Reflecting What Matters: Same-Surgery References for Postoperative Clinical Note Generation

Seungmi Yu[1,2] **Somin Song**[1,2] **Hyeryun Park**[5] **Jinwook Choi**[2,3,4]

[1]Interdisciplinary Program for Bioengineering, Graduate School, Seoul National University
[2]Integrated Major in Innovative Medical Science, Graduate School, Seoul National University
[3]Department of Biomedical Engineering, College of Medicine, Seoul National University
[4]Medical Research Center, Institute of Medical and Biological Engineering, Seoul National University
[5]Healthcare AI Research Institute, Seoul National University Hospital, Seoul, Republic of Korea

ysm0909@snu.ac.kr

Abstract

Generating the first postoperative outpatient note requires selecting which information from heterogeneous clinical records should be reflected in the final note, rather than merely summarizing all inputs. We evaluate this task on an original 170-case set of de-identified clinical cases collected at Seoul National University Hospital (SNUH), involving two surgical divisions and five physicians. We then extend the evaluation to 1,329 cases across 10 surgery groups within the same institution to examine whether the main trends persist at a larger scale. We introduce physician- and surgery-specific same-surgery references: structured reflection maps that capture how information is selected and rewritten in prior same-surgery cases. Experiments show that these references become most effective when carried into the merge stage, suggesting that the main bottleneck is deciding what to reflect across records. We further show that cross-patient references can induce unsupported transfer from prior cases. To analyze this trade-off, we use a human-validated two-source FactScore that checks each generated fact against both the source records and the physician-written note. A lightweight grounding constraint reduces hallucination while maintaining or improving alignment with physician-written notes across divisions and LLMs.

1 Introduction

The first postoperative outpatient note condenses a patient's diagnosis, operation, pathology, inpatient course, and follow-up plan into a single clinical document. Unlike single-document summarization, this task integrates the six source records for the current patient, written at different times and for different purposes, making cross-record information selection and organization essential.

Although a decompose-and-merge pipeline can reduce input length by summarizing each record before merging (Gidiotis and Tsoumakas, 2020), our experiments show that a naive decompose-and-merge pipeline performs worse than a single-call baseline. Intermediate summaries can discard details that later become important, such as dates, numeric findings, operative extent, pathology results, or follow-up plans. This suggests that the bottleneck is not simply processing long inputs, but identifying which information should be preserved from each source record.

We address this bottleneck using same-surgery references constructed from prior cases. For each physician–surgery pair, we encode recurring documentation patterns as structured reflection maps rather than raw few-shot examples. These references are not intended to supply additional patient information; instead, they guide which types of information to preserve and how that information is typically rewritten.

Cross-patient references, however, can introduce a fidelity risk. Information that frequently appears in prior notes, such as routine follow-up plans, may be incorrectly transferred into the current patient's note even when it is not supported by the source records. We therefore use a human-validated two-source FactScore that checks each atomic fact against both the source records and the physician-written note.

To reduce this risk, we apply a lightweight inference-time grounding constraint. The con-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 27314–27326
October 24-29, 2026 ©2026 Association for Computational Linguistics

straint allows the reference to guide structure, item selection, and writing flow, but requires clinical content, numbers, dates, and plans to be supported by the source records. Across two surgical divisions and three LLMs, this grounded variant reduces hallucination while maintaining or improving alignment with physician-written notes.

This paper makes two contributions. First, we introduce physician- and surgery-specific same-surgery references that capture how information from heterogeneous source records is typically reflected in postoperative outpatient notes at SNUH and show that these references are most effective when carried into the merge stage. Second, we analyze the fidelity cost of cross-patient documentation patterns using a human-validated two-source FactScore and show that a lightweight grounding constraint reduces the risk of incorrect cross-patient transfer while maintaining or improving alignment with physician-written notes.

2 Related Work

A prominent line of work studies clinical note generation from patient–clinician conversations. ACI-BENCH benchmarks outpatient note generation from visit dialogues, while MedicalSum incorporates medical knowledge into abstractive summarization of patient–doctor conversations (Yim et al., 2023; Michalopoulos et al., 2022). Our task instead generates the first postoperative outpatient note from six source records written at different stages of care, making cross-record information selection and organization central to the task.

Long-document summarization has explored divide-and-conquer processing, in which partial summaries are later recombined (Gidiotis and Tsoumakas, 2020). Our setting differs because the inputs are distinct clinical record types rather than segments of one document. We therefore study where cross-record guidance should be introduced into a decompose-and-merge pipeline. In our experiments, same-surgery references are most effective when carried into the merge stage.

Because surface overlap does not fully measure faithfulness, prior work has emphasized factuality-oriented evaluation (Maynez et al., 2020). FActScore evaluates long-form generation at the atomic-fact level (Min et al., 2023). We adapt this perspective by checking each generated fact against both the source records and the physician-written note, separating input-supported facts from target-aligned facts and revealing the risk of cross-patient transfer from references.

3 Task and Data

3.1 Task Definition

We study the task of generating the first postoperative outpatient note for a surgical case from six source records written at different stages of care: the initial outpatient note, preoperative outpatient note, operation note, discharge summary, pathology report, and diagnostic test report. Let $D = \{d_1, \ldots, d_6\}$ denote the source records, and let y denote the physician-written postoperative outpatient note. The goal is to generate a draft $\hat{y}$ that is grounded in D and semantically aligned with y.

This task differs from single-document summarization because it requires selecting and organizing information across heterogeneous records. In particular, same-surgery references should help identify physician- and surgery-specific documentation patterns without introducing information from prior patients that is not supported by the current patient's source records. We therefore evaluate both whether generated facts are supported by the source records and whether they align with the physician-written note, while also measuring coverage of the note's atomic facts.

3.2 Data Construction

We use de-identified clinical records collected at SNUH. The source records and physician-written postoperative outpatient notes contain both Korean and English text. All records were accessed in a read-only setting. Each patient case consists of six source records and one physician-written postoperative outpatient note. The dataset covers two surgical divisions: Breast–Endocrine Surgery and Colorectal Surgery, involving five physicians in total. We apply light rule-based preprocessing to remove metadata headers and formatting tags. After normalizing procedure names, we selected the four most frequent procedure groups in Breast–Endocrine Surgery and the three most frequent procedure groups in Colorectal Surgery and used cases corresponding to these procedure groups.

We construct two patient-disjoint case sets: an evaluation set for generation and evaluation, and a reference-construction set for building same-surgery references. The data cover seven surgery groups: four from Breast–Endocrine Surgery and three from Colorectal Surgery. For each physician–

surgery pair, we use 10 evaluation cases and 12 prior reference-construction cases, yielding 170 and 204 cases, respectively, as shown in Table 1. This separation ensures that no evaluation case is used to construct the same-surgery references.

Table 1: Evaluation and reference case counts.

Division	Phys.	Eval.	Ref.
Breast–Endocrine	2	$2\times4\times10 = 80$	$2\times4\times12 = 96$
Colorectal	3	$3\times3\times10 = 90$	$3\times3\times12 = 108$
Total	5	170	204

4 Method

Figure 1 summarizes the reference construction and generation variants.

4.1 Same-Surgery Reference Construction

For each physician–surgery pair, we construct a same-surgery reference from the corresponding reference-construction cases. Each reference case consists of six source records and the physician-written first postoperative outpatient note. The goal of the reference is not to provide prior notes as examples to copy, but to structure how information from each source record is typically reflected in the final outpatient note.

The reference is represented as a record-specific reflection map. For each source record type, it stores frequently consulted locations, repeatedly reflected items, and reflection modes such as copy, summary, reconstruct, or inferred. It is intended to guide what to preserve and how to rewrite it, rather than to serve as patient-specific evidence. Appendix A provides an example JSON format.

In our experimental setting, a separate offline reference-construction step, termed the Reference Writer, reads 12 prior cases for the corresponding physician–surgery pair and produces one JSON reflection map. It uses Claude Opus 4.6, which is distinct from the inference models used for note generation. The constructed reference is fixed for the corresponding physician–surgery pair and is used unchanged by the reference-based variants. C3 and C3+G use it only during per-record summarization, whereas C4 and C4+G use it during both per-record summarization and final merging. C1 and C2 are reference-free baselines.

4.2 Generation Variants

We compare four generation variants. C1 is a single-call baseline, where all six source records are provided at once and the model directly generates the final outpatient note. C2 is a decompose-and-merge baseline without references: the model first summarizes each source record separately and then merges the six summaries into the final note.

C3 uses the reference only during per-record summarization. Each summary call receives the source record together with the corresponding reference slice for that record type. C4 uses the reference during both per-record summarization and final merging. The summarization stage uses the same record-specific reference slices as C3, whereas the merge stage receives the same-surgery reference and can use it to select and organize information across the six intermediate summaries. Thus, C2–C4 share the same decompose-and-merge structure, while C3 and C4 differ specifically in whether the reference is available during final merging.

4.3 Grounding Constraint

Because same-surgery references are derived from other patients, they can encourage unsupported transfer. We therefore apply a simple grounding constraint to reference-based variants: the reference may guide structure, item selection, and writing flow, but clinical content, numbers, dates, and treatment plans must be supported by the source records. Items not verified in the source records are omitted.

This grounding constraint is applied at inference time and requires no retraining or modified decoding. We apply it to C3 and C4, denoted C3+G and C4+G, as an auxiliary safety constraint for same-surgery references rather than a new model architecture.

4.4 Models and Implementation

We use Claude Sonnet 4.5, Claude Opus 4.5, and Qwen3-235B as inference models, all run with temperature 0. Per-record summaries target approximately 300 characters, and final outputs follow the average physician-written note length of 1,388 characters. Model identifiers and implementation details are provided in Appendix B.

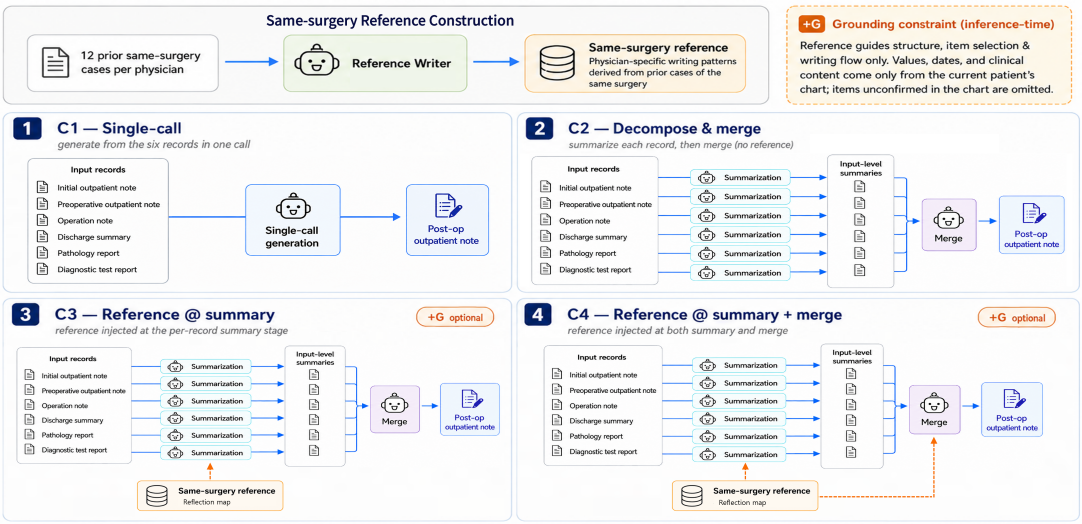


Figure 1: Method overview. We construct one same-surgery reference per physician–surgery pair from 12 prior cases. C1 uses a single call; C2 decomposes and merges without reference; C3 uses the reference during per-record summarization only; and C4 uses it during both per-record summarization and final merging. +G denotes the optional inference-time grounding constraint.

5 Evaluation

5.1 Automatic Metrics

We measure surface similarity to the physician-written target note using ROUGE-1/2/L, BLEU-1/4, and BERTScore (Lin, 2004; Papineni et al., 2002; Zhang et al., 2020). We report BLEU-1/4 together with the other surface metrics in Table 3 of Appendix E. These metrics are useful for comparison with prior generation work, but they may reward note-like surface overlap induced by references and do not fully capture clinical faithfulness (Maynez et al., 2020). We therefore complement surface metrics with fact-level faithfulness evaluation.

5.2 Two-Source FactScore

We decompose each generated note into atomic facts and verify each fact independently against two sources: the source records and the physician-written gold note. Here, input refers to the source records, whereas gold refers to the physician-written note. This yields four categories: input and gold, input only, gold only (standard-care prediction), and neither (hallucination). Let F_{gen} denote the atomic facts extracted from a generated note. We define FS_{in} as the fraction of atomic facts in F_{gen} supported by the input, and FS_{gt} as the fraction of atomic facts supported by gold. A schematic is provided in Appendix C.

This setup extends FactScore (Min et al., 2023) to two sources, allowing us to distinguish target-consistent facts from hallucinations, which a single-source evaluation may conflate (Cao et al., 2022). We use gpt-oss-120b as the LLM verifier, which is separate from the three inference models used for generation. The verifier uses only the provided source text—the source records or physician-written note—without external retrieval, with all judgments run at temperature 0.

5.3 Validation and Coverage Checks

To assess the reliability of LLM-based fact verification, we manually label 200 stratified fact–source pairs while blinded to the verifier output, with 100 pairs from each surgical division. The verifier agrees with the human labels on 0.91 of the pairs and achieves a Cohen's κ (Cohen, 1960) of 0.82 overall and in each division. It attains precision/recall of 0.95/0.88 against the human labels, suggesting that the verifier is reliable but conservative. Detailed verifier validation results are provided in Appendix D.

For gold-fact coverage, we first decompose the physician-written gold note into atomic facts, denoted by F_{gold}. We then use the same verifier to check whether each gold fact is covered by the generated note and report the fraction covered. This recall-oriented metric measures the proportion of physician-selected facts covered by the generated note, rather than the proportion of generated facts supported by the source records, helping determine whether a lower hallucination rate reflects shorter outputs or preservation of physician-selected con-

Table 2: Main results on the original 170-case evaluation, pooled across the three inference models (Sonnet 4.5, Opus 4.5, Qwen3-235B). R-L: ROUGE-L F1; BERT: BERTScore F1; FS_{in}/FS_{gt}: two-source FactScore against the source records and the physician-written note; Hall.: hallucination rate (category 4, %, lower is better); Cov.: gold-fact coverage. Best per column within each division in **bold**. Full metric details are provided in Appendices E–G.

Variant	R-L	BERT	FS_{in}	FS_{gt}	Hall. ↓	Cov.
Breast–Endocrine Surgery						
C1	.298	.707	**.828**	.387	14.7	**.441**
C2	.223	.666	.779	.346	19.0	.367
C3	.290	.697	.815	.381	15.3	.409
C3+G	.290	.699	.812	.378	15.7	.413
C4	.394	.740	.772	.499	15.9	.440
C4+G	**.396**	**.741**	.827	**.519**	**10.9**	.427
Colorectal Surgery						
C1	.166	.677	.825	.209	16.1	.299
C2	.133	.642	.768	.191	21.7	.271
C3	.163	.667	.787	.217	19.3	.300
C3+G	.171	.672	.805	.220	17.8	.310
C4	.241	.696	.752	.258	20.3	.328
C4+G	**.252**	**.702**	**.852**	**.287**	**10.8**	**.339**

tent. Because the physician-written note records information selected as necessary for a visit rather than every chart-supported fact, we interpret coverage together with FS_{in} and FS_{gt}.

6 Results and Analysis

Table 2 reports results pooled across the three inference models. Overall, we observe four patterns.

Decomposition alone underperforms. C2 performs worse than the single-call baseline C1 in both divisions. FS_{gt} decreases from .387 to .346 in Breast–Endocrine Surgery and from .209 to .191 in Colorectal Surgery, while hallucination increases. This suggests that naive decompose-and-merge can lose details needed for the final note during intermediate summarization.

References help primarily at the merge stage. C3, which uses the reference only during per-record summarization, remains close to C1. In contrast, C4 carries the reference into the merge stage and improves alignment with physician-written notes, raising FS_{gt} to .499 in Breast–Endocrine Surgery and .258 in Colorectal Surgery. This supports our hypothesis that same-surgery references mainly help select and organize information across records.

Cross-patient references introduce a fidelity cost. C4 improves target alignment but reduces input support: FS_{in} drops from .828 to .772 in Breast–Endocrine Surgery and from .825 to .752 in Colorectal Surgery. Hallucination also remains high under C4, at 15.9% and 20.3%. Reference-overlap analysis shows that some unsupported facts match the reference text, indicating cross-patient transfer rather than unconstrained invention.

Grounding mitigates the cost. C4+G yields the lowest hallucination rates in both divisions, 10.9% and 10.8%, while restoring FS_{in} to .827 and .852. FS_{gt} is maintained or improved relative to C4, reaching .519 and .287. Significance tests confirm fact-level gains without surface-metric loss, and coverage argues against a terseness-only explanation. Together, these results suggest that same-surgery references are best used as documentation-pattern guidance, not patient evidence; clinical content must remain grounded in the source records. Full metric tables and auxiliary analyses are provided in Appendices E–I.

Expanded-scale evaluation. To examine whether these trends persist at a larger within-institution scale, we applied the same generation and evaluation protocol with the inference models (Claude Sonnet 4.5, Claude Opus 4.5, and Qwen3-235B) to 1,329 cases across 10 surgery groups: 600 Breast–Endocrine cases and 729 Colorectal cases. The expanded evaluation included up to 300 cases per physician (129 for one physician with fewer available cases). We retained the same patient-disjoint reference construction and preprocessing, with no evaluation case used to build its reference. For Breast–Endocrine, C4+G improved FS_{in} from .775 to .830 and FS_{gt} from .571 to .596, while reducing hallucination from 15.1% to 10.4%; gold-fact coverage was .470 for C4+G versus .478 for C4. For Colorectal, C4+G improved FS_{in} from .786 to .838 and FS_{gt} from .233 to .252, while reducing hallucination from 17.2% to 12.0%; gold-fact coverage was .315 for C4+G versus .328 for C4. FactScore and hallucination values use atomic-fact-count pooling across the three models, whereas coverage is the case-level mean over generated case–model pairs. Because the expanded set retains the original selected cases and adds available cases, including three additional surgery groups for one physician, we treat this analysis as a within-institution scale check rather than an independent replication or a new significance test. Model-level, division-separated results for all six variants are reported in Appendix J.

Limitations

This study should be interpreted within its clinical and institutional scope. The experiments use de-identified clinical records collected at SNUH, spanning two surgical divisions and five physicians. While this setting provides real postoperative outpatient notes and physician-specific documentation patterns, future work should test whether the findings transfer to other institutions, specialties, languages, and documentation systems, as Korean medical LLM resources continue to develop (Seo et al., 2025).

The same-surgery references are constructed from 12 prior cases for each physician–surgery pair. This design allows us to capture recurring documentation patterns with a fixed and leakage-free reference set, but larger reference pools may be needed for rare procedures, physicians with limited prior cases, or settings where note-writing conventions change over time.

Our factuality evaluation relies on LLM-based atomic fact verification. We validate the verifier with blinded human annotation and observe high agreement, but the validation set remains limited. Future work could expand human evaluation and examine clinically acceptable inferences that may not be explicitly stated in the source records.

Finally, Appendix K reports a pilot few-shot comparison conducted on one physician, one division, and one model. We include it only as a preliminary comparison between raw prior-note examples and structured same-surgery references, rather than as a general conclusion about few-shot prompting.

Ethical Considerations

This study uses de-identified clinical records from SNUH. The use of these records was approved by the Seoul National University Hospital Institutional Review Board (IRB No. 2402-099-1512). All data were handled in a read-only research environment, and no direct patient identifiers are included in the experimental inputs, outputs, or examples. Because the data are clinical and not publicly shareable, we report aggregate results and provide prompts, model settings, and anonymized reference-format examples to support reproducibility without exposing patient information.

The generated notes in this study are research outputs and are not intended for clinical deployment or autonomous use. Any clinical use of similar systems would require institutional review, privacy safeguards, prospective validation, and physician oversight (Jung, 2025). In particular, same-surgery references should not be treated as patient-specific evidence; they are used only to guide documentation structure and item selection, while clinical content should be grounded in the source records.

References

Anthropic. 2025a. Claude Opus 4.5 System Card.

Anthropic. 2025b. Claude Sonnet 4.5 System Card.

Anthropic. 2026. Claude Opus 4.6 System Card.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Kyu-Hwan Jung. 2025. Large language models in medicine: Clinical applications, technical challenges, and ethical considerations. *Healthcare Informatics Research*, 31(2):114–124.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.

FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jean Seo, Sumin Park, Sungjoo Byun, Jinwook Choi, Jinho D. Choi, and Hyopil Shin. 2025. Advancing Korean medical large language models: Automated pipeline for korean medical preference dataset construction. *Healthcare Informatics Research*, 31(2):166–174.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

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. *Scientific Data*, 10(586).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Same-Surgery Reference Format

Each reference is a JSON object summarizing how the six source record types are reflected in physician-written outpatient notes across 12 prior same-surgery cases. The `input_reflection` field contains one entry per source record type. Its `granularity` adapts to the record structure: `section` for free-text notes, `table_row` for operation notes, `field_per_test` for pathology reports, and `test_x_timepoint` for laboratory results.

Within each entry, reflected items are annotated with a `located_in` hint, a `reflection_mode` (`copy`, `summary`, `reconstruct`, or `inferred`), and an observed `frequency` bucket (`high`, `med`, or `low`). At the top level, `output_section_breakdown` summarizes typical information flow into each output section, and `common_sentences` lists recurring expressions. The reference is built by Claude Opus 4.6 from the 12 prior cases alone and contains no information about the current patient. A simplified, anonymized excerpt is shown below.

```
{
 "input_reflection": {
   "operation_note": {
     "granularity": "table_row",
     "items": [
       {"item": "operative extent",
     "located_in": "extent <- operation",
        "reflection_mode": "copy",
        "frequency": "high"},
       {"item": "sentinel node biopsy",
        "located_in":
            "number <- SLNB <- axilla",
        "reflection_mode": "copy",
        "frequency": "high"}]},
   "pathology_report": {
     "granularity": "field_per_test",
     "items": [
       {"item": "final diagnosis",
        "reflection_mode": "copy",
        "frequency": "high"}]}
 },
 "output_section_breakdown": {
   "Findings":
     "pre-op dx -> op date+name -> ...",
   "Assessment & Plan":
     "staging -> follow-up plan"
 },
 "common_sentences":
    ["postop 1st visit", "wound: OK"]
}
```

The full prompts for reference construction, per-record summarization, merging, and grounding are included in the supplementary material.

B Models and Hyperparameters

Inference models are Claude Sonnet 4.5 (Anthropic, 2025b), Claude Opus 4.5 (Anthropic, 2025a), and Qwen3-235B (Yang et al., 2025). All model calls were made through Amazon Bedrock under the applicable provider and platform terms. All generations use temperature 0. Case selection used a fixed random seed of 42. Per-record summaries target approximately 300 characters; the final note targets approximately 1,388 characters. Same-surgery references are constructed by Claude Opus 4.6, which is not used as an inference model (Anthropic, 2026). The two-source FactScore verifier is gpt-oss-120b, also outside the inference set, run retrieval-free at temperature 0 (OpenAI, 2025).

C Two-Source FactScore Schematic

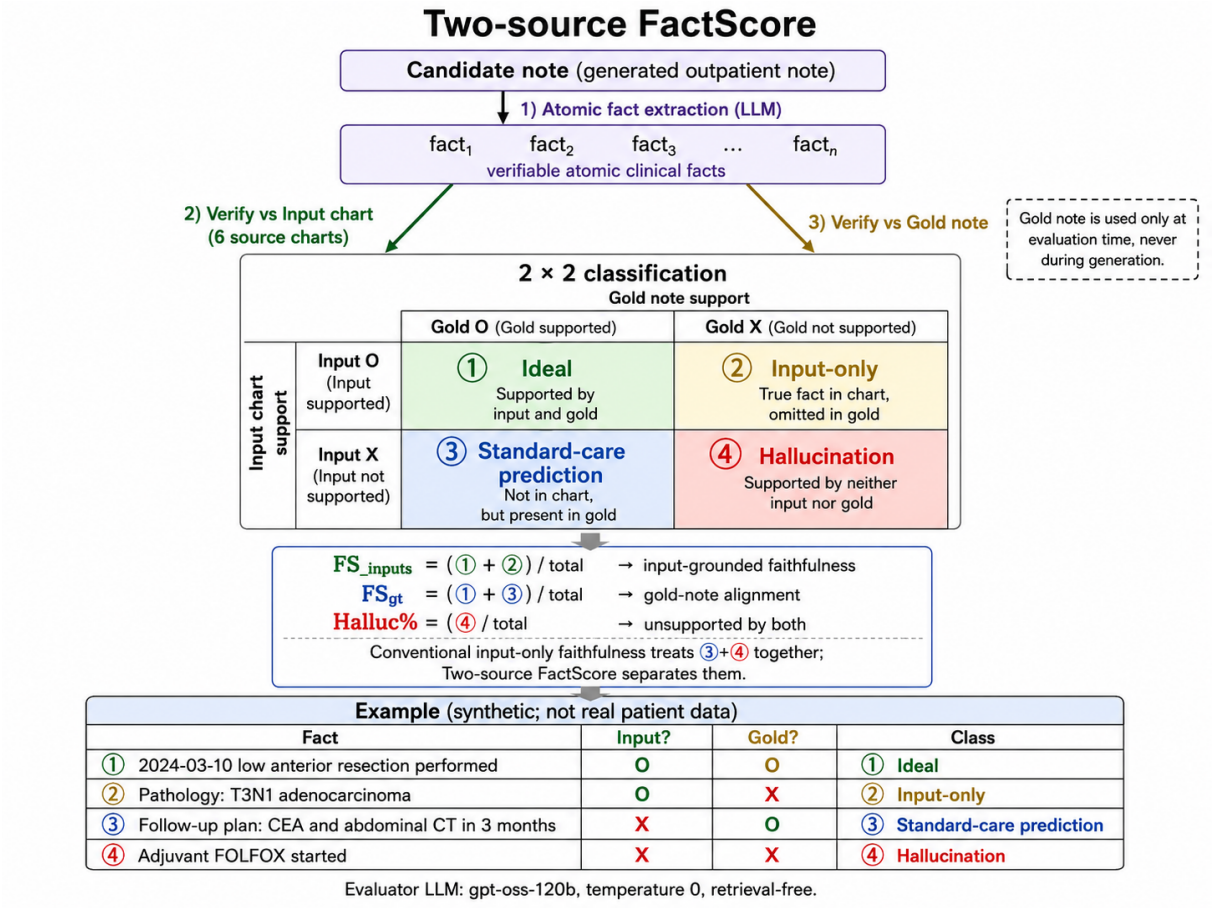


Figure 2: Two-source FactScore. Each atomic fact extracted from a generated note is verified against both the source records and the physician-written gold note, yielding four categories: (1) supported by both, (2) input only, (3) gold only, and (4) neither, i.e., hallucination. $\mathrm{FS}_{\mathrm{in}}$ and $\mathrm{FS}_{\mathrm{gt}}$ aggregate categories (1)+(2) and (1)+(3), respectively.

D Verifier Validation

We validate the two-source FactScore verifier against blind human annotation of 200 stratified fact–source pairs (100 per division). The verifier achieves 0.91 agreement with human labels, measured using Cohen's κ ($\kappa = 0.82$ overall; $\kappa = 0.82$ in each division), and attains precision/recall of 0.95/0.88 against human labels. These results suggest that the verifier is reliable but conservative.

E Full Automatic Metrics

Table 3: Full automatic metrics on Breast–Endocrine and Colorectal Surgery. Breast–Endocrine: n=80/variant; Colorectal: n=90/variant. R-1/2/L: ROUGE F1; BERT: BERTScore F1.

Model	Variant	R-1	R-2	R-L	BLEU-1	BLEU-4	BERT
Breast–Endocrine Surgery (n=80/variant)							
Sonnet 4.5	C1	.452	.281	.329	.248	.135	.729
	C2	.375	.170	.241	.104	.024	.673
	C3	.436	.270	.330	.177	.075	.701
	C3+G	.437	.268	.327	.187	.086	.704
	C4	.535	.394	.445	.331	.205	.758
	C4+G	**.543**	**.400**	**.450**	**.346**	**.232**	**.763**
Opus 4.5	C1	.460	.294	.344	.259	.148	.729
	C2	.385	.171	.249	.121	.027	.676
	C3	.438	.266	.317	.235	.127	.722
	C3+G	.434	.262	.318	.233	.126	.725
	C4	.508	.352	.412	**.322**	**.203**	**.750**
	C4+G	**.509**	**.361**	**.418**	.302	.197	**.750**
Qwen3-235B	C1	.332	.145	.220	.071	.009	.663
	C2	.286	.106	.180	.058	.003	.648
	C3	.347	.142	.223	.075	.007	.668
	C3+G	.343	.145	.226	.072	.008	.667
	C4	**.443**	**.262**	**.325**	**.187**	**.097**	**.712**
	C4+G	.439	.258	.320	.175	.083	.709
Colorectal Surgery (n=90/variant)							
Sonnet 4.5	C1	.290	.136	.187	.132	.044	.690
	C2	.235	.078	.137	.061	.009	.639
	C3	.265	.122	.172	.088	.023	.655
	C3+G	.273	.124	.172	.090	.023	.662
	C4	.346	.204	.258	.165	.075	.697
	C4+G	**.363**	**.217**	**.271**	**.172**	**.080**	**.709**
Opus 4.5	C1	.291	.134	.180	.132	.043	.695
	C2	.260	.091	.148	.073	.010	.656
	C3	.291	.135	.179	.123	.035	.691
	C3+G	.300	.146	.192	.127	.042	.697
	C4	.345	.190	.249	.180	.070	.711
	C4+G	**.368**	**.217**	**.276**	**.186**	**.082**	**.717**
Qwen3-235B	C1	.219	.071	.130	.049	.006	.647
	C2	.187	.050	.113	.032	.002	.630
	C3	.242	.085	.139	.056	.006	.655
	C3+G	.249	.091	.148	.059	.007	.658
	C4	**.310**	**.148**	**.215**	**.098**	**.020**	.679
	C4+G	.303	.143	.209	.093	**.020**	**.681**

F Full Two-Source FactScore

Table 4: Full two-source FactScore on Breast–Endocrine and Colorectal Surgery. FS_{in}/FS_{gt}: fractions of atomic facts supported by the source records and by the physician-written gold note, respectively. 2×2 cells (%): (1) input and gold; (2) input only; (3) gold only (standard-care prediction); (4) neither (hallucination).

Model	Variant	#facts	FS_{in}	FS_{gt}	(1)	(2)	(3)	(4)
Breast–Endocrine Surgery								
Sonnet 4.5	C1	44.5	.863	.399	37.4	48.9	2.5	11.2
	C2	43.0	.820	.344	31.5	50.5	2.9	15.1
	C3	46.2	.875	.399	36.9	50.6	2.9	9.6
	C3+G	46.0	**.881**	.376	35.0	53.1	2.5	**9.3**
	C4	41.0	.807	.525	46.3	34.4	6.1	13.2
	C4+G	38.1	.831	**.541**	**47.1**	36.0	7.0	9.9
Opus 4.5	C1	51.4	**.878**	.391	36.8	51.0	2.3	9.9
	C2	51.9	.826	.347	31.5	51.1	3.2	14.3
	C3	50.0	.855	.381	34.8	50.7	3.3	11.2
	C3+G	49.8	.828	.384	34.9	47.8	3.4	13.8
	C4	37.1	.800	.514	44.7	35.3	6.7	13.3
	C4+G	32.9	.864	**.556**	**49.6**	36.7	6.0	**7.6**
Qwen3-235B	C1	41.8	.742	.372	34.3	39.9	2.9	22.9
	C2	40.5	.692	.346	31.4	37.8	3.2	27.6
	C3	44.3	.716	.363	33.1	38.5	3.2	25.2
	C3+G	42.2	.728	.374	34.2	38.6	3.2	24.0
	C4	38.7	.708	.457	37.8	33.0	7.9	21.3
	C4+G	38.0	**.785**	**.461**	**39.9**	38.7	6.3	**15.2**
Colorectal Surgery								
Sonnet 4.5	C1	49.9	.855	.217	20.2	65.3	1.5	13.0
	C2	48.2	.833	.193	18.4	64.9	0.9	15.8
	C3	46.1	.819	.232	20.9	61.0	2.3	15.8
	C3+G	47.6	.855	.230	21.4	64.1	1.7	12.8
	C4	37.2	.757	.268	22.3	53.4	4.5	19.8
	C4+G	38.5	**.870**	**.282**	**24.1**	62.9	4.2	**8.9**
Opus 4.5	C1	58.6	.885	.213	19.7	68.9	1.7	9.8
	C2	61.2	.818	.198	17.7	64.1	2.0	16.2
	C3	55.6	.847	.216	19.5	65.2	2.1	13.3
	C3+G	55.3	.840	.232	21.1	62.9	2.1	13.9
	C4	43.3	.793	.257	22.0	57.2	3.7	17.0
	C4+G	36.2	**.905**	**.319**	**28.7**	61.8	3.2	**6.3**
Qwen3-235B	C1	46.1	.734	.196	18.6	54.8	1.0	25.6
	C2	47.6	.653	.181	16.4	48.9	1.7	33.0
	C3	48.3	.694	.203	18.5	50.9	1.8	28.8
	C3+G	47.9	.719	.198	18.3	53.6	1.5	26.7
	C4	42.9	.706	.248	19.6	51.1	5.3	24.1
	C4+G	41.6	**.782**	**.261**	**21.5**	56.7	4.6	**17.2**

G Gold-Fact Coverage

Table 5 reports gold-fact coverage (recall): the fraction of atomic facts in the physician-written note that are covered by the generated note.

Table 5: Gold-fact coverage (3-model pooled). C4+G maintains coverage comparable to the single-call baseline in Breast–Endocrine Surgery and higher than the baseline in Colorectal Surgery, despite producing fewer facts, indicating that its precision gains are not merely due to terser outputs.

Variant	Breast	Colorectal
C1	**.441**	.299
C2	.367	.271
C3	.409	.300
C3+G	.413	.310
C4	.440	.328
C4+G	.427	**.339**

H Significance Tests

We test paired differences with a 10,000-sample paired bootstrap and the Wilcoxon signed-rank test, pairing by case and model (breast n=240, colorectal n=270). C4+G improves over C4 on FS_{in}, FS_{gt}, and hallucination (cell 4) with $p < .01$ in both divisions, with no significant loss on surface metrics. Against the single-call baseline (C1), C4+G is significantly better on FS_{gt}, hallucination, ROUGE-L, and BERTScore ($p < .001$), and is comparable (breast) or significantly better (colorectal) on FS_{in}.

I Reference Leakage

To probe the source of hallucinations, we measure how many of C4's unsupported facts (cell 4) match the reference text. Of these facts, 34% (breast) and 20% (colorectal) overlap the reference verbatim, indicating transfer from the reference rather than free-form invention. Applying the grounding constraint reduces reference-originating hallucinations by 41% (breast) and 45% (colorectal).

J Expanded-Scale Evaluation

We repeated the generation and evaluation protocol on 1,329 cases spanning 10 surgery groups, including 600 Breast–Endocrine cases and 729 Colorectal cases. Four physicians contributed 300 cases each, and one physician contributed 129 cases because fewer eligible cases were available. The table below summarizes the expanded results separately by surgical division, model, and generation variant.

Table 6: Model-level results for the expanded-scale evaluation by surgical division and generation variant. Each model–variant condition uses 600 cases in Breast–Endocrine Surgery and 729 cases in Colorectal Surgery. FS_{in}, FS_{gt}, and Hall. use atomic-fact-count pooling within each model–variant condition, whereas Coverage is the case-level mean within each model–variant condition. Hall. is reported as a percentage; bold indicates the best value within each model–division block.

Model	Variant	FS_{in}	FS_{gt}	Hall.↓	Coverage
Breast–Endocrine Surgery (n = 600 cases per variant and model; 4 surgery groups)					
Sonnet 4.5	C1	.845	.470	12.0	.489
	C2	.847	.390	12.6	.406
	C3	**.878**	.453	**9.1**	.480
	C3+G	.876	.452	9.3	.481
	C4	.808	.617	12.4	.517
	C4+G	.842	**.636**	9.4	**.519**
Opus 4.5	C1	**.886**	.446	8.6	**.530**
	C2	.826	.385	14.0	.465
	C3	.858	.423	11.0	.484
	C3+G	.849	.438	11.6	.477
	C4	.799	.595	12.6	.494
	C4+G	.861	**.644**	**7.3**	.471
Qwen3-235B	C1	.749	.426	21.9	.415
	C2	.689	.381	27.5	.348
	C3	.716	.405	24.9	.390
	C3+G	.727	.412	24.0	.396
	C4	.717	.501	20.5	**.424**
	C4+G	**.789**	**.510**	**14.4**	.419
Colorectal Surgery (n = 729 cases per variant and model; 6 surgery groups)					
Sonnet 4.5	C1	.861	.183	12.2	.289
	C2	.826	.169	16.4	.253
	C3	.831	.199	14.9	.293
	C3+G	.837	.194	14.4	.297
	C4	.831	.241	12.8	**.343**
	C4+G	**.866**	**.249**	**9.4**	.339
Opus 4.5	C1	**.882**	.193	9.8	**.354**
	C2	.753	.159	22.3	.252
	C3	.813	.194	16.3	.316
	C3+G	.810	.204	16.6	.303
	C4	.822	.234	14.2	.326
	C4+G	.880	**.284**	**8.1**	.295
Qwen3-235B	C1	.734	.165	25.2	.252
	C2	.642	.165	33.9	.250
	C3	.688	.169	29.3	.265
	C3+G	.711	.171	27.1	.264
	C4	.707	.224	24.3	**.315**
	C4+G	**.781**	**.231**	**17.3**	.310

K Pilot Few-Shot Baseline

A natural alternative to structured same-surgery references is to provide raw prior notes from the same surgery as few-shot examples. We therefore conduct a small pilot comparison to test whether the gains from C4 can be explained simply by exposing the model to prior same-surgery notes.

This pilot is limited to one physician in Colorectal Surgery, one inference model (Claude Opus 4.5), and 30 evaluation cases. All compared variants are re-evaluated on the same 30 cases, so the absolute values are not directly comparable to the full-set results in the main tables. For the few-shot base-

line, denoted C5, we use the same decomposition setting as the no-reference baseline and provide three patient-disjoint physician-written prior notes from the same surgery group at the merge stage. The prior notes are selected from the reference pool using the same fixed seed and do not include the current patient's target note.

Table 7 reports the pilot results. C5 improves ROUGE-L over C1 (.162 vs. .133), suggesting that raw prior notes help surface form. However, its FS_{gt} remains comparable to C1 (.142 vs. .145), and its hallucination rate is higher (17.4% vs. 14.1%). In contrast, C4+G achieves substantially higher FS_{gt} (.217) and lower hallucination (7.0%). These results suggest that the benefit does not come merely from showing raw same-surgery examples, but from structuring the reference and grounding generated content in the source records. Because this is a pilot study limited to one physician, one division, and one model, we do not draw general conclusions about all few-shot prompting designs.

Table 7: Pilot raw few-shot comparison on one colorectal-surgery physician, Claude Opus 4.5, and 30 cases. C5 provides three raw same-surgery prior notes at the merge stage instead of a structured same-surgery reference. Hall.: hallucination rate; Cov.: gold-fact coverage.

Variant	#facts	Cov.	FS_{in}	FS_{gt}	Hall. ↓	R-L
C1	57.1	.195	.840	.145	14.1	.133
C2	59.5	.186	.770	.129	20.7	.116
C4	44.5	.178	.818	.178	15.6	.164
C4+G	34.1	.182	**.905**	**.217**	**7.0**	**.168**
C5	46.1	.173	.801	.142	17.4	.162