---:|---:|---:|---:|
| `hier_current` | 0.3365 | 0.3475 | 0.2000 | 0.3875 | 0.4375 |
| `hier_no_neighbors` | 0.4771 | 0.4000 | 0.2000 | 0.3875 | 0.4375 |
| `chapter_summary_chain` | 0.4771 | 0.4000 | 0.2000 | 0.3875 | 0.4375 |
| `hier_oracle_chapter` | 0.7844 | 0.6796 | 1.0000 | 1.0000 | 1.0000 |
| `hier_oracle_chapter_neighbors` | 0.7688 | 0.6925 | 1.0000 | 1.0000 | 1.0000 |
```

For the original hierarchical method, `hier_current`, the failure counts were:

```
| failure type | count |
|---|---:|
| `neighbor_dilution` | 32 |
| `wrong_chapter` | 27 |
| `right_chapter_wrong_chunk` | 7 |
| `ok` | 14 |
```

The conditional recall values also show that chapter selection mattered. For `hier_current`, conditional recall was 0.4000 when the expected chapter appeared in the top 5, and 0.2870 when it did not. For `hier_no_neighbors` and `chapter_summary_chain`, conditional recall was 0.6190 when the expected chapter appeared in the top 5 and 0.3667 when it did not.

Interpretation

The original error-compounding hypothesis was supported but incomplete.

Oracle chapter routing raised recall from 0.3365 to 0.7844. This supports the idea that first-stage chapter selection can damage downstream retrieval. When the correct chapter is forced, the in-chapter retrieval stage has much better access to evidence-bearing chunks.

However, removing neighbor expansion raised hierarchical recall from 0.3365 to 0.4771, matching chapter-summary chain. That means the original hierarchical pipeline was not failing only because of first-stage chapter selection. Neighbor expansion also diluted the context in many cases by crowding out or replacing better evidence chunks.

The oracle-neighbor variant reached recall 0.7688, slightly below the no-neighbor oracle recall of 0.7844. This suggests that neighbor expansion can still add context, but it is not automatically beneficial. In this protocol, neighbor expansion should be treated as a tunable retrieval stage, not a default improvement.

Safe Conclusion

For this 5-book corpus and 80-question protocol, chapter-summary chain was the strongest original retrieval method by context recall. Experiment C supports the diagnosis that the original hierarchical book RAG underperformed due to both first-stage chapter-selection limits and neighbor-expansion dilution. The study does not establish a universal rule against hierarchical retrieval; instead, it shows that multi-stage retrieval pipelines require stage-level ablation before their failures can be interpreted.

Limitations

This study uses a single private narrative corpus and 80 gold questions. The gold questions were written from the corpus context and are not an independently authored public benchmark. The evidence-term scoring is lightweight and does not replace full semantic judgment. No confidence intervals are reported because the package did not compute them. The oracle-chapter variants are diagnostic and not production-realistic because they use gold expected-chapter labels. The results should not be interpreted as a universal claim about hierarchical RAG.

Reproducibility Notes

The public package excludes full manuscript text and raw private corpus files. It includes the ablation script, summary tables, plots, reports, metadata, and packaging script. The ablation was run with:

```
python src\run_ablation.py
```

The Zenodo package can be rebuilt with:

```
python src\package_experiment_c_zenodo.py
```