From Hidden Causality to Verifiable Medical Explanation: The HIDDEN-RAD Shared Tasks at NTCIR-18 and NTCIR-19

Key-Sun Choi[1,2], **You-Sang Cho**[1]
[1]Konyang University, Republic of Korea
[2]Korea Advanced Institute of Science and Technology, Republic of Korea

Abstract

Radiology reports state findings and impressions but often leave their inferential connection implicit. This paper presents the HIDDEN-RAD shared-task series as a corpus-based framework for externalizing and auditing that hidden reasoning. The NTCIR-18 pilot operationalized causal-explanation generation from free-text reports and structured diagnostic questionnaires derived from MIMIC-CXR. Its 1,219/314 cases for Task 1 and 804/216 cases for Task 2 were evaluated with contextual and biomedical similarity, GPT-White, GPT-Black, and selective radiologist review. HIDDEN-RAD2 at NTCIR-19 changes the corpus object from a whole generated explanation to an evidence-linked explanation that can be classified, localized, and corrected. Task 1 predicts the A1–A5 diagnostic structure from chest radiographs; Task 2 evaluates valid and controlled-hallucination variants drawn from 200 MIMIC-CXR cases plus 17 IU X-Ray expert-validated cases. A two-case discourse analysis shows how report statements, thoracic localization, and confirmation-checklist evidence license distinct explanation relations, and why report-new content is not necessarily unsupported. Radiologist review accepted 28 of 32 explicitly judged finding–impression links and revealed error-type-dependent limits in 68 injected-hallucination candidates. The series therefore develops from a generation benchmark into a provenance-aware reliability framework in which medical explanations are represented, verified, and repaired against structured clinical evidence.

Keywords: causal explanation, radiology reports, corpus annotation, LLM evaluation, hallucination verification

1. Introduction

A radiology report usually places observed evidence in a Findings section and a condensed diagnostic conclusion in an Impression section. The discourse relation linking the two is clinically central but linguistically under-specified: a reader must infer why a density, distribution, device position, or anatomical pattern licenses a diagnosis. Generating another fluent report does not necessarily recover this relation. In a high-stakes setting, an explanation can be plausible in form while remaining unsupported, incomplete, or contradictory.

HIDDEN-RAD addresses this gap by treating the inferential bridge itself as a corpus object. The NTCIR-18 pilot asked systems to generate causal explanations from either free-text radiology reports or structured questionnaire answers. HIDDEN-RAD2 at NTCIR-19 retains the diagnostic structure but adds verification: explanations can be marked valid or invalid, erroneous spans can be localized and typed, and corrected language can be proposed. The central development is therefore not only an increase in data. It is a change from reference generation to evidence-grounded auditability.

We ask three questions: (RQ1) How can hidden causal reasoning in chest X-ray interpretation be represented as an annotatable corpus resource? (RQ2) How can language models recover that reasoning from reports, images, or structured diagnostic questionnaires? (RQ3) How should generated causal explanations be evaluated beyond surface similarity? We contribute a longitudinal account of the two shared tasks, detailed dataset and annotation descriptions, a provenance-aware discourse analysis of two complementary cases, and an expert-grounded analysis of GPT-based generation and evaluation.

2. Causal Explanation as a Corpus Object

2.1 From report sections to latent relations

We distinguish a causal explanation from a paraphrase or summary. A summary compresses propositions already stated; a causal explanation makes explicit a licensed relation between evidence and conclusion. In HIDDEN-RAD the minimal structure is evidence or finding → medically licensed inference → impression. The relation can include location, distribution, absence of competing signs, and confirmation-checklist evidence. Because several phrasings may express the same inference, a single reference text cannot exhaust the space of acceptable explanations.

Field	Corpus role	Example
A1	Initial candidate	Pneumonia
A2	Location	RUL parenchyma
A3	Thoracic range	T1–T4
A4	Final impression	Pneumonia
A5	Confirmation	Q7, Q8

Table 1: A1–A5 annotation fields.

2.2 From matching to verifiability

A clinically useful evaluation must examine evidence support, causal and logical coherence, clinical plausibility, and coverage of the final impressions. HIDDEN-RAD2 further defines verifiability as the ability to decide whether an explanation is supported, identify the unsupported span, assign an error category, and produce a correction. This representation turns evaluation from a global verdict into an auditable sequence: structured evidence → explanation → validity label → error span/type → correction.

3. NTCIR-18 HIDDEN-RAD

3.1 Source corpus and annotation scheme

The pilot dataset was derived from MIMIC-CXR, a de-identified chest-radiograph resource pairing images and reports (Johnson et al., 2024). Its annotation scheme follows a radiologist-oriented reading path: A1 records preliminary impressions; A2 identifies anatomical location and laterality; A3 represents thoracic-spine level; A4 records final impressions; and A5 is a 28-item confirmation checklist covering airways, lung fields, hila, vasculature, pleura, diaphragm, heart, mediastinum, bones, soft tissue, and upper abdomen.

Task 1 contained 1,219 training and 314 evaluation cases. Its input consisted of a report, with chest images optionally available, and its target was an expert causal-explanation report. Task 2 replaced the report with structured A1–A4

Evaluator	Primary target	Pilot evidence
BERT/Cosine	Surface/context overlap	Low-weight; divergent rankings
BioSentVec	Biomedical semantics	0.827 top Task 2
GPT-White	Reference coverage	0.827 top Task 2
GPT-Black	Validity/completeness	0.859 top Task 2
Experts	Clinical quality	0.816; ranking unchanged

Table 2: Complementary NTCIR-18 evaluation targets.

questionnaire responses while preserving explanation generation as the target; it contained 804 training and 216 evaluation cases. Multiple radiologists reviewed the structured causal annotations for accuracy and consistency. The diagnostic vocabulary and checklist make otherwise implicit reasoning observable at several linguistic and clinical levels.

3.2 Evaluation and pilot findings

Evaluation combined BERTScore (Zhang et al., 2020), report-level cosine similarity, BioSentVec (Chen et al., 2019), GPT-White, GPT-Black, and expert qualitative judgment. GPT-White compared system output with the ground-truth explanation under an external rubric. GPT-Black used an internal bonus–penalty scheme covering contextual similarity, finding–impression matching, disease coverage, causal appropriateness, logical consistency, and justification. The final score assigned 5% each to BERTScore and cosine similarity, 20% to BioSentVec, 50% jointly to the two GPT measures, and 20% to expert qualitative review.

Three teams submitted official runs. Expert review selected the top three runs under each of five automatic metrics and deduplicated them, yielding 18 Task-1 and 10 Task-2 runs for clinical assessment. Expert review did not change the final team ranking. The strongest Task-2 run obtained GPT-White 0.827, GPT-Black 0.859, and expert qualitative 0.816, whereas its BERTScore was only 0.099. This contrast illustrates the pilot's central finding: surface overlap can undervalue explanations whose content, coverage, and causal structure are judged strong.

4. NTCIR-19 HIDDEN-RAD2

4.1 Design changes motivated by the pilot

HIDDEN-RAD2 responds to four pilot limitations. First, a whole-text generation score did not show where a clinical error occurred; Task 2 therefore adds span localization. Second, a single reference explanation could not distinguish alternative wording from unsupported content; valid and controlled-invalid variants make evidential support the primary criterion. Third, global penalties did not characterize failure modes; the new taxonomy identifies ADD-PATH (unsupported pathology), ADD-DEVICE (unsupported device), NEG-FLIP (reversal of absence/presence), CONTRA (contradiction), and LAT-FLIP (laterality reversal). Fourth, automated judgment required stronger clinical grounding; new data pass through layered radiologist review.

Dimension	NTCIR-18	HIDDEN-RAD2
Core object	Generated explanation	Evidence-linked explanation
Task focus	Generation	Structured prediction + verification
Error unit	Whole text	Label, span, type, correction
Evidence	Report or A1–A4	Image, A1–A5, SCES
Expert role	Selective ranking review	Annotation + layered validation

Table 3: Evolution from HIDDEN-RAD to HIDDEN-RAD2.

4.2 Task 1: structured radiological reasoning

HIDDEN-RAD2 Task 1 predicts the A1–A5 interpretation directly from chest radiographs. A1 and A4 use a controlled vocabulary of 36 diagnostic labels; A2 encodes anatomical location and laterality; A3 uses T1–T12 ranges; and A5 records up to four groups of relevant checklist items. In the first 300-case release, 77 records have no final A4 diagnosis, including both pure-normal cases and cases where preliminary suspicion is not retained. Two Konyang University Hospital radiologists independently annotated cases and cross-reviewed one another's work before release. This double-annotation and reciprocal validation procedure explicitly separates preliminary candidates from final conclusions.

4.3 Task 2: causal verification and correction

Task 2 asks whether a machine-generated Causal Exploration paragraph is supported by the case evidence. For an invalid explanation, a system must detect the error, localize the offending span, classify the error type, and generate corrected text. The current corpus design uses 200 MIMIC-CXR cases split into 140 training and 60 evaluation cases; training is distributed in batches of 100 and 40. Seventeen IU X-Ray cases extend evaluation to 77 cases and add a second image source. Each base case is associated with valid and controlled-hallucination variants, enabling matched comparison while holding most linguistic context constant.

Data construction follows generate-then-validate. GPT-5.4 with vision generated A1–A5 structures, finding–impression pairs, Causal Exploration paragraphs, and controlled perturbations. Medical specialists then reviewed different layers. Two radiologists independently annotated and cross-reviewed Task 1. For the 17 IU X-Ray cases, a radiology professor reviewed finding–impression links, while a second radiologist independently reviewed injected hallucinations by type. Only expert-accepted perturbations were selected for the final benchmark subset.

5. Provenance-Aware Case Analysis

5.1 Annotation levels and relations

The source text is a medical report in the source database, produced from interpretation of a chest X-ray image. We distinguish three levels. A *section label* records where text occurs (e.g., FINDINGS or IMPRESSION); a *discourse-function label* records whether a unit functions as symptom, finding, impression, or recommendation; and an *evidence-provenance label* records whether a proposition comes from the report, image-derived A3 localization, or A5 confirmation. Thus a sentence inside the IMPRESSION section may function as a recommendation, and a fact absent from the report may still be grounded in expert-annotated image evidence.

We use short local identifiers such as `f1`, `i1`, and `g1` within each case. The released corpus should use globally unique keys such as `c1-f1` and `c2-f1`. Relations are directed from evidence to the unit it supports. The main inventory is ELABORATION, EVIDENCE–SUMMARY, EVIDENCE–INTERPRETATION, CONFIRMATION, ASSOCIATION, MOTIVATES–RECOMMENDATION, and JOINT–

SUMMARY. Rhetorical well-formedness and evidential validity are annotated separately: an explanation can read coherently while relying on the wrong source or an unsupported inference.

5.2 Case 1: mass and possible lymphadenopathy

The report describes a 4.6 cm × 3.7 cm rounded right-middle-lobe mass (`f3`), mild surrounding airspace disease and/or atelectasis (`f4`), and prominent right paratracheal soft-tissue density (`f2`). The impression (`i1`) retains the mass and interprets the paratracheal density as possible associated lymphadenopathy; `i2` recommends chest CT. The report-level graph is:

$$\texttt{f3} \xrightarrow{\textsc{Elaboration}} \texttt{f4}, \quad \{\texttt{f2}, \texttt{f3}\} \xrightarrow{\textsc{Evidence}} \texttt{i1} \xrightarrow{\textsc{Motivation}} \texttt{i2}.$$

The gold Causal Exploration adds T4–T6 localization from A3 and preserved hilar contour/visible vessels from A5. Its strongest hidden bridge is `f2+i1` → possible lymphadenopathy; the A5 hilar observations constrain this nodal interpretation as subtle rather than markedly distorting the hila. The conclusion summarizes a right-middle-lobe mass with possible mild paratracheal lymphadenopathy. This case separates three operations: restating the measured mass, interpreting a density as possible nodal enlargement, and motivating additional CT. It also shows why "surrounding parenchymal reaction" must remain hedged: spatial association is not proof of biological causation.

5.3 Case 2: effusion and cardiomegaly

The second report states that the cardiac silhouette is enlarged (`f1`), describes left-greater-than-right bibasilar opacities with a small right effusion (`f2`), and excludes pneumothorax (`f3`). The impression (`i1`) summarizes bibasilar airspace disease with "small effusions." The gold explanation draws on costophrenic-angle blunting from A5 and T8–T12 localization from A3 to explain dependent pleural fluid. A second branch maps enlarged cardiac silhouette to the normalized diagnostic label *cardiomegaly*. The two branches converge in a joint summary:

$$\begin{aligned} \texttt{g1} &\xrightarrow{\textsc{Localization}} \texttt{g2} \xrightarrow{\textsc{Association}} \texttt{g3}, \\ \texttt{f1} &\xrightarrow{\textsc{Interpretation}} \texttt{g4}, \\ \{\texttt{g1}, \ldots, \texttt{g4}\} &\xrightarrow{\textsc{Joint–Summary}} \texttt{g5}. \end{aligned}$$

This case exposes a number/laterality issue: the Findings section directly asserts a right effusion, whereas the Impression uses an unspecified plural and the gold text proposes both pleural spaces. Bilaterality is fully grounded only if the relevant A5 values support both sides. The final coordination of cardiomegaly and effusions does not by itself assert that one caused the other.

Both cases instantiate report finding + A3 localization + A5 confirmation ⇒ explicit evidence-to-impression explanation. They are nevertheless complementary. Case 1 demonstrates interpretive reclassification and a recommendation link; Case 2 demonstrates parallel evidence branches and a report-to-gold shift in number and laterality. Together they establish a central annotation principle: *report-new* does not mean *unsupported*, but every new claim must be traceable to the declared evidence bundle.

5.4 From gold explanation to verification item

The same representation supports Task 2 perturbation and repair. In a right-upper-lobe pneumonia case, A1/A4=pneumonia, A2=right-upper-lobe parenchyma, A3=T1–T4, and A5=Q7/Q8. The valid explanation links focal airspace disease and increased upper-zone density to a localized infectious process. An ADD-PATH variant inserts "with a small central cavitary component." Because no cavity is supported by the report or structured evidence, the gold correction deletes exactly that span. The instance consequently supports generation, evidence checking, span localization, error typing, and correction.

Dimension	Case 1	Case 2
Core bridge	Density → possible lymphadenopathy	Silhouette → cardiomegaly; basilar signs → effusion
A3	T4–T6	T8–T12
A5	Zonal/hilar confirmation	Costophrenic and pleural confirmation
Graph	Main chain + CT	Parallel branches + joint summary
Risk	Association read as cause	Number/laterality drift

Table 4: Complementary discourse phenomena in the two cases.

6. Evaluation Beyond Surface Similarity

6.1 Expert comparison of GPT-derived judgments

The reviewer requested evidence that GPT-based assessment is reliable relative to expert judgment. The available evidence is complementary rather than a single correlation coefficient. In NTCIR-18, experts examined 28 unique top-ranked runs across the two tasks using criteria aligned with GPT-White and GPT-Black; their qualitative judgments left the final ranking unchanged. The Task-2 top run's GPT and expert scores were close (0.827/0.859 versus 0.816), while its much lower BERTScore demonstrated why expert-aligned content measures were needed.

HIDDEN-RAD2 provides item-level evidence from the 17-case extension. Of 57 generated finding–impression pairs, 32 received an explicit accept/reject decision: 28 were accepted and four rejected, an 87.5% acceptance rate among adjudicated pairs. The rejected links were over-inferences, including treating projection or rotation effects as pathology and asserting stability without a valid comparison. The remaining 25 pairs were unlabelled rather than clinically rejected and are excluded from the denominator.

For injected hallucinations, the expert evaluated 68 candidates (17 cases × four tested types): 41 good (60.3%), 22 neutral/revision-needed (32.4%), and five discard (7.4%). ADD-DEVICE was accepted in 17/17 cases and ADD-PATH in 16/17. NEG-FLIP accounted for all five discards, usually because an elaborate statement of a nonexistent finding sounded unlike a real radiology report. CONTRA produced 16 neutral judgments, indicating that contradiction generation needed refinement. These type-specific results show both high reliability in selected categories and systematic limits that automated generation alone would conceal.

6.2 Interpretation and limits

These comparisons justify GPT-based metrics as scalable auxiliary evaluators, not replacements for radiologists. The pilot establishes ranking-level consistency on selected systems; HIDDEN-RAD2 establishes item- and error-type-level acceptance. Neither study yet provides a fully crossed, independently replicated rating matrix sufficient for robust inter-rater correlation or confidence intervals. Future re-

Expert comparison	n	Result
NTCIR-18 selected runs	28	Final ranking unchanged
Finding–impression pairs	32 judged	28 accepted (87.5%)
ADD-DEVICE	17	17 good
ADD-PATH	17	16 good, 1 neutral
NEG-FLIP	17	7 good, 5 neutral, 5 discard
CONTRA	17	1 good, 16 neutral

Table 5: Expert evidence for generated links and perturbations.

leases should have experts score a stratified random sample blindly, report Cohen's kappa or weighted kappa for categorical/rubric judgments, and estimate Spearman rank correlation with bootstrap intervals for continuous metrics.

6.3 Why verification changes evaluation

A global score can hide one clinically significant clause inside an otherwise excellent paragraph. Span and category annotations expose that clause and make disagreement reproducible. Correction further distinguishes detection from constructive repair. In this sense, HIDDEN-RAD2 changes the evaluation object from "How similar is this text?" to "Which claim is supported, where does support fail, what kind of failure is it, and how can it be repaired?"

7. Discussion

7.1 Corpus-linguistic contribution

HIDDEN-RAD externalizes a latent expert discourse relation as an annotation layer. Report prose, diagnostic questionnaires, image-derived structures, and causal explanations represent complementary registers of specialized reasoning. The controlled A1–A5 fields provide stable anchors, while explanation text preserves variation in how clinical relations are verbalized. Matched valid/invalid variants further support corpus-based analysis of laterality, negation, contradiction, evidentiality, and causal attribution in AI-generated medical language.

7.2 Medical-AI contribution

The framework shifts explainability from fluent rationale production toward evidence-grounded accountability. A model may generate, a second process may verify and localize, and an expert may adjudicate high-risk or ambiguous cases. The case analysis shows why report entailment and dataset grounding must be separated: thoracic level, hilar preservation, or costophrenic blunting may be absent from prose yet supported by A3/A5 image annotation. Conversely, a fluent relation such as mass → parenchymal reaction remains an inference rather than an observed fact and must preserve uncertainty. Structured evidence and relation labels together constrain plausible but unsupported elaboration.

7.3 Limitations and ethics

The work is restricted to chest radiography and inherits the documentation patterns and biases of MIMIC-CXR and IU X-Ray. Annotation is costly and medically contestable, especially for implicit or multifactorial causality. Current expert comparisons are selective, and GPT behavior depends on model version and prompts. In seven of the 17 IU cases, a specialist revised the structured checklist but the original GPT Causal Exploration was not regenerated; acceptance rates therefore characterize the reviewed GPT output, not complete propagation of every checklist correction through the pipeline. The benchmark is intended for research, not autonomous diagnosis; de-identified source data remain subject to their original access and licensing conditions. Public release should include data cards, model/prompt provenance, and clear separation between expert-authored, model-generated, and expert-validated fields.

8. Conclusion

RQ1 is addressed by representing diagnostic reasoning as A1–A5 evidence plus typed evidence–inference–impression relations. The two cases show that provenance must distinguish report-stated, image-localized, checklist-confirmed, and generated-inference content. RQ2 is addressed through generation from reports, structured questionnaires, and images. RQ3 is addressed by combining semantic metrics with rubric- and penalty-based LLM judgments, expert review, and—at NTCIR-19—validity, span, type, and correction annotations. Across the two shared tasks, HIDDEN-RAD evolves from a generation benchmark into a provenance-aware reliability framework in which medical explanations can be generated, localized, verified, and corrected against structured clinical evidence.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2022-II220078, Explainable Logical Reasoning for Medical Knowledge Generation).

Bibliographical References

Chen, Q., Peng, Y., and Lu, Z. (2019). BioSentVec: Creating sentence embeddings for biomedical texts. Proceedings of the IEEE International Conference on Healthcare Informatics, 1–5.

Cho, J.-M., Yi, H.-J., Kim, M.-K., Jeong, S.-J., and Na, S.-H. (2025). Optimizing causality-based radiology reporting with retrieval-augmented and structured reasoning approaches for the NTCIR-18 HIDDEN-RAD Task. Proceedings of NTCIR-18.

Choi, K.-S., Cho, Y.-S., and the HIDDEN-RAD Organizing Committee. (2025). Overview of the NTCIR-18 HIDDEN-RAD task. Proceedings of NTCIR-18.

Johnson, A., Pollard, T., Mark, R., Berkowitz, S., and Horng, S. (2024). MIMIC-CXR Database (version 2.1.0). PhysioNet. https://doi.org/10.13026/4jqj-jw95.

Ranjit, M., Kumar, R., Srivastav, S., Porya, A., and Ganu, T. (2025). RAD-PHI3 at the NTCIR-18 HIDDEN-RAD. Proceedings of NTCIR-18.

Won, Y., Hahm, Y., Yoon, C., and Kim, S. T. (2025). Teddysum at the NTCIR-18 HIDDEN-RAD Task: Using RAG and Tree-of-Thought for causal explanation generation. Proceedings of NTCIR-18.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. International Conference on Learning Representations.