LAIas Bench: A Parametric Evaluation Protocol for Radiology Report Generation by Language Models

Natan Paraíso Ribeiro[1*] Raquel Moreno[1] Francisco Akira[1]
[1]*Laudos.AI — São Paulo, Brazil* [*]*Corresponding author: natan@laudos.ai*

ABSTRACT

Language models are increasingly applied to medical text generation, yet evaluation frameworks for radiology report generation remain fragmented, subjective, or narrowly scoped. We introduce LAIas Bench, an open, CLI-first evaluation framework and leaderboard harness for assessing language models on the task of generating structured radiology reports from examination metadata and clinical findings. The benchmark comprises 3,075 evaluation cases (3,049 Portuguese, 13 English, 10 cross-locale verified, plus a 10-case lite subset) synthetically generated and individually reviewed by board-certified radiologists, spanning four imaging modalities (CT, MRI, US, XR) and seven anatomical regions. LAIas Bench evaluates generated reports across five clinically-grounded dimensions—hallucination resistance (CRIT), structural quality (QUAL), terminology correctness (TERM), anatomical coverage (GUIDE), and template fidelity (RAG)—using a dual-phase scoring architecture that combines deterministic rule-based checks with adversarial LLM-as-judge verification. The framework supports locale-pluggable evaluation for Brazilian Portuguese and American English, three provider backends (OpenRouter, OpenAI-compatible endpoints, and stdin/stdout command agents), strict submission validation, and grouped leaderboards that prevent incompatible runs from mixing. We describe the benchmark construction pipeline, the five-dimensional scoring framework, the conservative score combination strategy, and the track system for fair model comparison. LAIas Bench is designed for continuous extension, requiring minimal human intervention to add new cases, locales, or evaluation dimensions.

1 INTRODUCTION

Radiology report generation is a high-stakes application of language models (LMs) where errors carry direct clinical consequences. Unlike general-purpose text generation, radiology reports must satisfy strict structural constraints, use precise modality-specific terminology, faithfully preserve all input clinical findings without hallucination, and conform to institutional formatting standards. Despite rapid advances in medical LM capabilities (Singhal et al., 2023; Nori et al., 2023; Tu et al., 2024), the evaluation landscape for radiology report generation remains fragmented across three axes.

First, existing evaluations typically rely on surface-level metrics such as BLEU, ROUGE, or BERTScore (Zhang et al., 2020) that fail to capture clinically meaningful errors—a report that halluccinates contrast enhancement in a non-contrast study may score highly on lexical overlap while introducing a dangerous factual error. **Second**, most benchmarks are English-only, ignoring the reality that radiology is practiced worldwide with locale-specific terminology, section naming conventions, and stylistic norms. Brazilian Portuguese radiology, for example, uses distinct forbidden openers (*Presença de*, *Observa-se*), modality-specific vocabulary rules (e.g., *atenuação* not *densidade* for CT), and institution-level formatting requirements that English benchmarks cannot assess. **Third**, existing frameworks provide no mechanism for fair comparison across different model access patterns—a self-hosted vLLM endpoint, an OpenRouter-hosted model, and a custom multi-step RAG agent cannot be meaningfully ranked on the same leaderboard without controlling for scaffold, judge model, and evaluation locale.

To address these gaps, we introduce **LAIas Bench** (Laudos AI Assessment Benchmark), an open evaluation framework and leaderboard harness specifically designed for the task of text-to-report radiology generation. Given an examination descriptor (e.g., *tc abdome sc*) and a set of clinical findings (e.g., *esteatose moderada. cálculo vesícula 12mm*), a language model must produce a complete, structured HTML radiology report using only the allowed tags , <b>, and
.

LAIas Bench makes the following contributions:

1. A **five-dimensional scoring framework** (CRIT, QUAL, TERM, GUIDE, RAG) with clinically-grounded checks that go far beyond surface-level metrics, including hallucination filters for contrast language, modality vocabulary enforcement, anatomical coverage matrices, and finding preservation tracking.

2. A **dual-phase evaluation architecture** combining deterministic rule-based scoring with adversarial LLM-as-judge verification, where the final score is the conservative minimum of both phases, ensuring that neither phase alone can inflate results.

3. A **locale-pluggable** evaluation system with full implementations for Brazilian Portuguese (pt-BR) and American English (en-US), each defining complete section naming, forbidden terms, modality vocabulary, coverage matrices, preservation patterns, and system prompt builders.

4. A **3,075-case evaluation dataset** synthetically generated to cover realistic clinical scenarios and individually reviewed by board-certified radiologists, with a primary 3,000-case pt-BR corpus covering CT (1,755 cases), MRI (994), XR (197), and US (54) across head, chest, abdomen, pelvis, spine, and other regions, plus 49 curated reference cases, 13 en-US cross-locale cases, and tiered suites for different evaluation needs.

5. A **CLI-first harness** with three provider backends, submission validation, grouped leaderboards keyed by comparable dimensions, a track system for fair cross-architecture comparison, and built-in support for OpenAI-compatible self-hosted endpoints.

2 TASK FORMULATION

2.1 Input and Output

Each LAIas Bench task instance consists of an examination descriptor string and a clinical findings string, both in the locale's natural language. The model receives these along with a locale-specific system prompt generated from the examination metadata (modality, contrast status, anatomical region). The expected output is a complete radiology report formatted in a constrained HTML subset using only three tags: for the title, <b> for section headers, and
 for line breaks. This constraint reflects real-world radiology information systems (RIS) that typically accept only minimal HTML formatting.

2.2 Case Structure

Each case is defined by the following fields:

```
{
  "id": "R015",
  "label": "TC Abdome — Patológico",
  "exam": "tc abdome sup sc",
  "findings": "Esteatose moderada. Cálculo vesícula 12mm.
               Cisto renal simples à direita 25mm",
  "locale": "pt-BR",
  "tags": ["ct", "abdomen", "abnormal"]
}
```

The **exam** field uses natural-language shorthand as radiologists write in practice. The **findings** field contains condensed clinical observations—not the full structured report—simulating the real workflow where a radiologist dictates findings and the system generates the formatted report.

2.3 Evaluation Metrics

The primary metric is the **combined overall score** on a 1–5 scale, computed as the weighted average of five dimension scores, where each dimension score is the conservative minimum of the deterministic and adversarial phases. A case is classified as **PASS** (overall ≥ 4.2 with adversarial phase), **PARTIAL** (overall ≥ 3.0), or **FAIL**. Any critical check failure in either phase forces a FAIL verdict regardless of the overall score. Suite-level metrics include average overall score, strict pass rate (% of PASS verdicts), and relaxed pass rate (% of non-FAIL verdicts).

3 BENCHMARK CONSTRUCTION

3.1 Source Corpus

LAIas Bench cases were synthetically generated to reflect the full diversity of real-world radiology practice. The generation process produced findings across four imaging modalities—CT, MRI, US, and XR—covering a wide range of anatomical regions and clinical presentations, from normal studies to complex multi-finding pathological cases. Each case was designed following institutional reporting standards with structured sections (Technique, Analysis, Conclusion/Impression). All generated cases were individually reviewed and refined by board-certified radiologists to ensure clinical accuracy, appropriate terminology, and realistic finding distributions.

3.2 Case Extraction Pipeline

We construct benchmark cases through a four-stage pipeline:

Stage I: Modality-stratified generation. Cases are generated across the four primary radiology modalities (CT, MRI, US, XR) with proportional representation reflecting clinical practice volumes.

Stage II: Findings extraction. Using regular expression patterns matching section headers (*Análise*, *Achados*, *Descrição*), we extract the analysis section content up to the conclusion boundary. Reports with analysis sections shorter than 40 characters are excluded. This yields 13,470 reports with extractable findings.

Stage III: Radiologist review. Each case undergoes individual review by board-certified radiologists who verify: (a) clinical plausibility of the findings for the given exam type, (b) correct modality-specific terminology, (c) appropriate finding severity and distribution, (d) correct anatomical coverage for the specified region, and (e) absence of contradictory or clinically impossible finding combinations.

Stage IV: Stratified sampling. We sample cases to maximize diversity across a (modality × region × pathological status) matrix. The primary **corpus** dataset contains 3,000 pt-BR cases sampled proportionally: CT 1,755, MRI 994, XR 197, US 54, distributed across head (914), chest (615), abdomen (319), pelvis (309), spine (276), other (564), and urinary (3) regions, with 2,570 pathological and 430 normal studies. A curated 49-case **reference** subset provides manually-verified stratified coverage for targeted analysis.

3.3 Cross-Locale Cases

To enable cross-locale evaluation, we construct 13 synthetic en-US cases covering the same modality and region distribution. These cases use standardized English radiology terminology and serve as validation anchors for the en-US locale specification. The combined evaluation dataset comprises 3,075 cases across two locales and six suite configurations.

3.4 Suite Organization

Suite	Locale	Cases	Visibility	Purpose
corpus-public.pt-BR	pt-BR	3,000	public	Primary large-scale benchmark
reference-public.pt-BR	pt-BR	49	public	Curated reference benchmark
reference-public.en-US	en-US	13	public	Cross-locale validation
lite-public.pt-BR	pt-BR	10	public	Fast iteration
verified-public.mixed	mixed	10	verified	Harness integrity
leaderboard-private	pt-BR	—	private	Hidden leaderboard split

Table 1: Suite organization in LAIas Bench v1.0. The private suite ships only as a template manifest; cases are evaluated off-repo using the same harness contract.

4 SCORING FRAMEWORK

LAIas Bench employs a dual-phase evaluation architecture. The **deterministic phase** applies rule-based checks that are reproducible, fast, and cost-free. The **adversarial phase** uses an LLM judge to catch subtle semantic errors that rules cannot detect. The two phases are combined conservatively.

4.1 Deterministic Phase

The deterministic phase runs a battery of checks organized into five dimensions, each addressing a distinct axis of clinical report quality.

CRIT — Hallucination Resistance (weight: 0.30)

The CRIT dimension detects factual errors that could cause clinical harm. Checks include: (C01) contrast language detection in non-contrast studies—if the exam is non-contrast, any mention of *realce*, *impregnação*, *wash-out*, or phase-specific terms triggers a critical failure; (C02) umbrella phrase detection in conclusions, flagging vague language like *demais estruturas sem alterações*; (C03) banned phrase detection for known incorrect patterns; (C04) foreign formatting detection (markdown, HTML tags beyond the allowed set); (C05) measurement leakage into conclusions; (C06) abnormal-study/normal-conclusion mismatch detection; and (C07) input findings preservation, requiring ≥75% of expected findings patterns to appear in the generated report.

QUAL — Structural Quality (weight: 0.25)

QUAL evaluates the report's HTML structure: (Q01) centered bold title presence; (Q02) non-abbreviated title; (Q03–Q04) no double line breaks within analysis or conclusion sections; (Q05) section separators present; (Q06) only allowed HTML tags; (Q07) no placeholder tokens; (Q08) analysis section present; (Q09) ultrasound-specific rule prohibiting technique sections; and (Q10) conclusion-does-not-duplicate-analysis check.

TERM — Terminology Correctness (weight: 0.20)

TERM enforces locale-specific radiology terminology. For pt-BR, this includes 14 forbidden term pairs (e.g., *colônico→cólico*, *patente→pérvia*, *linfadenopatia→linfonodomegalia*), 9 forbidden sentence openers (*Presença de*, *Observa-se*, etc.), and modality-specific vocabulary rules: ultrasound reports must not contain CT/MRI terms like *atenuação* or *hipersinal*; MRI reports must not contain US terms like *ecogenicidade*; CT reports must use *atenuação* instead of *densidade*.

GUIDE — Anatomical Coverage (weight: 0.15)

GUIDE checks whether the report addresses the expected anatomical structures for the given exam type. A coverage matrix maps (modality, region) pairs to lists of expected anatomical tokens. For example, a CT abdomen study is expected to mention liver, gallbladder, pancreas, spleen, kidneys, adrenals, aorta, and lymph nodes. Missing more than one expected structure triggers a major failure.

RAG — Template Fidelity (weight: 0.10)

RAG evaluates whether the report faithfully reflects the input metadata: (R01) title contains expected modality and region tokens; (R02) laterality preservation (right/left/bilateral from findings must appear in report); (R03) spinal level preservation (e.g., L4-L5); (R04) measurement preservation; (R05) key findings preservation (≥80% threshold); and (R06) correct section ordering (Technique → Analysis → Conclusion for CT/MRI, or Analysis → Conclusion for US).

4.2 Dimension Scoring

Within each dimension, the score is computed as $score = round((pass / total) \times 5, 1)$, then clamped by severity-based caps: any critical failure caps the score at $\max(1, 3 - (critFails - 1))$; three or more major failures cap the score at 3.5. Dimensions with no applicable checks receive an UNSCORED verdict and are excluded from the weighted average. The dimension verdict is PASS if all checks pass, PARTIAL if score ≥ 4.0, or FAIL if any critical check fails or score < 4.0.

4.3 Adversarial Phase

The adversarial phase sends the exam, findings, and generated HTML to an LLM judge with locale-specific instructions. The judge returns a structured JSON response with per-dimension scores (1–5), critical failure annotations, lists of missing and hallucinated findings, spot checks, and a suggested fix. The judge response is parsed with tolerance for common LLM

output artifacts (trailing commas, markdown code fences).

4.4 Score Combination

The combined score for each dimension is the **minimum** of the deterministic and adversarial scores:

$$combined_{dim} = min(det_{dim}, adv_{dim})$$

The overall score is the weighted average over scored dimensions only. This conservative combination ensures that neither a lenient judge nor a permissive rule set can inflate the final score. If the adversarial phase is unavailable (e.g., due to cost constraints or endpoint failure), the result is marked as **degraded** with low confidence, and the maximum achievable verdict is PARTIAL.

The three-tier confidence system reflects evaluation completeness: **high** confidence requires both phases completed with no hallucinations or missing findings; **medium** indicates either a missing adversarial phase with high deterministic score, or an adversarial phase that detected issues; **low** indicates missing adversarial phase.

5 LOCALE-PLUGGABLE EVALUATION

LAIas Bench implements a locale specification type (*LocaleSpec*) that encapsulates all language-specific evaluation parameters. Each locale defines: section name patterns (Analysis, Conclusion, Technique), section labels, title abbreviation patterns, forbidden terms with corrections, forbidden sentence openers, contrast terminology patterns, umbrella terms, banned phrases, normal study patterns, modality vocabulary rules, title modality and region tokens, anatomical coverage matrices, finding preservation patterns, a region classification function, a system prompt builder, and judge instructions.

The pt-BR locale implements 14 forbidden term corrections derived from Brazilian College of Radiology (CBR) terminology guidelines. The en-US locale implements equivalent rules for American radiology conventions. Adding a new locale requires implementing a single *LocaleSpec* object and registering it in the locale index—no changes to the scoring engine, checks, or CLI are needed.

6 INFRASTRUCTURE AND HARNESS DESIGN

6.1 Provider Architecture

LAIas Bench supports three provider backends through a unified *GeneratorAdapter* interface: (1) **OpenRouter** for hosted model access with automatic cost tracking; (2) **OpenAI-compatible** for any endpoint implementing the /chat/completions API (vLLM, LM Studio, Fireworks, custom gateways), with configurable authentication, headers, and body parameters; and (3) **Command** for local agents communicating over stdin/stdout JSON. All HTTP providers implement exponential backoff retry with Retry-After header support for status codes 429, 500, 502, 503, and 504.

6.2 Track System

Track	Scaffold	Description
mini-agent	mini-laias-agent-v1	Canonical scaffold for apples-to-apples LM ranking
model	—	Direct-model experiments without scaffold parity
agent	—	Custom agents, RAG systems, multi-step pipelines

Table 2: Track system for fair comparison. The mini-agent track uses a canonical minimal scaffold to isolate model capability from system design.

6.3 Leaderboard Design

Leaderboard entries are grouped by a composite **comparable key** consisting of: benchmark version, suite ID, locale, track, scaffold ID, judge provider, and judge model. This ensures that runs using different judges, locales, or scaffolds are never ranked against each other. Within a group, entries are ranked by: (1) eligibility (valid submissions first), (2) average overall score descending, (3) strict pass rate descending, (4) relaxed pass rate descending, (5) total cost ascending, (6) average latency ascending. Invalid submissions (those failing validation) receive a null rank but can still be evaluated for

debugging.

6.4 Submission Validation

LAIas Bench provides strict submission validation for the predictions mode, checking for: missing case IDs, duplicate case IDs, extra case IDs, empty outputs, and malformed JSONL. Invalid submissions can still be evaluated but are flagged as ineligible for leaderboard ranking.

7 DESIGN COMPARISON WITH RELATED BENCHMARKS

LAIas Bench draws architectural inspiration from SWE-bench (Jimenez et al., 2024) while addressing a fundamentally different evaluation domain. Both benchmarks share several design principles: carefully constructed task instances grounded in domain practice (GitHub issues for SWE-bench, radiologist-reviewed clinical scenarios for LAIas Bench); execution-based evaluation rather than surface-level metrics; continuous updatability with minimal human intervention; and diverse, long inputs requiring the model to produce structured outputs.

However, key differences reflect the domain-specific requirements of radiology evaluation:

Aspect	SWE-bench	LAIas Bench
Domain	Software engineering	Radiology report generation
Input	Issue text + codebase	Exam descriptor + findings
Output	Patch file	HTML radiology report
Evaluation	Unit test execution	5-dim deterministic + adversarial
Verdict	Binary (pass/fail)	Graded (1–5 per dimension)
Locale	Python only	Pluggable (pt-BR, en-US)
Cases	2,294 (12 repos)	3,075 (4 modalities, 7 regions)
Source	GitHub PRs	Synthetic + radiologist-reviewed

Table 3: Design comparison between SWE-bench and LAIas Bench.

A fundamental distinction is that SWE-bench uses binary pass/fail evaluation (all unit tests must pass), while LAIas Bench provides graded scoring across five dimensions. This reflects the reality that radiology report quality exists on a spectrum—a report may have correct structure but incorrect terminology, or faithful findings but poor anatomical coverage. The graded scoring enables fine-grained model comparison and targeted improvement analysis.

8 RELATED WORK

Medical Report Generation.

Prior work on radiology report generation has focused primarily on image-to-report models (Jing et al., 2018; Li et al., 2018; Chen et al., 2020; Miura et al., 2021; Tanida et al., 2023) where a chest X-ray image is the input. LAIas Bench addresses a complementary but distinct task: text-to-report generation, where the input is structured clinical findings rather than images. This task arises naturally in voice-dictation radiology workflows and AI-assisted reporting systems where findings are extracted from speech or entered as structured data.

Medical NLP Benchmarks.

General medical benchmarks such as MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and MultiMedBench (Tu et al., 2024) evaluate medical knowledge broadly but do not address the specific constraints of structured report generation. RadBench (Wu et al., 2023) evaluates radiology understanding but focuses on question answering rather than generation. CheXpert labeling (Irvin et al., 2019) provides structured labels but not complete report evaluation.

LLM-as-Judge Evaluation.

The use of LLMs as evaluators has gained traction across NLP (Zheng et al., 2024; Liu et al., 2023). LAIas Bench combines LLM-as-judge with deterministic checks in a dual-phase architecture, using the conservative minimum rather than averaging or weighting, which provides stronger guarantees against score inflation.

Code Generation Benchmarks.

SWE-bench (Jimenez et al., 2024) introduced the paradigm of evaluating LMs on real-world repository-scale tasks with execution-based verification. LAIas Bench adapts this philosophy to radiology, replacing unit test execution with multi-dimensional clinical quality assessment while retaining the principles of real-world data sourcing, continuous updatability, and strict evaluation.

9 DISCUSSION

9.1 Limitations and Future Directions

Case volume. The current public dataset of 3,075 cases provides substantial statistical power for model ranking. The 3,000-case pt-BR corpus enables fine-grained comparisons between closely-performing models with confidence. We plan to continue expanding the private leaderboard split and generating additional cases covering underrepresented modality-region combinations.

Coverage matrices. The anatomical coverage matrices (GUIDE dimension) are currently sparse for certain (modality, region) combinations. Expanding these matrices with subspecialty radiology input would improve the GUIDE dimension's discriminative power.

Image-conditioned evaluation. LAIas Bench currently evaluates text-to-report generation only. Extending to image-conditioned evaluation (where the input includes DICOM images alongside findings) would enable assessment of end-to-end radiology AI systems.

Multi-language expansion. While the locale system supports arbitrary languages, only pt-BR and en-US are currently implemented. Priority languages for expansion include Spanish, French, and German, which have distinct radiology terminology conventions.

Regulatory alignment. Future versions should incorporate checks aligned with specific regulatory frameworks governing radiology practice, such as structured reporting standards from the Brazilian College of Radiology (CBR) and ACR guidelines for report content and format.

9.2 Ethical Considerations

All cases in LAIas Bench are synthetically generated and contain no patient data, physician identifiers, institutional references, or protected health information. No real patient reports were used in the construction of this benchmark. The benchmark is released under the MIT license. We emphasize that LAIas Bench evaluates report generation quality but does not constitute clinical validation—systems evaluated on this benchmark should undergo separate clinical validation before deployment.

10 CONCLUSION

We have presented LAIas Bench, an open evaluation framework for radiology report generation that combines clinically-grounded multi-dimensional scoring with adversarial LLM verification. With 3,075 synthetically generated and radiologist-reviewed cases, locale-pluggable evaluation for Portuguese and English, and implementing a fair leaderboard system with comparable-key grouping, LAIas Bench addresses critical gaps in the evaluation of language models for medical text generation. The benchmark is designed to be continuously extended with new cases, locales, and evaluation dimensions, serving as both a practical development tool and a rigorous comparison framework. We release the complete evaluation harness, case datasets, and CLI tooling as open-source software to accelerate progress in safe, accurate, and clinically-useful radiology AI.

REFERENCES

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating Radiology Reports via Memory-Driven Transformer. EMNLP.

Irvin, J., Rajpurkar, P., Ko, M., et al. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). SWE-bench: Can Language Models Resolve Real-World GitHub Issues? ICLR.

Jin, D., Pan, E., Oufattole, N., Weng, W., Fang, H., & Szolovits, P. (2021). What Disease Does This Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. EMNLP.

Jing, B., Xie, P., & Xing, E. (2018). On the Automatic Generation of Medical Imaging Reports. ACL.

Li, Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. NeurIPS.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP.

Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., & Jurafsky, D. (2021). Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. NAACL.

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on Medical Challenge Problems. arXiv:2303.13375.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large Language Models Encode Clinical Knowledge. Nature.

Tanida, T., Müller, P., Winkel, D., et al. (2023). Interactive and Explainable Region-guided Radiology Report Generation. CVPR.

Tu, T., Azizi, S., Driess, D., et al. (2024). Towards Generalist Biomedical AI. NEJM AI.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). Towards Generalist Foundation Model for Radiology. arXiv:2308.02463.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. ICLR.

Zheng, L., Chiang, W., Sheng, Y., et al. (2024). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS.

APPENDIX A: CASE DISTRIBUTION

The following tables present the case distribution for the 3,000-case pt-BR corpus and the 49-case curated reference subset.

Table A1: Corpus (3,000 cases)

Modality	Head	Chest	Abdomen	Pelvis	Spine	Other	Total
CT	568	546	203	160	157	121	1,755
MRI	344	3	116	149	119	263	994
XR	2	66	—	—	—	129	197
US	—	—	—	—	—	54	54
Total	**914**	**615**	**319**	**309**	**276**	**567**	**3,000**

Table A1: Distribution of the 3,000-case pt-BR corpus across modality and anatomical region. 2,570 cases are pathological, 430 are normal studies. All cases synthetically generated and reviewed by board-certified radiologists.

Table A2: Reference Subset (49 cases)

Modality	Head	Chest	Abdomen	Pelvis	Spine	Urinary	Other	Total
CT	6	5	6	4	2	—	—	23
MRI	4	—	3	3	3	—	2	15
US	—	—	3	2	—	2	2	9
XR	—	—	—	—	—	—	2	2
Total	**10**	**5**	**12**	**9**	**5**	**2**	**6**	**49**

Table A2: Distribution of the 49-case curated reference subset across modality and anatomical region.

APPENDIX B: DETERMINISTIC CHECK REGISTRY

ID	Dim	Severity	Description
C01	CRIT	critical	No contrast language in non-contrast exam
C02	CRIT	major	No umbrella phrase in conclusion
C03	CRIT	critical	No banned phrases
C04	CRIT	critical	No markdown or foreign HTML/XML
C05	CRIT	major	No measurements in conclusion
C06	CRIT	minor	Abnormal study ≠ normal conclusion lead
C07	CRIT	critical	Input findings preserved ≥ 75%
Q01	QUAL	critical	Starts with centered bold title
Q02	QUAL	critical	Title not abbreviated
Q03	QUAL	critical	No double breaks in analysis
Q04	QUAL	critical	No double breaks in conclusion
Q05	QUAL	major	Has section separators

Q06	QUAL	major	Only allowed HTML tags
Q07	QUAL	critical	No placeholder tokens
Q08	QUAL	critical	Analysis section present
Q09	QUAL	critical	US has no technique section
Q10	QUAL	major	Conclusion ≠ analysis duplication
R01	RAG	critical	Title preserves modality + region
R02	RAG	major	Laterality preserved
R03	RAG	major	Spinal levels preserved
R04	RAG	major	Measurements preserved in body
R05	RAG	critical	Key findings preserved ≥ 80%
R06	RAG	major	Section order preserved

Table B1: Complete deterministic check registry (CRIT, QUAL, RAG dimensions). TERM and GUIDE checks are dynamically generated from the locale specification.

APPENDIX C: SCORING WEIGHTS AND COMBINATION

Dimension	Weight	Critical Cap	Major Cap
CRIT	0.30	max(1, 3 – (n–1))	3.5 if ≥ 3
QUAL	0.25	max(1, 3 – (n–1))	3.5 if ≥ 3
TERM	0.20	max(1, 3 – (n–1))	3.5 if ≥ 3
GUIDE	0.15	max(1, 3 – (n–1))	3.5 if ≥ 3
RAG	0.10	max(1, 3 – (n–1))	3.5 if ≥ 3

Table C1: Scoring weights and severity-based score caps. Weights are renormalized over scored dimensions only when some dimensions are UNSCORED.