

Resource-Aware Retrieval Baselines for CBCT-to-Report Generation in ToothFairy4



DukeHealth

Zachary Mahmoud¹ and Dominic LaBella²

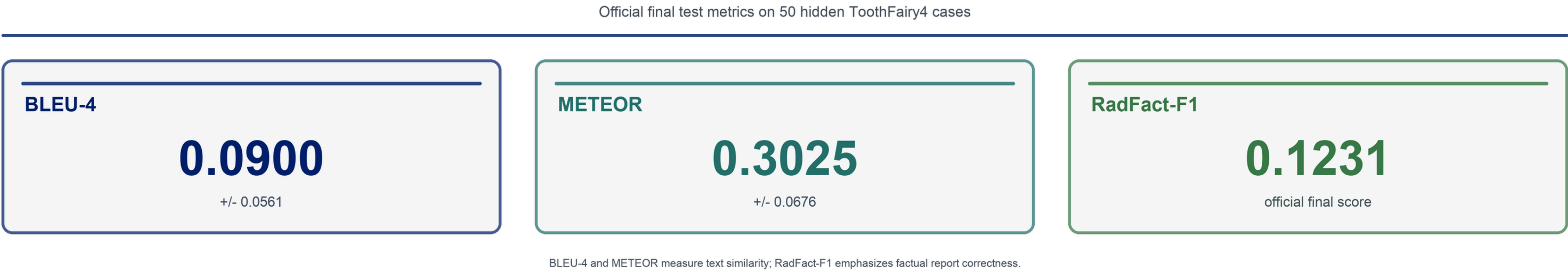
¹ Duke University

² Department of Radiation Oncology, Duke University Medical Center

ODIN 2026 / ToothFairy4 Task 1

Core Takeaway

A CPU-only retrieval system can provide a transparent and reproducible ToothFairy4 baseline. Its main limitation is factual rather than grammatical: copied reports can include unsupported patient-specific findings.



Clinical Problem:

ToothFairy4 extends dental CBCT analysis from segmentation to report generation. Given a 3D CBCT volume, the algorithm must produce a diagnostic imaging report suitable for maxillofacial and surgical planning. Unlike segmentation, report generation must decide which findings are present and describe them without adding unsupported details about teeth, implants, impacted teeth, granulomas, or anatomical risk relationships.

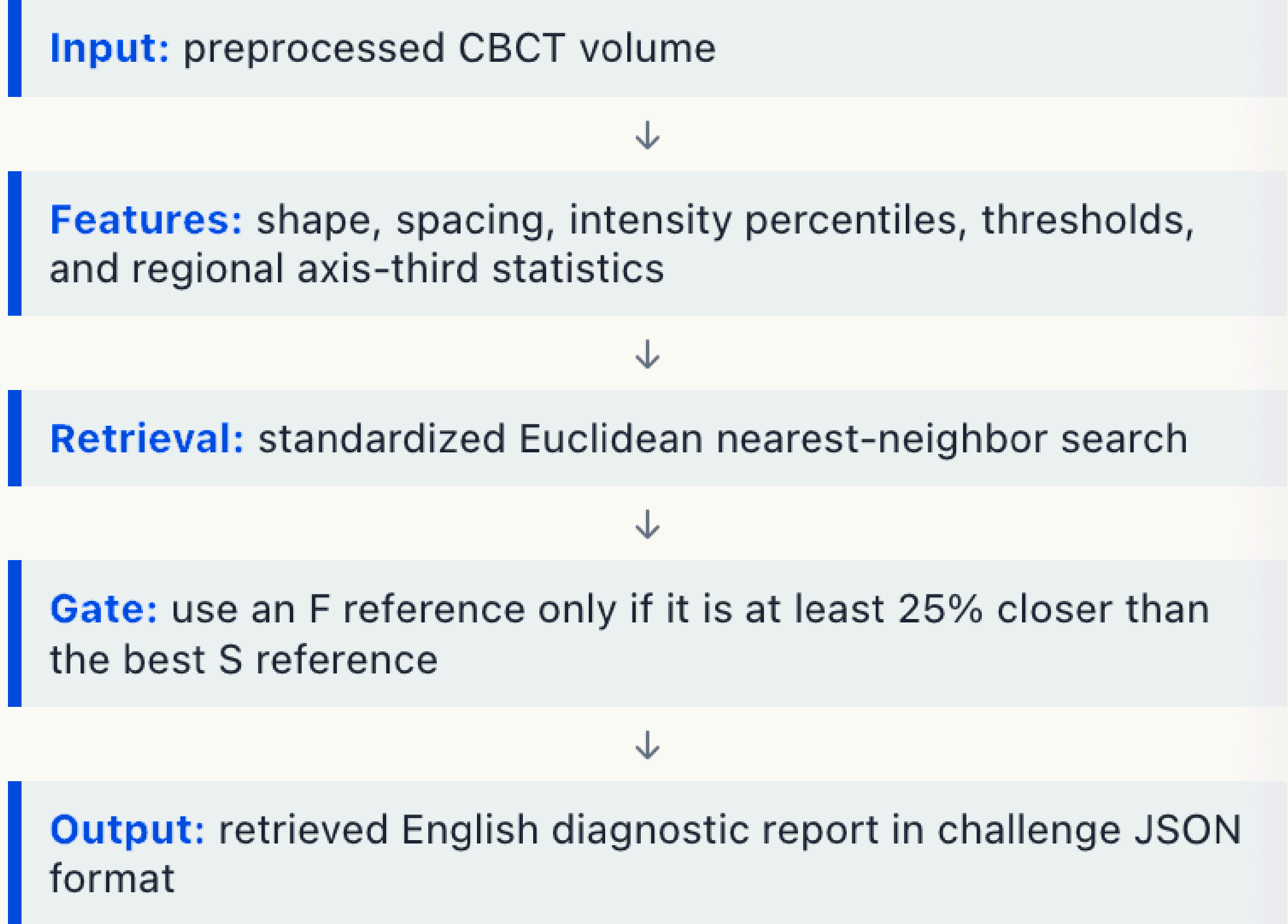
Resource-Aware Goal:

- Developed on a Mac laptop without an NVIDIA GPU
- Large 3D vision-language model training was outside practical scope
- The project focused on a transparent retrieval baseline that could be analyzed, packaged, and submitted

Data and Local Evaluation

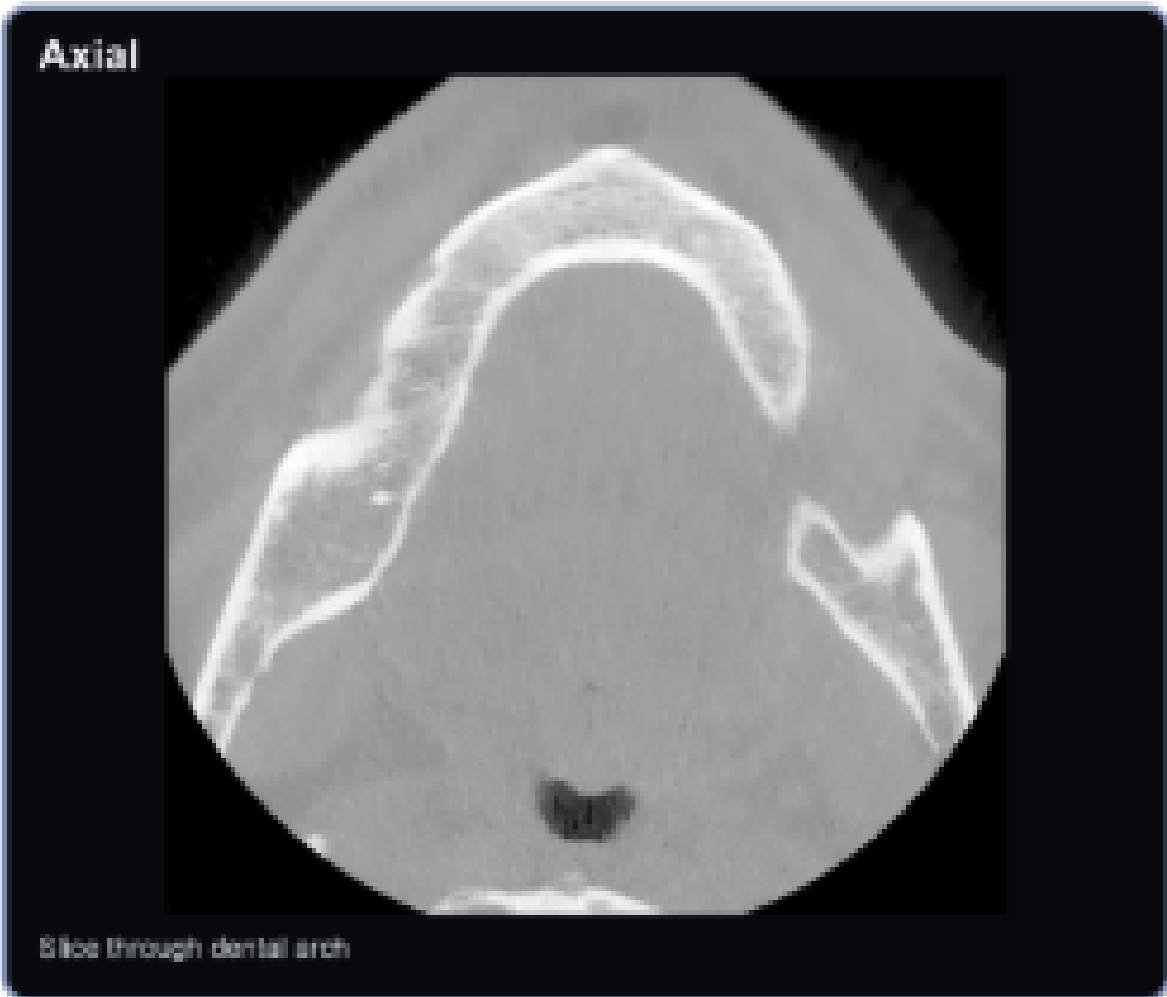
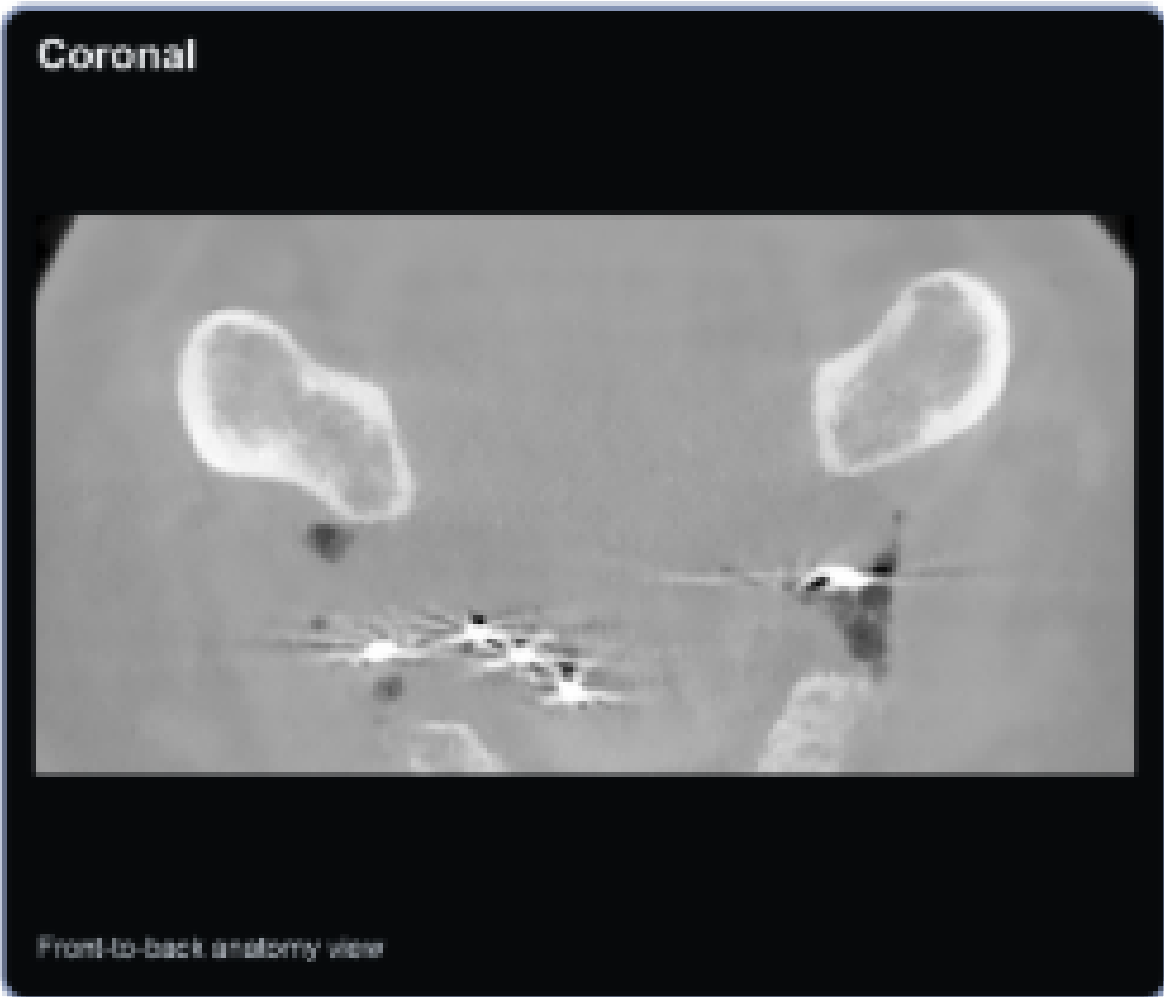
- 52 public S cases used for local development.
- Fixed 42/10 train-validation split for rapid testing.
- Leave-one-out analysis across all 52 S cases for a stability check.
- 63 F cases used only as an external reference library.

Retrieval Pipeline



Method Variants Tested

Variant	Why test it?
One-neighbor copied report	Simple baseline and container workflow validation.
Ranked key-term evidence	Measure sensitivity vs unsupported extra terms.
Natural report reranking	Preserve fluent report text while penalizing risky terms.
global_axis3 features	Add coarse regional information without segmentation labels.
Gated S+F retrieval	Use extra references only when image evidence is clearly stronger.



Case S0023 input CBCT: representative axial, coronal, and sagittal slices.

What Improved?

The most stable improvement came from better scan comparison. The global_axis3 representation improved local BLEU-4 and METEOR, and the gated S+F strategy produced the best local fixed-split metrics while avoiding blind pooling of all external references.

Error Analysis:

The main failure mode was whole-report copying. Retrieved reports were fluent because they were clinician-authored, but they sometimes carried over findings that were not supported by the input scan.

Ranked generic reports reduced extra key terms, but they became too short and generic, causing BLEU-4 and METEOR to fall.

Case Example: S0023

S-ONLY RETRIEVAL
0.242
Selected S0041. Missed endodontic and introduced teeth 47, and 48.

GATED S+F RETRIEVAL
0.314
Selected F006. Missed no evaluated key terms, but added granuloma and impacted.

Leave-One-Out Analysis

Method	Word	Missed	Extra	F selected
Original global retrieval	0.381	2.52	2.79	0/52
global_axis3 retrieval	0.401	2.42	2.90	0/52
Blind S+F global_axis3	0.393	2.52	2.69	4/52
Gated S+F global_axis3	0.404	2.40	2.85	2/52

Limitations

- The fixed local validation split contained only 10 cases.
- Features did not represent tooth-level anatomy, canal localization, or pathology.
- Key-term analysis used simple string matching and did not model negation or uncertainty.
- The system retrieved and copied real reports; it did not truly infer a structured dental report from the CBCT volume.

Future Work

A stronger lightweight system would first predict a structured checklist of findings, anatomical locations, and qualifiers, then render a report from that representation

Conclusion

Retrieval can be a practical ToothFairy4 baseline, but the results show that scan-comparison quality matters more than complicated report reranking. Future systems should move from whole-report copying toward structured finding prediction.