# Patient2Paper at #SMM4H-HeaRD 2026: Retrieval-Augmented Few-Shot Generation for Clinical Note Synthesis

**Timotei Andrei**[1] **Ioan-Tudor-Alexandru Anghel**[1]
**Bogdan Marian Comârdici**[1] **Carina Săicu**[1]

[1]Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania
timotei.andrei@s.unibuc.ro, ioan-tudor.anghel@s.unibuc.ro
marian-bogdan.comardici@s.unibuc.ro, carina.saicu@s.unibuc.ro

Authors are listed alphabetically by last name.

## Abstract

We present a retrieval-augmented few-shot system for the MedSynth Dial2Note shared task at SMM4H-HEARD 2026, placing 3rd on the official leaderboard (0.51 avg). Across 28 configurations, we find that retrieval design (hybrid BM25 + medical-domain dense fused via RRF) and prompt presentation format (few-shot examples as conversation turns) are the primary quality drivers, while model scale has surprisingly limited impact: Llama 3.2:3B, Llama 3.1:8B and GPT-4o mini remain within a narrow band on our locally computed scores. Our final submission used GPT-4o mini with $k$=3 few-shot examples retrieved by RRF over BioLORD-2023 embeddings. We report a full ablation, including negative results, to show where the gains come from and where further engineering stops paying off.

## 1 Introduction

Clinical documentation is a major burden for physicians (Hripcsak and Albers, 2011), motivating automated solutions that can reduce time spent on note-taking. The MedSynth Dial2Note task (Rezaie Mianroodi et al., 2025) at the SMM4H-HEARD 2026 Workshop (Lopez-Garcia et al., 2026) targets generation of structured SOAP-format clinical notes (Podder et al., 2022) from synthetic doctor-patient dialogues. The shared-task dataset contains more than 10,000 synthetic dialogue-note pairs spanning over 2,000 ICD-10 codes; synthetic generation avoids privacy concerns while preserving stylistic consistency. Systems are assessed using automatic metrics (BLEU, ROUGE-1/2/L, METEOR, and MedCon), though the task description also outlined LLM-as-a-judge and human expert review stages for top teams; in practice, only automatic metrics were used for the final ranking.

We describe a retrieval-augmented few-shot (RAG) system that placed 3rd on the official leaderboard. Our contribution is less the submission itself and more the analysis behind it: a systematic investigation of 28 configurations covering retrieval strategy, few-shot presentation, post-processing, and generator choice. The central empirical finding is that retrieval design and prompt format dominate, whereas model scale, from a 3B local model up to GPT-4o mini, offers marginal gains compared to the substantial performance leap provided by a superior retrieval strategy. Once retrieval and prompt format are optimised, the LLM behaves more as a format-aware assembler of retrieved structure than as a generator producing content from parametric knowledge, and further scaling yields diminishing returns.

## 2 Related Work

Dialogue-to-note generation has been studied on datasets such as ACI-Bench (Yim et al., 2023) and PriMock57 (Korfiatis et al., 2022), and in the MEDIQA-Chat 2023 shared task (Ben Abacha et al., 2023), where WangLab's winning entry (Giorgi et al., 2023) showed that few-shot prompting with strong retrieval (using GPT-4 as the generator) could rival fine-tuning. The MedSynth pipeline (Rezaie Mianroodi et al., 2025) uses multi-agent LLM setups to produce the synthetic data used here. Retrieval-augmented generation (Lewis et al., 2020), Reciprocal Rank Fusion (Cormack et al., 2009), and biomedical encoders such as BioLORD (Remy et al., 2023) are well-established components that we combine and ablate in a controlled fashion on this task.

## 3 System Description

### 3.1 Overview

Our system is a two-phase retrieval-augmented pipeline (Figure 1).[1] Given a test dialogue, we first retrieve $k$ similar dialogue-note pairs from the

[1]Code and prompts are available at https://github.com/bogdancomardici/SMM4H-26-SOAP-Generation.



*Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks*, pages 61–66
July 3, 2026 ©2026 Association for Computational Linguistics

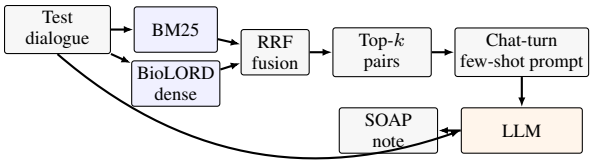


Figure 1: System architecture. RRF merges BM25 and BioLORD-2023 dense rankings; the top-$k$ pairs and the test dialogue are assembled into a multi-turn chat prompt for the LLM.

training split using a hybrid retrieval strategy. The retrieved pairs are formatted into a few-shot prompt and passed to a large language model, which produces the SOAP-format note in a single pass. Retrieval and generation are decoupled components that we vary independently; this modularity is what made the ablation below tractable.

### 3.2 Retrieval

We evaluated four families of retrievers:

- **BM25**: Okapi BM25 (Robertson and Zaragoza, 2009), a sparse lexical baseline scoring candidate dialogues on token overlap and term frequency.
- **Dense retrieval**: cosine similarity over dialogue embeddings, using both a general-purpose encoder (BAAI/bge-base-en-v1.5 (Zhang et al., 2023)) and a medical-domain encoder (BioLORD-2023 (Remy et al., 2023)).
- **Reciprocal Rank Fusion (RRF)** (Cormack et al., 2009): merges BM25 and dense rankings. The fused score for a document $d$ is

$$RRF(d) = \sum_{r \in R} \frac{1}{k_{RRF} + rank_r(d)}, \quad (1)$$

  with $R$ the set of rankers and $k_{RRF}$=60 the standard smoothing constant (distinct from the few-shot $k$ below).
- **ICD-10 hybrid**: RRF with an additive bonus for candidates that share the test dialogue's ICD-10 code. This turned out to be redundant in combination with BioLORD, whose embeddings already encode diagnostic similarity (Section 5).

Retrieval returns the top-$k$ training dialogue-note pairs with $k \in \{1, 3, 5\}$. We found that RRF with the medical-domain encoder gave the best balance of surface and clinical similarity, and that going from $k$=1 to $k$=3 produced large gains, with diminishing returns past $k$=3.

Dialogues exceeding BioLORD-2023's 512-token limit are silently truncated; this affects 99.5% of the evaluation set (mean length: 734 words). The truncated tail is almost entirely closing pleasantries, so clinically relevant content remains within the embedding window.

### 3.3 Prompt Construction

We compared two few-shot presentations:

- **Conversation turns (best)**: each retrieved example is a separate user/assistant turn in the chat history, and the test dialogue is the final user message.
- **In-prompt (baseline)**: examples are concatenated as flat text inside a single system/user message.

The turns format yielded a 12% relative improvement in average score over in-prompt, consistent with the hypothesis that the LLM's multi-turn attention preserves structural boundaries between examples more faithfully than flat concatenation.

### 3.4 Generation

We used three generators: Llama 3.2:3B and Llama 3.1:8B (both via Ollama, local inference) and GPT-4o mini (API). Temperature was fixed at 0.1 and the context window at up to 16,384 tokens. Once retrieval and prompt format were optimised, the choice of generator changed the locally computed average score by less than 0.03. The prompt was designed for Llama 3.2:3B and applied unchanged to all models; instruction-following differences between Llama and GPT families mean scale is not the sole variable, and model-specific tuning may yield larger gains. Our final leaderboard submission used GPT-4o mini with $k$=3; the remainder of the ablation was run with Llama 3.2:3B for cost reasons.

## 4 Experimental Setup

### 4.1 Dataset

The shared-task dataset (Rezaie Mianroodi et al., 2025), released by the organisers, contains 10,035 synthetic dialogue-note pairs spanning 2,001 ICD-10 codes with 5 pairs per code. We used the official 85/15 train/dev split (8,530 / 1,505 pairs) and the hidden test set for final evaluation. Dialogues average 932 tokens (∼55 sentences), notes average 621 tokens (∼23 sentences). All data is synthetic, produced by the organisers via a multi-agent GPT-4o pipeline (Rezaie Mianroodi et al., 2025).

### 4.2 Evaluation Metrics

The primary leaderboard ranking is the average of BLEU (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and MedCon. MedCon measures clinical content overlap: QuickUMLS (Soldaini and Goharian, 2016) extracts UMLS Concept Unique Identifiers (CUIs) from both the candidate and reference notes, and MedCon is the F1 over matched CUIs. Locally we scored outputs with SacreBLEU (Post, 2018), `rouge-score`, NLTK METEOR, and QuickUMLS (Docker backend); leaderboard scores are the ones returned by the shared-task evaluator on the hidden test set.

### 4.3 Experiment Design

We ran 28 configurations in three rounds:

- **Round 1**: retrieval baselines (BM25, BGE dense, RRF) $\times$ prompt format (in-prompt vs turns), with Llama 3.2:3B.
- **Round 2**: the medical-domain encoder (BioLORD-2023) inside RRF, with the best prompt format fixed.
- **Round 3**: post-processing (skeleton, vocabulary normalisation, length calibration, MedCon priming) and combinations thereof on top of Round 2's best retrieval.

We also compared Llama 3.1:8B and GPT-4o mini on the best retrieval configuration, and ran metric-targeted post-processing (ROUGE/BLEU boosters) as a separate sanity check. Negative results are reported alongside positive ones to keep the ablation honest.

## 5 Results and Analysis

### 5.1 Main Results

Table 1 summarises the main trial-and-error process that guided system design. The strongest local configuration was RRF(BioLORD) with turns-format prompting and $k$=5, while the most important qualitative lesson is that retrieval and prompt formatting matter more than downstream post-processing.

To satisfy the shared-task requirement of reporting validation and test performance, Table 2 gives the best validation result and the best test result. The best validation score came from Llama 3.1:8B, which reached 0.5502 on the labeled validation set. On the final test side, GPT-4o mini achieved 0.5078 on our local rerun of the organiser evaluator and 0.51 in the official submission (ID 675806).

### 5.2 Ablation and Key Findings

Figure 2 summarises the ablation as deltas against the best local configuration (05e, 0.546 avg).

**Retrieval is the dominant factor.** Zero-shot generation scores only 0.075; few-shot retrieval is what makes SOAP-format output possible at all. Among retrieval strategies, RRF over dense-only BGE yields a 14% relative gain (0.459 vs 0.403), and substituting BioLORD for BGE inside RRF pushes the average from 0.515 to 0.546. ICD-10 code boosting is redundant given BioLORD: 05a and 05b are indistinguishable.

**Prompt format is the second-largest lever.** Presenting few-shot examples as separate conversation turns rather than flat in-prompt text improves the score by 12% relative (0.515 vs 0.459). We attribute this to the LLM's multi-turn attention preserving structural boundaries more faithfully. The jump from $k$=1 to $k$=3 is dramatic (0.343 $\rightarrow$ 0.537); $k$=3 to $k$=5 adds only 1.6% on the overall average. The exception is MedCon, which falls from 0.8011 at $k$=3 to 0.7288 at $k$=5: the additional neighbours share the presenting complaint but diverge in specific clinical details, and the LLM may incorporate their patient-specific concepts into the output, reducing concept-level precision.

**Post-processing and alternative pipelines did not help.** Length calibration, skeleton prompting, vocabulary normalisation and their combinations all matched or slightly underperformed the unmodified pipeline. MedCon priming (pre-injecting UMLS concepts) boosted concept recall but hurt surface metrics, with a net-neutral to slightly negative effect on the average. Section-by-section generation (0.392) and extract-then-generate (0.135) were strictly worse, losing the structural patterns the few-shot examples carry. Multi-pass refinement added latency without improving scores.

**Generator scale is a weak influence.** Llama 3.2:3B, Llama 3.1:8B and GPT-4o mini land within 0.03 on locally scored configurations; with a strong RAG pipeline in place, scaling the generator is dominated by improving retrieval.

### 5.3 Error Analysis

Two failure modes recur across configurations. **Omissions**: clinically relevant details mentioned in the dialogue (family history, negative findings, specific test values) are occasionally dropped from the note, usually in favour of the structural patterns seen in the retrieved neighbours. **Hallucinations**:

| Exp | Method | BLEU | R-1 | R-2 | R-L | METEOR | MedCon | Avg |
|---|---|---|---|---|---|---|---|---|
| 05e | RRF(BioLORD) k=5, turns | 0.4149 | 0.6556 | 0.4071 | 0.5095 | 0.5597 | 0.7288 | 0.546 |
| 05a | RRF(BioLORD) k=3, turns | 0.3884 | 0.6315 | 0.3847 | 0.4653 | 0.5514 | 0.8011 | 0.537 |
| 05b | RRF(BioLORD)+ICD10 k=3 | 0.3884 | 0.6315 | 0.3847 | 0.4653 | 0.5514 | 0.8011 | 0.537 |
| 06b | k=5, +skeleton | 0.4136 | 0.6481 | 0.3843 | 0.4341 | 0.5783 | 0.7029 | 0.527 |
| 01g | RRF(BGE) k=3, turns | 0.3905 | 0.6410 | 0.3825 | 0.4322 | 0.5581 | 0.6840 | 0.515 |
| 01b | RRF(BGE) k=3, in-prompt | 0.3275 | 0.5755 | 0.3500 | 0.4288 | 0.4831 | 0.5891 | 0.459 |
| 01a | Dense(BGE) k=3 | 0.2880 | 0.4913 | 0.2986 | 0.3793 | 0.4407 | 0.5218 | 0.403 |
| 04a | Section-by-section | 0.2438 | 0.4983 | 0.2978 | 0.3489 | 0.4202 | 0.5426 | 0.392 |
| 01f | RRF(BGE) k=1 | 0.2152 | 0.4530 | 0.2550 | 0.3129 | 0.3757 | 0.4456 | 0.343 |
| 04b | Extract-then-generate | 0.0026 | 0.1951 | 0.1265 | 0.1496 | 0.1281 | 0.2070 | 0.135 |
| 00 | Zero-shot | 0.0003 | 0.1340 | 0.0680 | 0.0842 | 0.0861 | 0.0750 | 0.075 |

Table 1: Selected configurations scored with our local evaluator stack (generator: Llama 3.2:3B). Avg is a six-metric local average including MedCon, so it is not directly comparable to the official five-metric shared-task score.

| Split | Model | BLEU | R-1 | R-2 | R-L | METEOR | Avg |
|---|---|---|---|---|---|---|---|
| Validation best | Llama 3.1:8B | 0.4651 | 0.7040 | 0.4462 | 0.5178 | 0.6181 | 0.5502 |
| Test (local eval) | GPT-4o mini | 0.4138 | 0.6649 | 0.3948 | 0.4709 | 0.5946 | 0.5078 |
| Test (official sub.) | GPT-4o mini | 0.41 | 0.66 | 0.39 | 0.47 | 0.59 | 0.51 |

Table 2: Best validation and test results under the official five-metric scorer. The official test row is reported at the precision provided by the organisers.

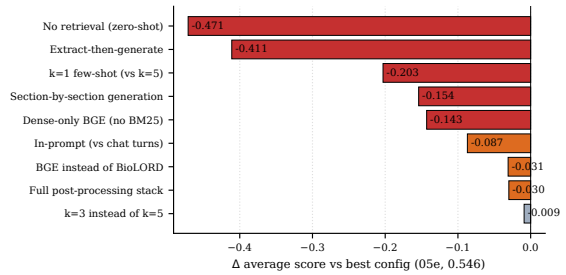


Figure 2: Ablation deltas against the best configuration (05e, 0.546). Each bar removes or downgrades one pipeline component.

the generator sometimes invents precise-looking numeric details, such as medication dosages, lab values, blood-pressure readings, that were not stated in the dialogue. Both error modes decrease as retrieval quality improves (RRF + BioLORD vs dense BGE) and as $k$ grows from 1 to 3, but they do not disappear even at our best configuration. Model scale had noticeably less effect on these errors than retrieval quality or prompt format did.

## 6 Discussion

Two things stand out from the ablation. First, *retrieval quality is the binding constraint*: what examples are retrieved and how they are presented determines most of the gap between a 0.34 system and a 0.55 system, while swapping the generator over more than an order of magnitude in parameter count moves the score by less than any single retrieval decision. In a deployment setting where compute is limited, this implies investing retrieval engineering effort before reaching for a larger model. Second, *small LLMs in a RAG setting behave more like format-aware assemblers than generators*: the few-shot examples carry the structural template, the dialogue carries the content, and the model's job is mostly to reorganise the latter into the former.

Once the core pipeline (RRF over BioLORD, conversation-turn prompting, $k \geq 3$) is in place, further modifications keep the averaged score inside a narrow 0.48–0.55 band. We read this as a soft ceiling imposed by the RAG on synthetic data combination: breaking past it likely requires orthogonal moves such as fine-tuning on the target style, retrieval from curated external knowledge bases, or evaluation on real clinical data.

**Limitations.** We did not fine-tune, so we cannot say where the parametric knowledge bottleneck sits. All numbers are on synthetic data, whose stylistic uniformity likely makes the task easier than real clinical transcription.

## 7 Conclusion

We presented a retrieval augmented few-shot system for the MedSynth Dial2Note shared task, placing 3rd on the official leaderboard (0.51 avg). A controlled sweep of 28 configurations isolates retrieval design ( hybrid BM25 + BioLORD fused with RRF ) and few-shot presentation as conversation turns as the primary drivers of quality, with generator scale from 3B up to GPT-4o mini contributing only marginally once these are in place. Natural next steps include fine-tuning a small generator on the best retrieval configuration, retrieval from external medical knowledge bases, best-of-$N$ selection, and evaluation on real, non-synthetic, clinical dialogues.

## References


Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An overview of the MEDIQA-Chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513.

Gordon V Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

John Giorgi, Augustin Toma, Ronald Xie, Sondra S Chen, Kevin R An, Grace X Zheng, and Bo Wang. 2023. WangLab at MEDIQA-Chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334.

George Hripcsak and David J Albers. 2011. Secondary use of electronic health records. *Journal of biomedical informatics*, 44:S1–S2.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2022. SOAP notes. In *StatPearls*. StatPearls Publishing.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

François Remy, Kris Demuynck, and Thomas Demeester. 2023. BioLORD-2023: Semantic textual representations fusing LLM and clinical knowledge graph insights. *arXiv preprint arXiv:2311.16075*.

Ahmad Rezaie Mianroodi, Hossein Amirkhani, and Mohammad Taher Pilevar. 2025. MedSynth: Realistic, synthetic medical dialogue–note pairs. *arXiv preprint arXiv:2508.01401*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-Bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. volume 10, page 586.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.


## A Qualitative Example

Table 3 shows an illustrative output from the best ablation configuration (RRF-BioLORD, $k$=3, conversation-turn prompting, Llama 3.2:3B) on a validation-set instance (ICD-10: early-stage age-related macular degeneration). We use the Llama 3.2:3B output here as the ablation was run with this model; the final leaderboard submission used GPT-4o mini on the hidden test set.

| **Dialogue excerpt (first 5 turns)** |
|---|
| `[doctor]`: Hi Ms. Greene, how are you today?<br>`[patient]`: Hi doctor, I'm doing okay, thank you.<br>`[doctor]`: Great to hear. So, I understand you're experiencing some mild visual disturbances. Can you tell me more about that?<br>`[patient]`: Yes, I've noticed my vision has been changing over the past 3 months. It's mostly blurred vision, difficulty reading, and some mild distortion in my central vision.<br>`[doctor]`: Are these symptoms constant, or do they come and go?<br>*(dialogue continues for 36 turns / 624 words total)* |
| **Reference note (key fields)** |
| *CC:* Mild visual disturbances<br>*HPI:* Ms. Greene, 68F, 3-month bilateral blurred vision, reading difficulty, central distortion; worse in bright light<br>*Objective:* BP 135/85; VA 20/40 OU; fundus: bilateral drusen in macula; OCT: confirmed drusen, no advanced AMD<br>*Assessment:* Bilateral degenerative drusen; early-stage AMD, no wet AMD<br>*Plan:* AREDS2 supplements (2 caps/day); 6-month follow-up; referral to retina specialist (Dr. Ling); smoking cessation; UV protection |
| **System output (RRF-BioLORD $k$=3, Llama 3.2:3B)** |
| *CC:* Mild visual disturbances<br>*HPI:* Female patient, 3-month bilateral blurred vision / reading difficulty *(age: 62 vs. 68 — minor demographic hallucination)*<br>*Objective:* BP 135/85 ; VA 20/40 OU ; drusen in macula ; OCT pending<br>*Assessment:* Early-stage AMD<br>*Plan:* AREDS2 × 2 caps/day ; 6-month follow-up ; referral Dr. Sarah Ling ; smoking cessation |

Table 3: Validation-set example (ICD-10: early-stage AMD). The system correctly reproduces the diagnosis, treatment plan, and referral. The characteristic error is a minor demographic hallucination (age 62 vs. reference 68), consistent with the retrieval mechanism: the top-$k$ neighbours share the presenting complaint but differ in patient demographics, and the model occasionally blends them.

## B Post-Processing Configurations (Negative Results)

Table 4 reports the negative Round 3 post-processing results. These runs confirm that retrieval and prompt format, rather than post-processing, are the binding constraints.

| ID | Post-processing applied | Avg |
|---|---|---|
| 05a | None | 0.537 |
| 06a | Length calibration | 0.534 |
| 06b | Skeleton prompting | 0.527 |
| 06c | Vocabulary normalisation | 0.531 |
| 06d | MedCon priming | 0.529 |
| 06e | Skeleton + length | 0.524 |
| 06f | Skeleton + vocab. norm. | 0.521 |
| 06g | MedCon + length | 0.526 |
| 06h | Full stack | 0.519 |

Table 4: Round 3 post-processing ablation on RRF-BioLORD $k$=3, turns format, Llama 3.2:3B. None improves over the unmodified baseline.