

Beyond Templates: A Compositional Model and Lower Bound for Radiology Report Variability

Natan Paraiso Ribeiro, MD Petrus Paraiso Ribeiro, MD Stephanie Alba Herrera, MD
Francisco Akira, MD Raquel Moreno, MD

Laudos.AI Research • oi@laudos.ai

Draft version – May 2026

Abstract

Radiology reporting is often framed as template selection, dictation, or free-text generation. This framing is incomplete. Diagnostic reports are rendered from a high-dimensional clinical state space involving modality, protocol, anatomy, indication, technique, findings, quantitative attributes, temporal comparison, diagnostic impression, recommendations, and local reporting style. Static template libraries improve standardization, but they attempt to pre-author report outputs in a space whose clinically meaningful variants grow combinatorially. We introduce a formal model of radiology report variability, define the Report Space Lower Bound (RSLB) and a log-scale Radiology Report Variability Index (RRVI), and compare the asymptotic scaling of template enumeration with compositional report generation. The model shows that even conservative subsets of mutually compatible reporting variables can exceed the practical capacity of manually maintained template catalogs by several orders of magnitude. We then propose a compositional report-engine architecture based on protocol ontologies, anatomy graphs, finding libraries, attribute schemas, clinical constraints, impression logic, consistency validation, and report rendering. The claim is not that templates are useless. Rather, templates are useful as presentation surfaces, but they are the wrong unit of computational scale. A scalable reporting system should represent report-relevant clinical state and render reports from constrained semantic components.

Keywords. radiology reporting; structured reporting; report generation; templates; clinical informatics; combinatorial modeling; semantic components; large language models.

1 Introduction

Radiology reports are compact textual artifacts derived from complex clinical and imaging states. A report for a single procedure may depend on indication, protocol, technique, anatomic coverage, findings, measurements, temporal comparison, diagnostic impression, recommendations, and institution-specific style. Conventional reporting systems frequently treat this task as free dictation, template selection, or automated text generation. Each view captures part of the problem, but none directly represents the underlying combinatorial structure of reporting.

Structured reporting initiatives have demonstrated the value of standardization. RSNA’s RadReport library provides reporting templates for common radiology procedures and many templates incorporate RadLex terminology and RadElement common data elements [1]. The LOINC/RSNA Radiology Playbook and RadLex initiatives address the variation in procedure naming and terminology across radiology departments [4, 2]. The European Society of Radiology has emphasized

the potential of structured reporting while also noting that clinical adoption remains incomplete, cross-institutional template use is limited, and consensus on technical standards is still missing [5]. These facts suggest that reporting is not merely a content-authoring problem; it is a representation and composition problem.

The central thesis of this paper is that radiology reporting is not primarily a template-selection problem. It is a constrained combinatorial generation problem. Static templates can be effective for narrow protocols and for standardizing final presentation. However, when the system must support many procedures, findings, attributes, temporal comparisons, and reporting styles, pre-authoring every clinically meaningful output variant is not scalable. A compositional report engine instead represents reusable semantic units and renders a report from a validated clinical state.

This paper makes four contributions:

1. A formal definition of report-relevant clinical state and report space.
2. A lower-bound estimate, the Report Space Lower Bound (RSLB), with a log-scale Radiology Report Variability Index (RRVI).
3. A scaling comparison between static template enumeration and compositional report generation.
4. A system architecture and evaluation protocol for a compositional radiology report engine.

The intended contribution is conceptual and formal. We do not claim to enumerate all possible reports or to replace physician judgment. We estimate conservative lower bounds for clinically meaningful report states and use them to motivate a different computational unit of scale.

2 Background and Related Work

2.1 Structured reporting and terminology standards

Structured reporting has long been proposed as a mechanism to reduce ambiguity, improve completeness, enable downstream analytics, and standardize communication. RadReport, RadLex, RadElement common data elements, and the LOINC/RSNA Radiology Playbook represent important infrastructure for this effort [1, 2, 4]. Common data elements are particularly relevant because they define standardized questions and allowable answers for observations, supporting reporting, decision support, and data analysis [3].

The ESR’s 2023 update describes both the promise and the limited routine adoption of structured reporting. It notes that structured reporting has substantial potential to improve service quality, but that cross-institutional template applications and usage incentives remain limited; it also highlights interoperable formats such as FHIR and the future relevance of large language models for workflow integration [5]. This supports the view that the next step is not simply more static templates, but a reporting representation that can interoperate with clinical data standards and adapt to local workflow.

2.2 Automated radiology report generation

Recent work in automated radiology report generation focuses on deep learning architectures, multimodal models, datasets, knowledge integration, and evaluation metrics [6]. Structured radiology report generation has become an active subproblem. Delbrouck et al. introduced a structured radiology report generation task that reformulates free-text reports into a standardized format, with new datasets and evaluation methods [8]. Niu et al. developed a dynamic-template-constrained large

language model for fully structured lung cancer screening reports, reporting high cross-institutional performance in converting free-text reports into structured reports [7]. Other recent work has emphasized rich clinical context, fine-grained structured evaluation, and category-wise modular decoding [9, 10, 11]. Information extraction resources such as RadGraph-XL further illustrate the importance of structured entities and relations in radiology text [12].

These works are complementary to the present paper. They address generation, structuring, extraction, and evaluation. The contribution here is to formalize why static output enumeration is an insufficient scaling primitive for radiology reporting and to define a compositional alternative.

3 Problem Definition

Let p denote a radiology procedure. A procedure-specific reporting model can be represented by the tuple

$$\mathcal{R}_p = (M_p, P_p, A_p, I_p, T_p, F_p, Q_p, C_p, D_p, S_p), \quad (1)$$

where M_p is the imaging modality, P_p is the procedure or protocol, A_p is the relevant anatomy set, I_p is the set of indications, T_p is the set of technique variants, F_p is the set of finding slots, Q_p contains finding-level attribute domains, C_p is the temporal comparison domain, D_p is the diagnostic impression domain, and S_p denotes style or institution-specific rendering conventions.

A report-relevant clinical state is a structured object

$$x = (i, t, c, s, z_1, z_2, \dots, z_n), \quad (2)$$

where $i \in I_p$, $t \in T_p$, $c \in C_p$, $s \in S_p$, and each z_j is the state of a finding slot $f_j \in F_p$. A finding slot may be normal, absent, not evaluated, or abnormal with attributes such as size, number, location, laterality, severity, density, signal, enhancement, morphology, certainty, or interval change.

Not all combinations are clinically valid. Let

$$\Gamma_p(x) \in \{0, 1\} \quad (3)$$

be a constraint predicate that accepts only internally consistent and procedure-compatible states. The exact report space is therefore

$$\Omega_p = \{x \in X_p : \Gamma_p(x) = 1\}, \quad (4)$$

where X_p is the unconstrained Cartesian product of all report-relevant variables.

A report renderer is a function

$$\rho_p : \Omega_p \rightarrow \mathcal{Y}_p, \quad (5)$$

where $\mathcal{Y}_p$ is the set of textual reports or structured report documents. A validator is a function

$$\nu_p : \mathcal{Y}_p \times \Omega_p \rightarrow \{0, 1\}, \quad (6)$$

which determines whether a rendered report is semantically consistent with the source state and with reporting constraints.

Two report states x_a and x_b are clinically distinct if changing from x_a to x_b changes at least one report-relevant fact that could alter interpretation, follow-up, management, or downstream data extraction. This definition intentionally ignores purely stylistic variants unless style affects interpretability or institutional compliance.

4 Report Space Lower Bound

The exact cardinality $|\Omega_p|$ is difficult to compute because clinical constraints can be complex. For scalability analysis, exact enumeration is unnecessary. A lower bound is enough.

Let $B_p = \{b_1, \dots, b_k\}$ be a set of mutually compatible report-relevant variables for procedure p . Each variable b_j has a finite domain V_j of clinically distinct values. The variables in B_p are chosen so that every combination in $\prod_j V_j$ satisfies the relevant subset of constraints. Then the Report Space Lower Bound is

$$\text{RSLB}(p; B_p) = \prod_{j=1}^k |V_j|. \quad (7)$$

For interpretability across procedures, define the Radiology Report Variability Index as a log-scale score:

$$\text{RRVI}(p; B_p) = \log_{10} \text{RSLB}(p; B_p). \quad (8)$$

An RRVI of 6 means at least 10^6 clinically distinct report states under the selected conservative variable subset.

For a finding slot f with abnormal attributes $Q_f = \{q_1, \dots, q_m\}$, a conservative state domain can be represented as

$$V_f = \{\text{normal/absent}\} \cup \left(\prod_{r=1}^m Q_{f,r} \right), \quad (9)$$

where $Q_{f,r}$ is the value set for attribute r . This definition counts one normal or absent state plus abnormal combinations. If this overstates dependencies among attributes, a smaller subset of valid abnormal states can be selected. Because RSLB is a lower bound, the construction should favor conservative, mutually compatible domains.

4.1 Scaling proposition

Proposition. If a procedure p contains k mutually compatible report-relevant variables with cardinalities $|V_1|, \dots, |V_k|$, any static template catalog that stores one complete output for each clinically distinct state requires $\Omega(\prod_{j=1}^k |V_j|)$ stored outputs. A compositional representation that stores the variable domains, rendering components, and constraints requires $O(\sum_{j=1}^k |V_j| + |\mathcal{C}_p| + |\mathcal{G}_p|)$ representational units, where $\mathcal{C}_p$ is the constraint set and $\mathcal{G}_p$ is the rendering grammar or generation logic.

Proof sketch. By definition, the chosen variables are mutually compatible and clinically distinct. Therefore the Cartesian product contains $\prod_j |V_j|$ distinct report states. A catalog that stores a complete output for each state must contain at least one output per state, producing the product lower bound. A compositional system stores values and rules separately and combines them at render time. Its storage grows with the number of components, constraints, and grammar rules rather than with the number of complete combinations. This does not eliminate validation complexity, but it changes the unit of scale from complete reports to reusable semantic components. $\square$

5 Illustrative Lower-Bound Calculations

Table 1 gives illustrative conservative calculations. These are not estimates of the true number of possible reports. They intentionally use small domains to show that the lower bound becomes large even before adding many real-world factors such as institution-specific phrasing, recommendations,

uncertainty, multi-finding interactions, structured measurements, or guideline-specific management logic.

Table 1: Illustrative conservative RSLB calculations.

Procedure abstraction		Conservative assumptions	RSLB	RRVI
Single-region sound	ultra-	4 indications, 2 technique variants, 3 comparison states, 8 independent finding slots, 5 states per slot	9,375,000	6.97
Routine CT protocol		5 indications, 4 technique variants, 4 comparison states, 12 independent finding slots, 6 states per slot	174,142,586,880	11.24
Routine MRI protocol		6 indications, 4 technique variants, 4 comparison states, 10 independent finding slots, 7 states per slot	27,117,623,904	10.43

The first row exceeds nine million distinct states using only eight finding slots and five states per slot. The CT and MRI abstractions exceed 10^{10} states. These numbers do not imply that a radiologist writes billions of unique prose reports. They imply that a reporting system must be able to represent many clinically distinct states without pre-authoring one static template per state.

6 Why Static Template Libraries Do Not Scale

A static template library stores report outputs or report skeletons. The primary maintenance problem is coverage: as protocols, findings, measurements, comparison states, and recommendations expand, the library must either add more templates or embed increasingly complex branching logic inside templates. Both strategies shift complexity into manual authoring and maintenance.

A compositional engine stores semantic components. The core objects are findings, attributes, constraints, impression rules, and rendering rules. Templates may still exist, but only as final presentation surfaces. The computational core is not the template; it is the validated clinical state.

The distinction is shown in Table 2.

Table 2: Template enumeration versus compositional reporting.

Dimension	Static template catalog	Compositional report engine
Primary stored unit	Complete or partially complete report text	Semantic component, attribute, constraint, renderer
Scaling behavior	Product of clinically distinct combinations	Sum of reusable components plus constraints
Failure mode	Template proliferation, omissions, stale variants	Constraint gaps, renderer bugs, validation gaps
Update pattern	Add or edit templates for each use case	Add component once, reuse across protocols
Validation target	Textual completeness and formatting	State validity, semantic consistency, rendered output
Role of templates	Core representation	Presentation layer

This does not make templates obsolete. It reframes them. A template is useful for layout, sectioning, and local style. It is not the right primitive for representing the combinatorial report state space.

7 Compositional Report-Engine Architecture

A compositional report engine can be organized into seven layers:

1. **Protocol ontology.** Represents modality, procedure, acquisition variants, contrast phases, clinical indications, and protocol-specific constraints.
2. **Anatomy graph.** Represents organs, subregions, laterality, adjacency, containment, and procedure-specific anatomic coverage.
3. **Finding library.** Stores finding slots, normal states, abnormal states, synonyms, structured descriptors, and links to terminologies such as RadLex or common data elements.
4. **Attribute schema.** Defines value domains for size, number, location, laterality, severity, morphology, signal, density, enhancement, certainty, and temporal change.
5. **Constraint engine.** Rejects impossible, contradictory, unsupported, or underspecified states before rendering.
6. **Impression generator.** Maps findings and temporal states into clinically coherent impressions and recommendations under explicit rules.
7. **Consistency validator and renderer.** Produces the final report and verifies agreement between state, findings, impression, and output text.

This architecture can be implemented with deterministic rules, probabilistic models, large language models, or hybrid systems. The key requirement is not the use of a particular model class. It is the preservation of an auditable intermediate semantic state that can be constrained and validated.

8 Evaluation Protocol

A compositional report engine should be evaluated against static templates and against unconstrained generation. The following protocol is designed to be feasible without private patient data.

8.1 Coverage and scalability

For each procedure, define a reference set of report-relevant variables and constraints. Measure:

- RSLB and RRVF under a declared conservative variable subset.
- Component reuse ratio: fraction of components reused across protocols.
- Time-to-add-protocol: expert time required to add a new protocol and pass validation tests.
- Coverage delta: number of additional report states supported after adding a component.

8.2 Validity and consistency

Generate synthetic clinical states from the validated state space and adversarial invalid states from constraint violations. Measure:

- Valid render rate: proportion of valid states that render without missing required content.
- Contradiction rejection rate: proportion of invalid states rejected before rendering.
- Semantic consistency: agreement between rendered findings and impression.
- Determinism or bounded variability: whether equivalent states render equivalent clinical content.

8.3 Clinical review

A blinded radiologist panel can score reports generated from the same state by a static-template baseline, a compositional engine, and an unconstrained text generator. Suggested dimensions are clinical correctness, completeness, clarity, absence of contradiction, actionability of impression, and perceived edit burden. The evaluation should separate stylistic preference from clinically meaningful state preservation.

8.4 Interoperability

For each generated report, measure mapping coverage to controlled terminologies and structured data standards: proportion of findings, attributes, and recommendations with a corresponding RadLex, CDE, LOINC/RSNA Playbook, or FHIR-compatible representation where applicable.

8.5 Reference implementation: LAIBench

The protocol above is implemented as an open evaluation harness, *LAIBench* (<https://laibench.laudos.ai>). LAIBench operationalizes the compositional thesis by scoring finding-text-to-report generation on five weighted dimensions (critical-finding preservation, clinical quality, terminology, guideline coverage, evidence fidelity) under hard failure gates: a missed or fabricated critical finding caps the score and forces failure rather than being averaged away, and there is no prose or aesthetic axis. This makes LAIBench a concrete instrument for the validity-and-consistency tests of Section 8: synthetic valid/invalid report states (derivable without private patient data) become positive and adversarial cases, and the critical-finding gate measures exactly the preservation property a compositional engine must guarantee. We distinguish three evaluation tiers to keep claims honest: a public synthetic smoke set (contaminable; for harness inspection only), a gated controlled set scored aggregate-only as a first-party result, and a hidden holdout reserved for externally adjudicated claims. The results reported by the harness are first-party and aggregate; external inter-rater adjudication and a hidden-holdout split are required before any clinical-validation claim, and are explicitly out of scope here.

9 Discussion

The main risk in this framing is overstatement. Static templates are useful and can be highly effective in narrow domains. The claim is narrower: static template libraries are not a scalable computational core for broad radiology reporting when clinically distinct states grow as products

of findings, attributes, comparisons, and impressions. The scalable unit is the clinically constrained semantic component.

Another limitation is independence. Real clinical variables are not fully independent. Some findings are mutually exclusive; some attributes are conditionally required; some combinations are anatomically impossible. The RSLB construction addresses this by using only mutually compatible subsets. Stronger empirical versions of the model should estimate constraints from expert-authored schemas or structured reporting corpora.

The proposed architecture also introduces new risks. Constraint engines can omit important rules. Renderers can produce grammatical but clinically misleading text. Impression logic can over-prioritize or under-prioritize findings. Therefore, the system must preserve auditability, versioned components, clinical validation tests, and physician oversight.

Finally, large language models can be useful inside the architecture, but they should not replace the semantic state. An unconstrained language model may generate plausible text without preserving the full report state. A safer design constrains generation through explicit state, schemas, validation, and deterministic or bounded rendering logic.

10 Conclusion

Radiology reporting should be modeled as constrained combinatorial generation. Static templates remain useful for layout and presentation, but broad clinical coverage requires a compositional representation of protocol, anatomy, findings, attributes, temporal comparison, impression logic, and constraints. The Report Space Lower Bound and RRVI provide a simple way to quantify why manual template enumeration does not scale. The practical implication is direct: scalable radiology reporting systems should store semantic components and render validated clinical states, rather than attempting to pre-author the report space.

Ethics and clinical safety

This paper presents a formal and architectural model. It does not evaluate patient outcomes, does not claim autonomous diagnosis, and does not recommend removing radiologist oversight. Any implementation intended for clinical use would require prospective validation, local governance, privacy review, regulatory assessment, and continuous quality monitoring.

References

- [1] Radiological Society of North America. *RadReport reporting templates*. Accessed May 11, 2026. <https://www.rsna.org/practice-tools/data-tools-and-standards/radreport-reporting-templates>.
- [2] Radiological Society of North America. *RadLex radiology lexicon*. Accessed May 11, 2026. <https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon>.
- [3] Radiological Society of North America. *RadElement common data elements*. Accessed May 11, 2026. <https://www.rsna.org/practice-tools/data-tools-and-standards/radelement-common-data-elements>.

- [4] Regenstrief Institute and Radiological Society of North America. *LOINC/RSNA Radiology Playbook User Guide*. 2025. Accessed May 11, 2026. <https://loinc.org/kb/users-guide/loinc-rsna-radiology-playbook-user-guide/>.
- [5] European Society of Radiology. ESR paper on structured reporting in radiology - update 2023. *Insights into Imaging*, 14, 2023. <https://doi.org/10.1186/s13244-023-01560-0>.
- [6] Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated Radiology Report Generation: A Review of Recent Advances. *IEEE Reviews in Biomedical Engineering*, 2024. arXiv:2405.10842. <https://arxiv.org/abs/2405.10842>.
- [7] Chuang Niu, Md Sayed Tanveer, Md Zahirul Islam, Parisa Kaviani, Qing Lyu, Mannudeep K. Kalra, Christopher T. Whitlow, and Ge Wang. Development and Validation of a Large Language Model for Generating Fully-Structured Radiology Reports. 2024. arXiv:2409.18319. <https://arxiv.org/abs/2409.18319>.
- [8] Jean-Benoit Delbrouck et al. Automated Structured Radiology Report Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 26813–26829, 2025. <https://aclanthology.org/2025.acl-long.1301/>.
- [9] Seongjae Kang, Dong Bok Lee, Juho Jung, Dongseop Kim, Won Hwa Kim, and Sunghoon Joo. Automated Structured Radiology Report Generation with Rich Clinical Context. 2025. arXiv:2510.00428. <https://arxiv.org/abs/2510.00428>.
- [10] Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. S-RRG-Bench: Structured Radiology Report Generation with Fine-Grained Evaluation Framework. 2025. arXiv:2508.02082. <https://arxiv.org/abs/2508.02082>.
- [11] Shantam Srivastava, Mahesh Bhosale, David Doermann, and Mingchen Gao. CWCD: Category-Wise Contrastive Decoding for Structured Medical Report Generation. 2026. arXiv:2604.10410. <https://arxiv.org/abs/2604.10410>.
- [12] Jean-Benoit Delbrouck et al. RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12902–12915, 2024. <https://aclanthology.org/2024.findings-acl.765/>.