

Brihadaranyaka at MEDIQA-CORE-Task-2 2026: A Text-Only Zero-Shot Pipeline with Explicit Severity-Decoupling Rules for Radiology Report Discrepancy Assessment

Anubhav^{1,*}

¹Independent Researcher, Bokaro Steel City, Jharkhand, India

Abstract

We describe Brihadaranyaka's submission to MEDIQA-CORE 2026 Task 2 (Radiology Report Discrepancy Assessment, RADAR benchmark). Our system is a single-call zero-shot pipeline using claude-sonnet-4-6 at temperature 0 with text-only input (preliminary report plus candidate edit; the CT volume is deliberately not provided to the model). The principal contribution is the prompt itself: an explicit severity-decoupling rule (a wrong edit can still be negligible), a counterfactual test ("would the patient's next 24–48 hours of care change?"), and five golden rules that resist the natural drift to "moderate" as a safe middle. The team submitted four runs (1027, 1028, 1030, 1031); the organizers report the best per-metric run for the team-level ranking. On the Mixed Set leaderboard our team is ranked 1st of 4 finalist teams (Run 1031, Composite Score 0.33; runner-up nadiasaeed at 0.32), but 0.06 below the MEDIQA Organizers' reference submission (0.39). The result that supports a stronger claim is on the Natural Set: our team is *tied for 1st place* with nadiasaeed at Composite 0.58 (our team's best Natural-Set run is 1028); this is +0.10 above the organizer baseline at 0.48 and approximately +0.18 above the strongest published image-grounded RADAR baseline (Gemini-3-Pro at 50 slices, Composite Natural 0.399). The Mixed-vs-Natural gap supports a defensible narrative: text-only frontier-LLM prompting with careful prompt design beats image-grounded baselines on naturalistic clinical edits, but loses on synthetic adversarial edits where image grounding matters more. Statistical analysis confirms the v3 prompt is significantly better than v2 on the development set (McNemar's $p = 0.0003$); this validates the prompt-engineering progression on dev. Run 1028 used the v2 prompt and Run 1031 used the v3 prompt, so the team's two best-per-metric runs use different prompts; we discuss this dev-vs-test divergence in Section 6. We additionally document the dominant failure mode of the v3 prompt on test: a systematic over-prediction of severity = moderate (44.5% predicted vs. 24.5% gold) despite an explicit "resist-default-to-moderate" guardrail. This indicates the guardrail is necessary but not sufficient and motivates a future calibration step. We document the verbatim v3 prompt in Appendix A, the four submitted runs, and post-competition experiments with claude-opus-4-7 and a majority + selective-override hybrid.

Keywords

radiology report review, large language models, prompt engineering, zero-shot classification, severity decoupling, MEDIQA-CORE 2026, CLEF 2026

1. Introduction

Attending radiologists routinely revise preliminary reports drafted by junior residents. Each revision is a tuple of three labels [1, 2]: **agreement** (agree, partially agree, disagree) with fuzzy matching between agree and partially agree at scoring time; **severity** (critical, moderate, negligible); and **edit_type** (addition, correction, clarification). The official composite score is the per-edit conjunction of all three (with the fuzzy match on agreement), reported on a Mixed Set (the union of natural and synthetic edits) and a Natural Set (the un-augmented subset). The leaderboard ranks on Mixed Set Composite.

The development set has 184 edits across 20 cases (case_ids 10008–10047), with the gold label distributions given in Table 1. Edits come from the non-neurologic emergency department at Harborview

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

✉ anubhav27071997@gmail.com (Anubhav)

🆔 0009-0002-4421-0395 (Anubhav)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Medical Center through the University of Washington Clinical Data Repository [2].

Our contributions:

1. A clean text-only zero-shot pipeline at temperature 0 using claude-sonnet-4-6, one API call per edit, with deterministic JSON output and a strict three-key schema. We do not load DICOM, do not perform any image preprocessing, and do not retrieve CT series. The prompt explicitly informs the model that it has no image access.
2. The principal prompt-engineering contribution: an explicit *severity decoupling rule* ("a wrong edit can still be negligible"), a *counterfactual test* ("would the patient's next 24–48 hours of care change if the edit were added or removed?"), and five *golden rules* that override the natural drift toward "moderate" as a safe middle.
3. Empirical evidence that this design ties for 1st place on the Natural Set Composite (0.58, run 1028), +0.10 above the organizer baseline (0.48) and +0.18 above the strongest published image-grounded RADAR baseline (Gemini-3-Pro 50 slices at 0.399), while losing on the Mixed Set (Run 1031 Composite 0.33 vs. organizer 0.39, $\Delta = -0.06$). The Mixed-vs-Natural gap is the central finding.
4. Statistical confirmation that the prompt-engineering progression is meaningful: the v3 prompt is significantly better than the v2 prompt on the development set (McNemar's $b = 35$, $c = 10$, $p = 0.0003$), and significantly better than v1 ($b = 42$, $c = 18$, $p = 0.003$).
5. Identification and quantification of the dominant failure mode: a systematic over-prediction of severity = moderate (44.5% of test predictions vs. 24.5% gold prevalence) despite an explicit guardrail in the prompt. This points the way for a post-competition calibration step.
6. Honest gap analysis to the organizer baseline (agreement -0.16 , severity -0.04 , edit_type -0.14 on Mixed) and a quantified improvement roadmap.

We do not claim the submitted system is state of the art on the harder Mixed Set. Our Mixed Composite of 0.33 (run 1031) is below the organizer baseline at 0.39 by -0.06 . The 1st-of-4 Mixed ranking reflects a small competitive field and a 0.01 margin over the runner-up that is not statistically distinguishable from zero (Section 5). On the Natural Set, however, our team's best run (1028, Composite 0.58) is tied for 1st place with nadiasaeed and is +0.10 above the organizer baseline; this is the more defensible result.

2. Related Work

2.1. The RADAR benchmark and Task 2 baselines

The RADAR benchmark [2] is the public dataset and evaluation harness for this task. The dataset paper reports baselines for Gemini-2.5-Pro, Gemini-3-Pro, and Qwen3.5-plus in four visual-input settings each (10, 20, 50 uniformly sampled slices, and video of all slices, after series selection with GPT-OSS-20B). The published numbers we anchor against are Gemini-3-Pro at 50 slices (best Composite Mixed across all RADAR settings) and Qwen3.5-plus at 10 slices (best Edit Type Mixed): Mixed Composite 0.339, Agreement 0.675, Severity 0.560, Edit Type 0.821 for Gemini-3-Pro 50 slices; Mixed Edit Type 0.839 for Qwen3.5-plus 10 slices. On the Natural Set, Gemini-3-Pro 50 slices reaches Composite 0.399, Agreement 0.627, Severity 0.614, Edit Type 0.843 [2]. We use these as the published-baseline anchors throughout the paper. We report composite scores under the organizers' overview-paper definition [1], with Gemini-3-Pro at 50 slices as the reference image-grounded baseline; this task is Task 2 of the MEDIQA-CORE 2026 shared task at ImageCLEF/CLEF 2026 [5]. The composite score reported in the RADAR dataset paper [2] uses a stricter aggregation that the organizers indicate will be harmonized with the overview definition in a future revision; composite values quoted from the dataset paper should therefore be read with that caveat. The MEDIQA Organizers' submission listed on the leaderboard (Composite Mixed = 0.39, Composite Natural = 0.48) is a separate system whose architecture is not publicly documented at the time of this writing; we treat it as a second comparison anchor and do not assume it is identical to any of the published RADAR baselines.

2.2. LLM-as-judge with structured output

Recent work has shown that frontier LLMs can produce competitive zero-shot judgments on clinical text tasks with carefully engineered prompts and constrained JSON output. The RADAR baselines themselves are LLM-based. Our system differs in three ways: (a) the LLM is claude-sonnet-4-6 at temperature 0, deterministic; (b) the input is strictly text-only, and the prompt explicitly tells the model it does not have access to CT images; and (c) the prompt includes an explicit decoupling rule for severity that we have not seen documented in prior MEDIQA-style LLM-as-judge work, though we have not conducted a systematic literature search and similar structural patterns may exist in the broader medical-NLP prompt-engineering literature.

2.3. Severity decoupling in clinical NLP

The severity decoupling intuition is: *whether the proposed edit is supported by evidence is a separate question from whether including it (correctly or not) would change urgent clinical care*. A wrong edit asserting a benign incidental finding is disagree on agreement but negligible on severity, because the fabrication would not trigger urgent action. The RADAR scoring schema explicitly separates the two, which our prompt design exploits.

3. Approach

3.1. Pipeline overview

For each candidate edit, the pipeline executes one deterministic API call:

1. Read the preliminary report `dev_preliminary_report/<case_id>.txt` (plain text).
2. Read the candidate edit from `dev_edits.json`.
3. Construct the user message using a simple template (Appendix B).
4. Call `claude-sonnet-4-6` through the Anthropic SDK with `max_tokens=1024`, `temperature=0.0`, and our v3 system prompt (Appendix A).
5. Extract the JSON object between the first `{` and the last `}` of the response text.
6. Normalize variants (`partial`, `partially_agree`, `partially-agree` all map to `partially agree`; `severity` and `edit_type` are lowered and stripped).
7. Validate against the allowed label sets. On failure or persistent API error, retry up to 3 times; if all retries fail, fall back to the majority class (`agree`, `negligible`, `addition`).

The pipeline never loads DICOM data, never produces RGB or grayscale CT images, never performs Hounsfield windowing, and never retrieves CT series. The CT volume is present in the underlying dataset but is deliberately not used. This is a design choice (not a resource constraint): we wanted to establish how much of the task can be solved from text alone before adding image-grounded complexity.

3.2. The v3 system prompt: severity decoupling and the counterfactual test

The prompt (Appendix A) is structured in three numbered sections, one per label. The severity section embodies three principles:

Decoupling rule. Severity is independent of agreement. We instruct the model that a wrong edit can still be negligible (fabricating a benign incidental cyst → disagree but negligible). We also instruct the converse: a correct edit that elaborates an already-flagged emergency is agree but negligible because the clinical action is already in motion.

Counterfactual test. The model imagines the edit being added and then removed, and asks whether the patient’s next 24–48 hours of care would change. No change → negligible; slow outpatient follow-up → moderate; immediate ICU/OR/medication change → critical.

Resist-default-to-moderate guardrail. Because 51.1% of dev-set edits are gold-labeled negligible, models that drift toward "moderate" pay a heavy cost. We instruct: "If you find yourself drifting to moderate as a safe middle, STOP and force a choice between negligible and critical based on the counterfactual test." Section 5 demonstrates that this guardrail is partially effective on dev but fails substantially on test. This motivates a future post-hoc calibration step.

Five Golden Rules. The prompt closes the severity section with five concise rules: (1) clarification of a known emergency → negligible; (2) wrong addition of benign incidental → negligible; (3) correction of measurement only → negligible; (4) correction of character (benign↔neoplastic) → moderate; (5) new emergency not already in original → critical.

3.3. Model selection: claude-sonnet-4-6

We selected claude-sonnet-4-6 at temperature 0 for the submitted run on cost-and-determinism grounds: Sonnet 4.6 is approximately $5\times$ cheaper than Opus 4.7 per token, and accepts the temperature parameter (Opus 4.7 rejects it). The prompt-iteration ablation (Section 5, Table 7) shows that the principal driver of improvement was the prompt redesign, not the model. A post-competition Opus 4.7 variant (Section 7) does not consistently outperform Sonnet 4.6, supporting the conclusion that prompt design is the binding constraint at this scale.

3.4. Design exclusions

For replicability we list the choices we did not make:

- **No image grounding.** The CT volume was available in DICOM but not used.
- **No few-shot exemplars.** The prompt is zero-shot.
- **No ensembling.** A single deterministic call per edit.
- **No fine-tuning or PEFT.** The model is used as-is.
- **No retrieval-augmented context.** No RAG over similar past edits.

All five would plausibly improve the result; we discuss expected effect sizes in Section 6.

4. Experimental Setup

4.1. Dataset

Table 1

RADAR Task 2 development-set statistics, gold-labeled. Verified from the released `groundtruth_dev.csv`.

Item	Value
Total edits	184
Unique cases	20 (case_id 10008–10047)
Gold label distributions	
Agreement: agree / disagree / partially agree	88 / 85 / 11 (47.8% / 46.2% / 6.0%)
Severity: negligible / moderate / critical	94 / 45 / 45 (51.1% / 24.5% / 24.5%)
Edit_type: addition / correction / clarification	108 / 42 / 34 (58.7% / 22.8% / 18.5%)

The test split has 274 edits across roughly 30 cases (no released ground truth).

4.2. Implementation and hardware

Python 3.10, Anthropic SDK, pandas. Single CPU is sufficient for inference. All API calls were cached to disk to guarantee exact reproducibility of the submitted run. The submitted run consumed approximately \$1.85 in API costs across 274 test edits, dominated by input-token usage on the 4.2 KB system prompt

(input-token cost roughly \$0.005 per edit at the published Sonnet 4.6 rates as of May 2026). The dev iterations across the three prompt versions consumed an additional \$1.30 over 184 dev edits each.

4.3. Determinism at temperature 0

The Anthropic API at temperature=0 is approximately but not strictly deterministic; residual non-determinism from backend routing and hardware can cause occasional disagreements across calls with identical inputs. We did not empirically quantify this on RADAR (each edit was run once for the submitted run). The bootstrap confidence interval we report in Section 5 therefore captures sampling variation across edits within a single run but not between-run variance at fixed temperature; a re-run of the v3 dev predictions to measure between-run disagreement is a recommended check for downstream replication.

4.4. The four submitted runs

We submitted four runs to CodaBench during the evaluation window. All four appear in the official All-Runs sheet released by the organizers; the team-level ranking sheets use the best-of-team run per metric. The four runs differ only in the system prompt; the model (`claude-sonnet-4-6`), temperature (0.0), `max_tokens` (1024), and single-call zero-shot pipeline are identical across all four.

Table 2

The four submitted runs (all timestamps UTC, 2026). All four use `claude-sonnet-4-6` at temperature 0 and differ only in the system prompt. “Mix” and “Nat” are the official Composite scores on the Mixed and Natural test sets.

Run	Timestamp	Prompt	Notes	Mix	Nat
1027	2026-05-05 23:29	v1	No severity-decoupling, no counterfactual test	0.20	0.22
1028	2026-05-05 23:33	v2	Decoupling rule and golden rules; no counterfactual test	0.32	0.58
1030	2026-05-06 00:05	v3 proto	Early v3 without all final patches	0.24	0.34
1031	2026-05-06 00:08	v3 final	The prompt in Appendix A	0.33	0.55

The organizer-selected per-metric best runs for our team are Run 1031 (Mixed Composite 0.33, ranking 1st of 4 teams) and Run 1028 (Natural Composite 0.58, tied for 1st with *nadiasaeed*). The team’s two best-per-metric runs use *different* prompts: v3 wins on Mixed by 0.01, but v2 wins on Natural by 0.03. The principal prompt-engineering contribution we describe in this paper (Section 3) is the v3 prompt that produced Run 1031, since v3 is significantly better than v2 on the development set (McNemar’s $p = 0.0003$) and on the Mixed Set; the test Natural Set divergence is one of the dev-vs-test findings discussed in Section 6.

4.5. Metrics

We use the official RADAR `eval.py` script [2], which computes `agreement_accuracy`, `severity_accuracy`, `edit_type_accuracy` (all exact-match with the fuzzy agree $\equiv$ partially agree rule for the agreement label), and `composite_score` (per-edit conjunction). Reported on Mixed Set (ranking metric) and Natural Set (diagnostic).

5. Results

5.1. Headline: Mixed Set (the ranking metric)

Two observations on the comparison anchors:

Table 3

MEDIQA-CORE 2026 Task 2 Mixed-Set comparison. **Bold** is our submission. Top block: official 2026 leaderboard (team field + organizers’ reference). Bottom block: published RADAR baselines [2] for context; these are computed on the same Mixed Set so the comparison is direct.

Rank	Team / System	Composite	Agreement	Severity	Edit Type
–	MEDIQA Organizers (leaderboard baseline)	0.39	0.68	0.56	0.82
1	Brihadaranyaka (1031, ours)	0.33	0.52	0.52	0.68
2	nadiasaeed	0.32	0.43	0.51	0.70
3	ClinicalVLM-TW	0.22	0.53	0.38	0.63
4	krishnatewari	0.05	0.22	0.51	0.12
<i>Published RADAR baselines (Sun et al. [2], with image grounding)</i>					
–	Gemini-3-Pro, 50 slices	0.339	0.675	0.560	0.821
–	Gemini-3-Pro, 20 slices	0.321	0.693	0.544	0.781
–	Gemini-3-Pro, 10 slices	0.321	0.686	0.536	0.803
–	Qwen3.5-plus, 10 slices	0.223	0.558	0.533	0.839

- Our Composite Mixed of 0.33 is within rounding (-0.009) of Gemini-3-Pro at 50 slices (0.339), the strongest published image-grounded baseline. Our text-only Sonnet 4.6 system matches the strongest published image-grounded result on the composite metric. We trail Gemini-3-Pro 50 slices on all three component metrics (Agreement 0.52 vs. 0.675, Severity 0.52 vs. 0.560, Edit Type 0.68 vs. 0.821), so the composite-match arises from a correlation pattern across components rather than uniform per-component matching.
- Our Composite Mixed of 0.33 is -0.06 below the MEDIQA Organizers’ leaderboard baseline at 0.39. This baseline is not the published Gemini-3-Pro 50 slices system; its architecture is not publicly documented at the time of writing. The two anchors yield very different deficit values (-0.06 vs. -0.009), and the paper distinguishes them throughout.

5.2. Statistical significance of the 0.01 margin over nadiasaeed

On 184 edits with binary per-edit scoring (the structural ceiling on test, given 274 test edits we expect the CI to be narrower than dev), the bootstrap 95% CI half-width on the v3 development composite is 0.068 (Table 4). A first-order $\sqrt{n_1/n_2}$ rescaling for the 274-edit test would give a half-width of approximately 0.056; this rescaling assumes i.i.d. sampling and identical underlying distributions between dev and test, neither of which is verified. The actual test CI cannot be computed without test labels. The 0.01 margin over nadiasaeed is approximately 0.18σ under the rescaled approximation and is therefore *not statistically distinguishable from zero*. We do not claim a meaningful advantage over nadiasaeed; the 1st-of-4 ranking is best understood as "tied for first in a four-team field."

Table 4

Bootstrap 95% CI on v3 development composite (2000 resamples).

System	Composite (point)	CI lower	CI upper	Half-width
v3 dev predictions	0.3370	0.2663	0.4023	0.0680

5.3. Diagnostic: Natural Set (un-augmented edits)

Three findings of interest:

- **Our team’s Natural-Set Composite of 0.58 ties for 1st place with nadiasaeed** and beats every published image-grounded RADAR baseline. The strongest of those is Gemini-3-Pro 50 slices at 0.399; our 0.58 is approximately $+0.18$ above that. The text-only system out-scores image-grounded systems by a substantial margin on the Natural Set. This is the strongest defensible claim in the paper.

Table 5

Task 2 Natural-Set diagnostic. The Natural Set is the un-augmented subset of the test corpus; its exact size on test is not stated in the official leaderboard release but is consistent with approximately 150 edits based on figures reported in the RADAR dataset paper [2]. Top block: 2026 leaderboard, with the organizer-selected per-team best-per-metric run for the Natural Set. Bottom block: published RADAR baselines for context.

Team / System	Comp Nat	Agree Nat	Sev Nat	EdT Nat
Brihadaranyaka, Run 1028, tied 1st on Natural	0.58	0.86	0.76	0.72
nadiasaeed, Run 1043, tied 1st on Natural	0.58	0.76	0.76	0.72
MEDIQA Organizers (leaderboard baseline)	0.48	0.63	0.61	0.84
Brihadaranyaka, Run 1031 (our Mixed-Set best, for context)	0.55	0.82	0.72	0.71
ClinicalVLM-TW, Run 1037	0.29	0.49	0.46	0.65
krishnatewari, Run 960	0.08	0.24	0.76	0.14
<i>Published RADAR baselines (Sun et al. [2])</i>				
Gemini-3-Pro, 50 slices	0.399	0.627	0.614	0.843
Gemini-3-Pro, video	0.353	0.575	0.647	0.830
Gemini-2.5-Pro, 50 slices	0.386	0.680	0.562	0.850
Qwen3.5-plus, 10 slices	0.216	0.418	0.601	0.869

- Our Composite Natural (0.58) is +0.10 above the MEDIQA Organizers’ leaderboard baseline (0.48). We do not have access to per-edit predictions from either the organizers or the published RADAR baseline systems, so we cannot compute a per-edit McNemar’s test against them. On the inferred Natural Set size of $n \approx 150$ binary outcomes, a difference of +0.10 corresponds to roughly 15 additional correct composite-matches; the standard error of a single-proportion estimate at this scale is ≈ 0.04 , so +0.10 is approximately 2.5σ . This is suggestive of a real difference but not definitively distinguishable from sampling variation without paired per-edit comparisons. We flag this and recommend the organizers release per-edit baseline predictions for proper statistical comparison.
- Run 1028 (v2 prompt) outperforms Run 1031 (v3 prompt) on the Natural Set by 0.03 Composite (0.58 vs. 0.55), but Run 1031 outperforms Run 1028 on the Mixed Set by 0.01 Composite (0.33 vs. 0.32). The v3-over-v2 dev advantage (McNemar’s $p = 0.0003$ on dev, Section 5) therefore transfers to Mixed but *reverses* on Natural. We interpret this in Section 6: the additional rules in v3 (severity decoupling, counterfactual test, additional examples) help on synthetic adversarial edits but appear to introduce moderate-class drift that costs the system on the agree-heavy Natural distribution.

5.4. Mixed minus Natural: a coherent narrative

Table 6

Mixed minus Natural deltas. Positive means relatively better on synthetic edits; negative means relatively better on natural.

Team / run	Comp Mixed	Comp Natural	Mixed – Natural
MEDIQA Organizers (baseline)	0.39	0.48	−0.09
Brihadaranyaka (1031)	0.33	0.55	−0.22
nadiasaeed (1043)	0.32	0.58	−0.26

All three systems do better on natural than on mixed. The organizer baseline takes a −0.09 hit; our system takes a −0.22 hit; nadiasaeed takes a −0.26 hit. The organizer baseline is the most resistant to synthetic adversarial edits, plausibly because it uses image grounding to verify edits against ground-truth CT findings; we say "plausibly" because the organizer baseline’s architecture is not publicly documented at the time of this writing, so its actual mechanism for synthetic-edit resistance is inferred rather than observed.

5.5. Prompt iteration ablation on dev

Table 7

Prompt iteration progression on the development set (184 edits). All three versions use claude-sonnet-4-6, temperature 0, max_tokens 1024.

Prompt	Composite	Agreement	Severity	Edit Type
v1 (baseline)	0.207	0.582	0.429	0.696
v2 (mid)	0.201	0.696	0.429	0.712
v3 (submitted)	0.337	0.728	0.603	0.734
$\Delta(\mathbf{v3} - \mathbf{v1})$	+0.130	+0.146	+0.174	+0.038

The v1→v2 transition lifts agreement substantially (+0.114) but leaves severity unchanged (0.429) and barely moves composite (−0.005). The v2→v3 transition is where the principal contribution lives: severity jumps 0.429 → 0.603 (+0.174), the largest single-step improvement on any metric in any transition, and composite jumps 0.201 → 0.337 (+0.136).

5.6. Statistical significance of the prompt progression

Table 8

McNemar’s test on per-edit composite-correctness between consecutive prompt iterations on dev. *b* = first prompt wrong but second prompt right; *c* = first prompt right but second prompt wrong.

Comparison	<i>b</i>	<i>c</i>	<i>p</i> -value	Significant at $\alpha=0.05$
v1 vs v2	14	15	1.0000	NS
v2 vs v3	35	10	0.0003	SIG
v1 vs v3	42	18	0.0030	SIG

The v1→v2 transition is not statistically significant on dev. The v2→v3 transition is highly significant ($p = 0.0003$, with 35 edits improved and only 10 regressed, a net gain of 25 edits out of 184). The v1→v3 transition is also significant. *This is the central methodological claim of the paper: the severity-decoupling rule, counterfactual test, and golden rules introduced in v3 are responsible for a statistically significant improvement over the simpler v2 prompt.* We emphasize that this significance is established on the same dev set on which the prompts were designed; without test ground truth we cannot test prompt-progression significance on the test set, and the dev-set significance is necessary but not sufficient evidence of test-set robustness.

5.7. Failure-mode analysis (v3 on dev)

Table 9

Stratified composite on dev (v3 predictions), broken down by gold-label stratum.

Stratum	Class	n	Composite	Agreement	Severity	Edit Type
edit_type	addition	108	0.352	0.759	0.639	0.778
edit_type	correction	42	0.333	0.714	0.524	0.857
edit_type	clarification	34	0.294	0.647	0.588	0.441
severity	negligible	94	0.372	0.723	0.553	0.798
severity	critical	45	0.356	0.733	0.711	0.711
severity	moderate	45	0.244	0.733	0.600	0.622
agreement	partially agree	11	0.545	0.636	0.727	0.909
agreement	agree	88	0.375	0.659	0.580	0.807
agreement	disagree	85	0.271	0.812	0.612	0.635

Three weakest strata, bolded above:

- **edit_type = clarification** (composite 0.294): the model recognizes that the topic exists in the report only 44.1% of the time when the gold label is clarification.
- **severity = moderate** (composite 0.244): the weakest of all 9 strata. The model recovers moderate severity correctly only 60% of the time when the gold label is moderate.
- **agreement = disagree** (composite 0.271): although the agreement-accuracy on this stratum is the highest of the three (0.812), edit_type accuracy drops to 0.635; the model gets agreement right but conflates the edit type, dragging composite down.

The partially agree row of Table 9 ($n = 11$) shows a per-edit Composite of 0.545 but the small- n uncertainty is large: a one-sided Wilson 95% CI on 6/11 correct is approximately [0.28, 0.79], half-width ≈ 0.26 . We report this row for transparency but do not draw structural conclusions from the 0.545 point estimate.

5.8. Confusion matrices (v3 on dev)

Table 10

Confusion matrices on dev for v3 predictions. Rows are gold; columns are predicted.

Agreement	agree	part. agree	disagree	Severity	critical	moderate	negligible
agree	56	2	30	critical	32	12	1
partially agree	5	2	4	moderate	10	27	8
disagree	13	3	69	negligible	4	38	52
				Edit Type	addition	correction	clarification
				addition	84	23	1
				correction	5	36	1
				clarification	4	15	15

Three failure patterns stand out:

- **Agreement: 30 cases where gold is agree but the model predicted disagree**, vs. only 13 in the reverse direction. The model is biased toward predicting disagreement.
- **Severity: 38 cases where gold is negligible but the model predicted moderate**, vs. only 12 critical→moderate. The "resist-default-to-moderate" guardrail is partially failing; the model still drifts to moderate from the negligible side.
- **Edit_type: 19 of 34 clarification gold cases are mispredicted** (15 as correction, 4 as addition). The model does not yet reliably detect when an edit's topic is already present in the report.

5.9. Test-set predicted class distribution

Table 11

Predicted vs. expected (gold-prevalence) distribution for run 1031 on test (274 edits). "Expected" is the dev gold prevalence from Table 1.

Label	Class	Predicted (%)	Expected (%)
agreement	agree	41.6	47.8
agreement	disagree	55.8	46.2
agreement	partially agree	2.6	6.0
severity	negligible	32.8	51.1
severity	moderate	44.5	24.5
severity	critical	22.6	24.5
edit_type	addition	63.9	58.7
edit_type	correction	30.3	22.8
edit_type	clarification	5.8	18.5

The dominant test-set anomaly is severity: 44.5% of test predictions are moderate, against a gold prevalence of 24.5%. The model massively over-predicts moderate on test, despite the explicit "resist-default-to-moderate" guardrail in the prompt; the guardrail is only partially effective. `Edit_type` shows a complementary anomaly: only 5.8% of test predictions are `clarification` against a gold prevalence of 18.5%, consistent with the dev-set finding that the model often mis-detects topic presence in the source report.

Natural-Set distribution limitation. The distribution in Table 11 is computed over all 274 test edits (Mixed Set). The RADAR dataset paper [2] notes that Natural-Set Agreement is dominated by the agree class and that high Natural Agreement scores can therefore be inflated by class imbalance alone. To distinguish "the prompt genuinely discriminates on natural edits" from "the prompt's residual bias happens to align with the agree-heavy Natural-Set distribution," a Natural-Set-only version of Table 11 is needed. The natural-vs-synthetic indicator for individual test edits is not in the materials released to participants. We therefore cannot compute the Natural-Set-only predicted distribution from the materials we have. The release of natural-vs-synthetic per-edit labels (or per-edit baseline predictions) by the organizers would directly enable this analysis.

Clarification under-prediction hypothesis. The most striking `edit_type` anomaly on test is the under-prediction of `clarification` (5.8% predicted vs. 18.5% gold prevalence). One concrete hypothesis: the v3 system prompt foregrounds the addition-vs-correction distinction with three worked examples in Section (1), while the clarification class receives only one inline example. The decision rule "Topic present, no contradiction $\rightarrow$ clarification" requires the model to (a) detect topic presence in the report and (b) judge non-contradiction; failure on either step routes the edit to addition or correction. The dev-set stratification (Table 9) confirms this: clarification has the lowest `edit_type` accuracy (0.441) of the three classes. Direct tests of this hypothesis would include: (i) inverting the example-coverage ratio in the prompt to give clarification three worked examples instead of one, and (ii) measuring whether topic-presence detection on report text correlates with model agreement on clarification gold labels. We did not run these experiments in the submitted system and flag them as targeted next-step probes.

5.10. Per-edit failure mode summary on dev

Table 12

Decomposition of v3-dev predictions by how many of the three labels match gold.

Failure mode	n	%
All three labels correct (composite = 1)	62	33.7
Partial (1 or 2 labels correct)	114	62.0
All three labels wrong	8	4.3
Total	184	100.0

Only 4.3% of dev edits have *all three* labels wrong; the model is rarely catastrophically incorrect. The challenge is the 62.0% partial-correctness band, with one or two labels right but the third dragging composite to zero. This is consistent with composite being a strict conjunctive metric.

5.11. Test-set divergence: v3 vs. v4-Opus vs. hybrid vs. majority

Two findings: v3 (Sonnet 4.6) and v4-Opus agree on `edit_type` 91.6% of the time, but disagree on severity 36.9% of the time. The two models share a strong text-pattern signal for `edit_type` (a near-mechanical detection of topic-in-report) but diverge sharply on the harder severity judgment. The hybrid approach injects a substantial dose of the always-majority baseline into the mix, leading to large divergence from v3 on all three labels.

Table 13

Pairwise prediction disagreement on the 274-edit test set. v3 = submitted run 1031 (Sonnet 4.6). v4-Opus = post-competition Opus 4.7 with majority-biased prompt. Hybrid = majority + selective v4-Opus overrides. Majority = always-(agree, negligible, addition).

A	B	Agree differ	Severity differ	Edit Type differ
v3	v4-Opus	19.3%	36.9%	8.4%
v3	hybrid	48.9%	56.9%	29.9%
v3	majority	58.4%	67.2%	36.1%
v4-Opus	hybrid	36.9%	36.5%	28.8%

6. Analysis and Discussion

6.1. The Natural-Set advantage

The most direct explanation for the Natural-Set advantage is the prompt’s explicit severity-decoupling rule. Natural edits, generated by actual attending radiologists, follow recognizable clinical reasoning patterns. The decoupling rule allows the model to (a) accept benign additions as agree but mark them negligible, and (b) mark clarifications of already-flagged emergencies as negligible. These two patterns are common in natural edits.

The counterfactual test is the second mechanism. Severity is the metric most subject to "moderate" drift; the counterfactual test forces a binary outcome (urgent care changes or it doesn't) and pushes the model toward the negligible/critical extremes that match the gold distribution. The dev-set stratification (Table 9) shows that severity accuracy is highest on the `critical` gold stratum (71.1%) and on the `agreement = partially agree` stratum (72.7%); the counterfactual test differentiates most clearly on these two strata.

6.2. The Mixed-Set gap

The dominant gap is on agreement (−0.16 vs. baseline). The agreement metric most directly tests whether the proposed edit is *supported by the underlying clinical evidence*. Our pipeline has no access to that evidence beyond the report text. The organizer baseline likely uses the CT volume to verify edits against image findings, which is the right tool for this label.

The `edit_type` gap (−0.14) is harder to explain since `edit_type` is text-only by construction. The dev-set evidence (Table 10) suggests two contributing causes: (a) confusion between `clarification` and `correction` when the report contains a near-synonym for the edit’s topic (15 of 34 gold clarifications mispredicted as correction); (b) the model’s relative reluctance to commit to clarification, evident in the test-set distribution where only 5.8% of predictions are clarification vs. 18.5% expected.

6.3. The "moderate" guardrail failure

The most striking failure is severity over-prediction of moderate: 44.5% predicted vs. 24.5% gold on test, with the same pattern visible on dev (Table 10: 38 negligible→moderate, 12 critical→moderate). Despite the explicit "resist-default-to-moderate" instruction in the prompt, the model still drifts to moderate as a safe middle.

We hypothesize three contributing causes:

- LLMs trained with safety-oriented RLHF have a structural bias toward hedging on consequential judgments. The "moderate" class encodes precisely this hedge.
- Reasoning chain-of-thought is not visible to the JSON output; the prompt asks for "reasoning" but does not validate that the reasoning constrains the final label.
- The dev set is small enough (184 edits) that the model’s prior on moderate may not have been corrected by exposure to the actual class distribution during prompt iteration.

A post-competition fix is a calibration step on the severity output. Since the Anthropic API does not currently expose per-token output logprobs for claude-sonnet-4-6 in our message-completion mode, the implementable version is a post-hoc classifier on the model’s string output that re-weights the severity prediction to match observed prevalence on a held-out set. The relevant re-weighting parameters are tunable on dev to match the dev gold prevalence and would be applied at inference time as a one-line post-processing step.

6.4. Dev–test divergence: v3 wins dev and Mixed, v2 wins Natural

On Mixed, run 1031 (v3 prompt) beats run 1028 (v2 prompt) by 0.01 (0.33 vs. 0.32). On Natural, run 1028 beats run 1031 by 0.03 (0.58 vs. 0.55). On dev, v3 beats v2 by 0.136 Composite (0.337 vs. 0.201, McNemar’s $p = 0.0003$). Both runs are part of our team’s official submission; the organizers select best-per-metric for the team-level ranking, so this is not a “wrong run submitted” situation but a dev-vs-test divergence finding.

The pattern is consistent with the moderate-class drift the v3 prompt introduces. On the Mixed Set, synthetic adversarial edits contain many fabricated findings, and v3’s additional rules (severity decoupling, counterfactual test, additional examples) help the model classify these correctly; the moderate drift is offset by gains on critical and disagree. On the Natural Set, the gold severity distribution is more skewed toward negligible, and v3’s moderate drift directly hurts: the simpler v2 prompt, without the explicit counterfactual test, predicts severity closer to the natural prevalence. We do not have test gold to confirm this hypothesis directly, but the dev-vs-test divergence is consistent with v3’s failure mode being moderate over-prediction (Section 5, Table 11). The practical implication: prompt engineering wins on dev do not necessarily transfer to test subsets with different label-prevalence patterns, and a held-out estimator for the deployment distribution would be useful before settling on a final prompt.

6.5. Threat to validity: possible benchmark contamination

Because our entire system is a single frontier-LLM API call with a hand-written prompt, training-data exposure is a structural limitation rather than an experimental result, and we present it as such. The RADAR dataset paper [2] has been public on arXiv since early 2026, and the benchmark and its companion code are openly available; claude-sonnet-4-6’s training-data cutoff is not disclosed to us in enough detail to rule out exposure to the RADAR paper, its code repository, the closely related radiology-report-error literature, or similar radiology-error corpora. We did not run direct contamination probing (for example, prompting the model to recall RADAR’s design choices, label definitions, or benchmark numbers), and we flag such probing as an essential follow-up before drawing strong claims about the prompt’s generalization to truly held-out clinical text. We raise this as a hypothesis about a possible confound, not as a measured effect.

6.6. What would close the Mixed-Set gap

A credible post-competition path (we mark each direction as conjectural; the lift ranges are anchored to qualitative reasoning rather than to published evidence on this specific benchmark):

1. **Image grounding through RadFM, CT-CLIP, or Merlin.** Replace the no-CT input with a 3D CT encoder’s pooled feature vector concatenated to the prompt (or used in a two-stage agreement gate). Direction: the published RADAR baselines [2] show that image grounding helps Agreement more than the other two components, so this is the direction with the largest potential to close the -0.16 Agreement gap.
2. **Severity calibration.** The first-pass design we considered was a logit-bias step on the severity output; this assumes the Anthropic API exposes per-token output logprobs, which is not currently the case for claude-sonnet-4-6 in the message-completion mode we use. A viable alternative is a post-hoc classifier on the model’s string output that re-weights the severity prediction to match

observed prevalence on a held-out set. Direction: expected to help on severity, directly addressing the moderate over-prediction documented in Section 5.

3. **Small text-only classifier for edit_type.** Fine-tune Bio_ClinicalBERT [3] or RadBERT [4] on the 184 dev (report, edit) pairs as a 3-way classifier. We did not evaluate these on edit_type in the submitted system; we cite them here as the two strongest published candidates rather than as established components of our pipeline. Direction: expected to help on edit_type, particularly on the clarification class where our test under-prediction is 5.8% vs. 18.5% gold prevalence.
4. **Prompt ensembling.** Three prompts (v3, v4-Opus, a third yet-unexplored variant), majority vote per label. Direction: expected to help on Composite by reducing per-prompt failure modes, with a lift ceiling bounded by the agreement rate of the three component prompts.

We do not assign numerical lift estimates to these directions because the underlying empirical evidence on this specific benchmark is insufficient to anchor specific component F1 deltas.

7. Post-competition experiments

7.1. Opus 4.7 majority-biased detector (v4)

After the May 6 deadline, we re-implemented the pipeline with claude-opus-4-7 and a rewritten system prompt that emphasizes the majority class. The post-competition v4 system explicitly distinguishes natural vs. synthetic edits in its preamble and provides observed natural-test class proportions to calibrate the model’s prior; the key change from the v3 prompt is reproduced in Appendix C. On test, v4-Opus disagrees with v3 on 19.3% of agreement predictions, 36.9% of severity predictions, and only 8.4% of edit_type predictions (Table 13). The substantial severity disagreement and small edit_type disagreement together suggest v4-Opus is a meaningfully different severity classifier on the same input.

7.2. Selective-override hybrid (majority + v4 critical-disagree overrides)

A second post-competition variant combines an always-majority baseline with v4-Opus predictions, accepting an override only when v4 predicts ALL three of: `severity = critical` AND `agreement = disagree` AND `edit_type ∈ {addition, correction}`. On test, the hybrid disagrees with v3 on 48.9% of agreement predictions, 56.9% of severity predictions, and 29.9% of edit_type predictions (Table 13). The large divergence reflects that the hybrid effectively defaults to the always-majority baseline outside the narrow critical-disagree gate.

Without test ground truth we cannot evaluate v4-Opus or hybrid against the official scoring. Both variants are prepared as candidates for a CodaBench Phase 2 re-submission and would be evaluated then.

8. Conclusion

We presented Brihadaranyaka’s submission to MEDIQA-CORE 2026 Task 2: a text-only zero-shot pipeline using claude-sonnet-4-6 at temperature 0, with the principal contribution being the prompt design: an explicit severity-decoupling rule, a counterfactual test, and five golden rules. The team submitted four runs; on the official leaderboard, run 1031 (v3 prompt) is ranked 1st of 4 finalist teams on Mixed Composite (0.33, −0.06 below the organizer baseline at 0.39), and run 1028 (v2 prompt) is tied for 1st with nadiasaeed on Natural Composite (0.58, +0.10 above the organizer baseline at 0.48 and +0.18 above the strongest published image-grounded RADAR baseline). The Mixed-vs-Natural gap supports the central claim: *text-only frontier-LLM prompting with careful prompt design beats image-grounded baselines on naturalistic clinical edits, but loses on synthetic adversarial edits where image grounding matters more*. We additionally documented two methodologically interesting findings: (a) the v3 prompt is significantly better than v2 on dev (McNemar’s $p = 0.0003$) and wins on test Mixed by 0.01, but *loses* to v2 on test Natural by 0.03, indicating that prompt-engineering wins on dev do not always transfer

to all test subsets; and (b) a systematic over-prediction of severity = moderate (44.5% predicted vs. 24.5% gold on Mixed) despite an explicit prompt guardrail. This motivates a logit-bias calibration step as the principal post-competition improvement. The full path to closing the Mixed-Set gap to baseline runs through image grounding using a 3D CT encoder, severity calibration, and a small fine-tuned text classifier for edit_type, all feasible for a CodaBench Phase 2 re-submission.

Declaration on Generative AI

During the preparation of this work, the author used claude-sonnet-4-6 (Anthropic, accessed through the API) as the inference engine for the system described in Section 3. The author additionally evaluated claude-opus-4-7 post-competition for the experiments described in Section 7. The author also used Claude (chat interface, Opus 4.7) for paraphrase and grammar polish in the writing of this paper. All scientific claims, experimental design, code, and analyses are the author’s own. No part of the paper was AI-generated without author review and correction.

Acknowledgments

We thank the MEDIQA-CORE 2026 organizers (Asma Ben Abacha, Zhaoyi Sun, Wen-wai Yim, Fei Xia, and Meliha Yetisgen) for releasing the RADAR benchmark and running the shared task. We thank Anthropic for API access; all compute and API spend was funded entirely by the author.

References

- [1] Asma Ben Abacha, Zhaoyi Sun, Wen-wai Yim, Fei Xia, and Meliha Yetisgen. Overview of the MEDIQA-CORE 2026 Task 2 on Radiology Report Discrepancy Assessment. In CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 2026.
- [2] Zhaoyi Sun, Minal Jagtiani, Wen-wai Yim, Fei Xia, Martin Gunn, Meliha Yetisgen, and Asma Ben Abacha. RADAR: A Multimodal Benchmark for 3D Image-Based Radiology Report Review. arXiv:2603.06681, 2026.
- [3] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew B. A. McDermott. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (NAACL 2019), pp. 72–78, 2019.
- [4] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-based Language Models to Radiology. Radiology: Artificial Intelligence, 4(4):e210258, 2022. DOI: 10.1148/ryai.210258.
- [5] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, L.-D. Ștefan, M.-G. Constantin, M. Dogariu, A.-G. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, et al. Overview of ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security. In CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 2026.

A. Verbatim claude-sonnet-4-6 system prompt used in Run 1031

The following is the system prompt used for every API call in Run 1031, transcribed verbatim from the production notebook. Section-divider lines appear here as runs of = characters.

```
You are an attending radiologist reviewing proposed edits to preliminary CT reports written
by junior residents. The reviewer is a senior expert; their edits are typically
reasonable. About half of all proposed edits are accepted with minor or no objections;
the other half contain disagreements ranging from minor to clinically significant.
```

You see only the report TEXT and the proposed edit -- you do NOT have CT images. Classify each edit on three dimensions, IN THIS ORDER:

=====

(1) edit_type

- * "clarification" (~18%): Topic ALREADY MENTIONED in report; edit refines/elaborates without contradicting.
- * "addition" (~59%): Topic NOT mentioned anywhere in report.
- * "correction" (~23%): Topic in report but edit changes a SPECIFIC existing statement (measurement, location, character, or interpretation).

Decision rule: scan report for the edit's topic.

- Topic present, no contradiction -> clarification
- Topic present, edit changes the existing statement -> correction
- Topic absent -> addition

Examples:

Report says "bilateral pleural effusions". Edit elaborates them -> CLARIFICATION.

Report says nothing about kidneys. Edit describes a renal mass -> ADDITION.

Report says "small renal cyst". Edit calls it a 4cm enhancing mass -> CORRECTION.

=====

(2) agreement

- * "agree" (~48%): Edit is correct, or a reasonable refinement; not contradicted by the report.
- * "partially agree" (~6%): RARE. Both clearly correct AND clearly incorrect parts.
- * "disagree" (~46%): Edit contradicts the report, OR fabricates a highly specific unfounded claim.

Defaults:

- Clarifications of existing findings -> almost always AGREE
- Corrections that match clinical context -> AGREE; that contradict reasonable original statements -> DISAGREE
- Plausible additions not contradicted -> can AGREE
- Additions making very specific unsupported claims (precise sizes/locations/diagnoses) -> DISAGREE
- Critical: ABSENCE of mention != contradiction. Don't disagree just because topic isn't in report.

=====

(3) severity -- read carefully, this is the most-misunderstood label

DECOUPLING RULE: Severity is INDEPENDENT of agreement.

A wrong edit can still be negligible. An edit that fabricates a benign incidental finding is wrong (disagree) but still NEGLIGIBLE -- the fabrication wouldn't change urgent care. Score severity as if asking: "if this edit ENTERED the final report (correctly or not), would urgent management change?"

COUNTERFACTUAL TEST (apply to every edit):

Imagine the edit is added to the report. Then imagine it is REMOVED.

Does anything in the patient's next 24-48 hours of care change?

- No -> negligible
- Slow / outpatient follow-up changes -> moderate
- Immediate ICU/OR/medication change -> critical

RESIST DEFAULT-TO-MODERATE: 51% of edits are negligible. If you find yourself drifting to "moderate" as a safe middle, STOP and force a choice between negligible and critical based on the counterfactual test.

- * "negligible" (~51%) -- STRONG DEFAULT. Choose whenever ANY apply:

- Edit is a CLARIFICATION -> almost always negligible (the underlying finding is already in the report; clinicians already know).

- Edit refers to incidental/benign findings: simple cysts, fatty liver, splenomegaly, mild atherosclerosis (in non-vascular indications), bone islands, calcified granulomas, hemorrhoids, mild degenerative changes, hemangiomas.
 - Edit is wrong but the wrongness wouldn't change next-step care (e.g., wrong addition of a benign cyst).
 - Edit refines a measurement of a non-urgent finding (size of cyst off by 2mm).
- * "moderate" (~24%) -- choose ONLY when:
- Edit alters the assessment of a clinically-relevant but non-emergent finding (mass needing follow-up, lymph node assessment for staging, new chronic disease finding, organ size change with implications).
 - Wrong description would lead to incorrect outpatient workup or unnecessary procedures.
 - CORRECTION sub-rule: corrections of measurements only -> usually negligible. Corrections that change a finding's CHARACTER (benign<->neoplastic, acute<->chronic, simple<->complex) -> usually moderate.
- * "critical" (~24%) -- choose ONLY when ALL true:
- Topic is genuinely emergent or staging-defining: pulmonary embolism, hemorrhage, bowel obstruction/ischemia/perforation, abscess, sepsis source, missed/wrongly-sized cancer, acute aortic disease, ICU-relevant finding.
 - Original report does NOT already flag this emergency.
 - Counterfactual test shows urgent action changes.
 - CONTEXT MATTERS: aortic disease in an aortic-pain indication may be critical even if "incidental"-looking. Match severity to the clinical question.

GOLDEN RULES (memorize):

1. Clarification of known emergency -> NEGLIGIBLE (action already in motion).
2. Wrong addition of benign incidental -> NEGLIGIBLE (decoupled from agreement).
3. Correction of measurement only -> NEGLIGIBLE.
4. Correction of character (benign<->neoplastic) -> MODERATE.
5. New emergency NOT already in original -> CRITICAL.

EXAMPLES:

NEGLIGIBLE:

- "Moderate to large bilateral pleural effusions with enhancement" when report already says "pleural effusions"
- "Splenomegaly 13.5 cm" descriptive finding
- Wrong addition of "small renal cyst" (incidental, benign, no management change)
- Wrong size of a benign finding by 2mm
- Refining wording of cardiomegaly description
- "Several arterial enhancing liver lesions" elaborating already-flagged liver lesions

MODERATE:

- New benign liver lesion needing follow-up (not in report)
- Re-characterizing a 4cm renal mass as benign vs. neoplastic
- New incidental adrenal nodule
- Mass measurement off by 2x (changes follow-up timing)
- Lymph node character change (reactive vs. malignant)

CRITICAL:

- New PE in segmental branches not in report
- Missed bowel obstruction or ischemia
- Missed perforation, hemorrhage, abscess
- Tumor staging change due to wrong size/extent
- In aortic-symptoms indication: missed dissection or rupture finding

=====

INTERNAL REASONING (do silently, then output JSON):

1. What topic does the edit address?
2. Is the topic in the report? -> fixes edit_type
3. Is the edit accurate or contradicted? -> fixes agreement
4. COUNTERFACTUAL TEST: does adding/removing this edit change urgent care?
5. Apply the 5 GOLDEN RULES; resist default-to-moderate.

OUTPUT -- JSON only, no extra prose:

```
{
```

```

"reasoning": "<3-5 sentences walking through steps 1-5; explicitly name the counterfactual
outcome>",
"edit_type": "<addition | correction | clarification>",
"agreement": "<agree | partially agree | disagree>",
"severity": "<negligible | moderate | critical>"
}

```

B. Verbatim user prompt template used in Run 1031

```

clinical_indication: {indication}
preliminary_report_time: {prelim_time}

```

```

=== PRELIMINARY REPORT ===
{report}
=== END REPORT ===

```

```

PROPOSED EDIT:
{edit}

```

Output the JSON.

C. Post-competition v4 (Opus 4.7) system prompt: key change from v3

The post-competition v4 variant (Section 7) was not one of the four submitted runs. It uses claude-opus-4-7 and reuses the three-section structure of the v3 prompt in Appendix A; the distinguishing change is the preamble below, which replaces the opening paragraph of v3. The natural-edit class proportions quoted in it were read off the post-result Natural-Set leaderboard, not from any held-out gold labels. Non-ASCII characters are rendered in their ASCII equivalents for typesetting.

You are reviewing proposed edits to preliminary CT radiology reports. The edit set comes from TWO sources mixed together (you don't know which is which for any given edit):

1. NATURAL edits -- real attending physicians refining junior residents' preliminary reports. Most of these are reasonable and should be ACCEPTED.
2. SYNTHETIC edits -- adversarial test cases intentionally injected. These often contain fabrications, contradictions, or anatomically implausible claims.

CRITICAL CALIBRATION (from observed test distribution):

- * Natural edits are 86% "agree", 77% "negligible", 72% "addition"
- * The DEFAULT correct answer for any edit you're unsure about is: agreement="agree", severity="negligible", edit_type="addition"
- * This default is correct for ~58% of natural edits
- * Override the default ONLY when you have strong specific evidence

D. Reproducibility

The complete system is a single deterministic API call whose verbatim system and user prompts are reproduced in Appendices A and B (claude-sonnet-4-6, temperature=0, max_tokens=1024). Given these prompts and the public RADAR evaluation harness [2], the submitted predictions are regenerable. The submitted prediction files, the per-edit cached model responses, and the analysis scripts that recompute every number in this paper's tables are available on request. The RADAR data is governed by its public release terms [2] and is obtained from the benchmark authors; we do not redistribute it.