Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated Intracranial Aneurysm Detection and Size Extraction from Imaging Reports Using a Dual-Model Transformer Framework

Syed Hasib Akhter Faruqui
Sam Houston State University
Amr R. Alwakeal
Northwestern University
Dillan Prasad
Northwestern University
Rishi Jain
Northwestern University
Michael Oblich
Northwestern University
Nikhil Prasad
Northwestern University
Emily Winterhalter
Northwestern University
Donald Cantrell
Northwestern University
Matthew Potts
matthew.potts@nm.org
Northwestern University

Research Article

Keywords: Large Language Model, Cranial Aneurysm Detection, Endovascular Imaging Reports, Natural Language Processing (NLP), Neurosurgery, Radiology

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11435195/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Automated Intracranial Aneurysm Detection and Size Extraction from Imaging Reports Using a Dual-Model Transformer Framework

Syed Hasib Akhter Faruqui*, PhD[1], Amr R. Alwakeal*, MD[2], Dillan Prasad, MS[2], Rishi Jain, BA[2], Michael Oblich, BS[2], Nikhil Prasad[2], Emily Winterhalter[2], MD, Donald Cantrell[3], MD, PhD, Matthew Potts, MD[2]

Affiliations:

1. Department of Engineering Technology, Sam Houston State University, 1019 Bowers Boulevard, Huntsville, TX 77341-2088, United States
2. Department of Neurological Surgery, Northwestern University Feinberg School of Medicine, 676 N. St. Clair Ave, Suite 2210, Chicago, IL 60611, United States
3. Department of Radiology, Northwestern University Feinberg School of Medicine, 676 N. St. Clair Ave, Suite 800, Chicago, IL 60611, United States

Corresponding Author:
Matthew Potts, MD
Associate Professor, Neurological Surgery
Director, Neurological Surgery Residency Program
Department of Neurological Surgery
Northwestern University Feinberg School of Medicine
676 N Saint Clair Ave
Chicago, IL 60611, United States
Email: matthew.potts@nm.org

Keywords: Large Language Model, Cranial Aneurysm Detection, Endovascular Imaging Reports, Natural Language Processing (NLP), Neurosurgery, Radiology
*Authors contributed equally.

Abstract:
Background: The unstructured and variable nature of radiology reports remains a barrier to automated extraction of clinically relevant information for large-scale research and AI applications, particularly for intracranial aneurysms (IAs).
Objective: We aimed to develop and evaluate a dual-model transformer framework for intracranial aneurysm information extraction and assess whether structured processing improved classification and size-extraction performance compared with unstructured text.
Methods: We analyzed 3,183 neuroimaging reports annotated by neuroradiologists and neurosurgeons. Four encoder-based transformer models were evaluated for aneurysm-presence classification using 5-fold stratified cross-validation. A zero-shot Llama 3.1 framework was used to extract aneurysm size using direct and structured-report prompting. RadDistilBERT and Llama 3.1 were evaluated on 20 temporally distinct reports using full-report and impression-only inputs, with and without structured preprocessing.
Results: RadDistilBERT achieved the highest mean classification performance, with accuracy of 96.4% (±1.8%), F1 score of 91.7% (±10.3%), precision of 94.0% (±4.6%), and recall of 90.7% (±15.6%). F1 scores were 87.5% for BERT-base-uncased and 31.1% for ALBERT-base-v2. In the 20-report set, aneurysm-presence classification accuracy was 100% across all input and processing conditions. For aneurysm size extraction, structured report processing increased accuracy from 61.90% to 66.67% using the full report and from 80.95% to 85.00% using the impression section.
Conclusions: This study demonstrates the feasibility of a dual-model transformer framework for aneurysm-presence classification and size extraction from unstructured neuroimaging reports. RadDistilBERT achieved the strongest classification performance, while Llama 3.1 enabled zero-shot size extraction. These findings support further external validation and expansion to additional aneurysm characteristics with dedicated annotation and validation.

1.0 Introduction

Intracranial aneurysms (IAs) are commonly encountered in endovascular and interventional radiology due to their clinical significance(1-3). Their evaluation relies heavily on imaging modalities such as computed tomography angiography (CTA), magnetic resonance angiography (MRA), and digital subtraction angiography(4, 5), which serve as the gold-standard modality.

However, radiology reports describing IAs are typically written in free text, with substantial variability in terminology and structure (see **Figure 1**). As a result, these reports vary between institutions and even between providers within the same institution(6-8). Manual review and extraction of key clinical findings is time-consuming and prone to inconsistencies. These challenges are especially relevant to retrospective chart reviews and constructing datasets for research and clinical registries. In particular, identifying and aggregating aneurysm cases from large archives of imaging reports requires significant labor, limiting the scalability of large-scale clinical and research analyses.

These challenges have driven growing interest in using natural-language processing (NLP) and transformer-based language models to automate the interpretation and summarization of radiological reports(7, 8). Recent advances in LLM-based systems have demonstrated their ability to extract clinically relevant information from unstructured text with high accuracy(9, 10). Models such as BioBERT, ClinicalBERT, and GPT-derived architectures excel at recognizing domain-specific terminology, including anatomical structures and pathological features, thereby enabling efficient data extraction, standardization, and summarization to improve diagnostic workflows(7, 8). Despite these promising developments, a gap remains in applying LLMs specifically to aneurysm characterization and reporting. In particular, the use of instruction-following generative models for zero-shot extraction of aneurysm characteristics remains insufficiently studied.

This study introduces a dual-model approach combining classification and information extraction for automated aneurysm analysis from radiology reports: 1) A fine-tuned classification model specifically designed for binary classification of radiology reports as aneurysm-positive or aneurysm-negative. Trained on a curated dataset of 3,183 de-identified reports, each annotated by clinical experts. 2) A zero-shot approach using **Llama 3.1** to extract clinically relevant details such as aneurysm size without task-specific fine-tuning.

Detailed prompts have been explored to improve the extraction of size information from the report. We implemented and tested two complementary strategies: a) Direct extraction from raw reports, and b) A structured breakdown of the report into semantically distinct sections prior to extraction. This allowed us to assess how reasoning-augmented prompts affect the accuracy and consistency of the model's outputs.

Notably, we perform all information extraction without task-specific fine-tuning, showcasing the flexibility and power of contemporary instruction-following LLMs. Together, these models aim to streamline aneurysm case identification, reduce reporting variability, and provide scalable AI-driven solutions for IA analysis in diverse clinical and research settings. Figure 2 shows the overall work procedure and the proposed methodology.

2.0 Methods

2.1 Ethical Considerations

The Northwestern University Institutional Review Board (IRB) reviewed and approved this retrospective study (protocol STU00204913). The approval preceded study-specific data extraction, de-identification, model development, evaluation, and statistical analysis. The institutional review board waived the requirement for written informed consent. Patient identifiers were removed before analysis in accordance with institutional privacy procedures and applicable Health Insurance Portability and Accountability Act of 1996 (HIPAA) requirements. All study methods were performed in accordance with the applicable institutional guidelines and relevant regulations, including HIPAA requirements governing the protection of patient health information.

2.2 Data Acquisition

A retrospective dataset of 3,183 de-identified imaging reports was obtained from the Northwestern Medicine Enterprise Data Warehouse (NM-EDW). Inclusion criteria were age >18 years, evaluation by the neurovascular service, and availability of a neurovascular imaging report assessing intracranial vascular pathology, irrespective of aneurysm status. The dataset spanned 2010 to 2023 and included CT, MRI, and digital subtraction angiography reports. Other forms of clinical free-text documentation were excluded. Among the reports, 2,476 (77.8%) contained at least one intracranial aneurysm, while the remaining 707 (22.2%) were aneurysm-negative.

The aneurysm-positive reports encompassed a wide range of vascular territories, including both anterior and posterior circulations, and treatment statuses (clipped, coiled, stented, and untreated). This heterogeneity reflected the diversity of clinical presentations and imaging findings encountered in neurovascular practice. Demographic information, including age and sex, was retained only in aggregate for descriptive purposes and was not incorporated into model training, ensuring that classification relied exclusively on report text.

2.3 Data Annotation and Pre-processing

A multidisciplinary team of neuroradiologists, neurosurgeons, and clinical informatics experts manually reviewed and annotated each report. The primary annotation objective was to classify reports as aneurysm-positive or aneurysm-negative based on explicit or inferred evidence of intracranial aneurysms (IAs). For aneurysm-positive reports, lesion size (recorded in millimeters or centimeters) was also captured. Annotators were instructed to consider both explicit mentions of aneurysms and inferred cases based on clinical language (e.g., “clip in place,” “treated aneurysm”) while excluding incidental mentions unrelated to the intracranial vasculature. Reports with conflicting interpretations between annotators were flagged and adjudicated during structured consensus meetings. Final labels were determined by agreement of at least two senior clinical reviewers following consensus adjudication. Supplementary **Appendix 1** provides a detailed annotation guide of aneurysm characteristics.

For final input data processing, two strategies were employed: (i) using only the impression section of the report, which provides a concise summary of clinical findings; and (ii) using the entire report to extract comprehensive information for more detailed analysis. Tokenization was performed using relevant tokenizers tailored to selected models.

2.4 Model Architecture

This study employs a two-stage approach:

(a) **Aneurysm classification:** We used RadDistilBERT, a study-specific classifier fine-tuned from DistilBERT-base-uncased, for binary intracranial aneurysm classification. Three additional encoder-based transformer models (RoBERTa-base-SQuAD2, BERT-base-uncased, and ALBERT-base-v2) were independently fine-tuned on annotated radiology reports for comparative evaluation.

(b) **Structured report generation and aneurysm size extraction:** Llama 3.1 (8B) was used without task-specific fine-tuning to generate structured representations and extract aneurysm measurements from complete radiology reports using direct and structured extraction workflows.

The classification stage utilizes fine-tuned transformer models, while the extraction stage leverages a large language model's instruction-following capabilities.

2.4.1 Classification Models

The following models were used to finetune and develop the report classifier.

RadDistilBERT (66M parameters): Our specialized fine-tuned version of DistilBERT-base-uncased, which is a smaller, faster, and lighter version of BERT. DistilBERT retains 97% of BERT's performance while being 60% smaller and 60% faster. We fine-tuned this model specifically on our radiology report dataset, creating RadDistilBERT, for binary intracranial aneurysm classification(11).

RoBERTa-base-SQuAD2 (125M parameters): A variant of RoBERTa (Robustly Optimized BERT Pretraining Approach) fine-tuned on the SQuAD 2.0 dataset for question-answering tasks and subsequently adapted for binary classification of intracranial aneurysm status in radiology reports.

BERT-base-uncased (110M parameters): The foundational bidirectional transformer model that reads text in both directions simultaneously. BERT serves as a versatile backbone for many NLP tasks, including text classification, named entity recognition, and question answering(12).
ALBERT-base-v2 (A Lite BERT) (11M parameters): Google's parameter-efficient alternative to BERT that uses factorized embedding parameterization and cross-layer parameter sharing to achieve similar performance with fewer parameters. This compact variant is particularly valuable when computational resources are limited(13).

2.4.2 Information Extraction with Llama 3.1

This study employs **Llama 3.1** for two primary tasks.
First, we use Llama 3.1 to transform complete radiology reports into structured, meaningful information organized into six categories: abnormal findings, diagnoses, clinical relevance, reasoning, clinical history, and aneurysm characteristics. This structured information was subsequently used as an alternative input to our classification model. **Table 1** shows one such example of an original report and its corresponding extracted structured report, while **Table 2(a)** presents the prompt used to generate the structured information.
Second, we implement two distinct prompting strategies for aneurysm size extraction: (a) Direct extraction employs a straightforward prompt that instructs the model to identify specific information such as lesion size directly from the radiology report. (b) Structured breakdown uses the structured representation generated from the original report using the prompt presented in **Table 2(a).** This is the same prompt used to generate structured information for the classification model. The resulting structured representation is then processed using the aneurysm size extraction prompt presented in **Table 2(b).**
Importantly, our implementation leverages **LLaMa 3.1**'s instruction-following capabilities without requiring fine-tuning, demonstrating the model's inherent ability to adapt to specialized medical tasks and extract clinically relevant information from radiology reports.

2.4.3 Training Configuration

All classification models were fine-tuned using a learning rate of 1×10^{-5}, a batch size of 8, five training epochs, and the AdamW optimizer with a weight decay of 0.01.
Model performance was evaluated using fivefold patient-level stratified cross-validation, with all reports from the same patient assigned to the same fold to prevent information leakage between training and validation sets. The best-performing checkpoint was selected based on validation performance, and class balancing was applied during model development to address the unequal distribution of aneurysm-positive and aneurysm-negative reports.
Training and inference were conducted on an NVIDIA RTX 3060 GPU with 12 GB of GPU memory and a system equipped with 32 GB of RAM.

3.0 Results

3.1 Classification Performance

Table 3 presents the 5-fold cross-validation results for all classification models. Performance was assessed using accuracy, F1 score, precision, and recall. Among the tested models, RadDistilBERT achieved the strongest overall performance, with a mean accuracy of 96.4% (±1.8%), F1 score of 91.7% (±10.3%), precision of 94.0% (±4.6%), and recall of 90.7% (±15.6%). It also showed the lowest variability across folds, with a standard deviation of ±1.8% in accuracy. BERT-base-uncased ranked second, with a mean accuracy of 93.2%. RoBERTa-base-SQuAD2 showed moderate performance, with an accuracy of 70.8%, whereas ALBERT-base-v2 had the lowest performance among the evaluated models, with an accuracy of 31.7%.

Overall, these results identified RadDistilBERT as the best-performing classifier for binary classification of angiography reports as aneurysm-positive or aneurysm-negative in this dataset.

3.2 Out-of-Distribution Evaluation and Size Extraction Performance

After model selection, we evaluated the fine-tuned RadDistilBERT classifier on an independently collected out-of-distribution [12] dataset of 20 previously unseen reports. Special care was taken to make sure these reports are collected separately compared to the training data. Table 4 summarizes the OOD classification ($N = 20$ reports) results and the corresponding performance of the Llama 3.1 zero-shot extraction framework.

When the full report text was used as input, both approaches achieved perfect classification accuracy (100%). The structured information extraction approach showed measurable improvements when processing abbreviated report sections. The aneurysm size prediction improved by 4.77 percentage points (from 61.9% to 66.67%) when structured breakdown has been provided for full report. In contrast, when this is done only for the impression section, the models' overall performance improves. The size extraction accuracy improved again by 4.05 percentage points (from 80.95% to 85%). These improvements, while modest in absolute terms, represent meaningful enhancements in clinical scenarios where only summary sections of reports are available or when processing efficiency is prioritized over comprehensive analysis.

3.3 Error Analysis

The error analysis identified two main sources of discrepancies: (i) ambiguities in report phrasing (such as overlapping terminology between aneurysmal and non-aneurysmal vascular findings) and (ii) inconsistent representation of numerical data (e.g., mixed units and partial descriptors). Notably, the structured segmentation strategy (extracting details from raw reports) successfully mitigated many of these issues by isolating the “impression” section, a part of the report where radiological findings are most standardized.

4.0 Discussion

We have applied a dual-model transformer framework for extracting intracranial aneurysm classification and characterization (aneurysm size) from DSA angiography reports. By integrating a fine-tuned, domain-adapted classifier (RadDistilBERT) with a zero-shot LLM pipeline (Llama 3.1), our model leverages the complementary strengths of both approaches to address the inherent variability, redundancy, and complexity of clinical text. This framework not only demonstrated state-of-the-art performance in distinguishing aneurysm-positive from aneurysm-negative reports but also robustly extracted aneurysm size, a key attribute that is often inconsistently documented in free-text radiology reports(6, 14, 15).

Notably, the combination of domain-specific fine-tuning and zero-shot generalization illustrates a scalable strategy for radiology NLP: RadDistilBERT ensures precision and reliability in classification, while Llama 3.1 enhances adaptability and interpretability in structured information extraction. Together, these results highlight the potential of hybrid transformer pipelines to overcome limitations of single-model systems and establish a foundation for clinically deployable, AI-driven report analysis(16).

The structured information extraction (Using pre-trained non-fine-tuned Llama 3.1) approach employed in this study allows for a more comprehensive understanding of patient reports, enabling the development of more accurate classification models. This not only supports the development of more accurate classification models but also offers practical benefits in settings with limited annotated data. Moreover, integrating clinical rules and guideline-based logic into this framework can facilitate real-time risk stratification, allowing for rapid identification and categorization of aneurysms, thereby enhancing patient safety. Standardizing the extraction and formatting of radiological findings also holds promise for harmonizing terminology across institutions(17). This can improve interoperability and foster clearer communication between radiologists and referring physicians, ultimately contributing to more coordinated and effective patient care.

4.1 Clinical Utility

The clinical applicability of our framework is manifold. First, an automated classifier with high sensitivity and specificity for aneurysm detection can reduce manual burden and improve turnaround time for clinical reports, thereby enabling clinicians to focus on complex cases that require expert judgment. Second, the reliable extraction of aneurysm attribute information is a critical step toward creating structured databases for longitudinal studies, risk stratification, and decision support in interventional planning. Third, the efficiency gains from using a dual-model approach are particularly beneficial in settings where annotated data are limited, as the system leverages both fine-tuning on a moderate dataset and the adaptable zero-shot capabilities of large general-purpose LLMs.

4.2 Limitations and Challenges

This study has several significant limitations. First, all training and evaluation data were obtained from a single institution. While this provided a large and clinically rich dataset, it may limit the generalizability of our findings to other healthcare systems with different reporting practices, patient populations, imaging protocols, and physician documentation styles. Future studies should prioritize collecting multi-institutional datasets to increase data diversity and improve external validity(18).

Second, the evaluation of structured information extraction was performed on a relatively small out-of-distribution sample (N = 20). Although these reports were drawn from a temporally distinct cohort, they were still sourced from the same institution and thus may not capture the full variability of clinical language encountered across broader practice settings.

Third, the present study was limited to binary classification of reports as aneurysm-positive or aneurysm-negative and extraction of aneurysm size, while other aneurysm characteristics, such as location, were not evaluated. We intentionally deferred location extraction because anatomical sites are described heterogeneously across reports, including variation in vessel nomenclature, segment designation, laterality, and treatment-related terminology. Future work will require the development and validation of a standardized location schema, with extracted terms mapped to controlled clinical vocabularies such as SNOMED CT or RadLex before location-specific analyses are undertaken. Such standardization may improve consistency and interoperability, although its reliability will require dedicated annotation and external validation.

Fourth, Llama 3.1 may occasionally generate outputs that are clinically plausible but factually incorrect, particularly when faced with ambiguous or incomplete radiology reports(19, 20). This underscores the need for careful validation and the incorporation of human-in-the-loop oversight to ensure safe and reliable clinical deployment. Together, these limitations highlight the importance of external validation, larger multi-institutional studies, and the development of hybrid systems that combine automated extraction with expert review to maximize both efficiency and clinical reliability.

4.3 Ethical Considerations and Responsible AI

While these models are helpful in supporting practitioners with quick outcomes, these systems should be evaluated for potential biases. Specifically, the training data should be well balanced to avoid any ill-informed conclusions related to patients, integrate institution-specific practices, or reporting styles that could lead to dissimilar outcomes.

5.0 Conclusion

This study presents a novel dual-model framework for automated IA detection and their size determination from radiology reports. By combining a domain-specific, fine-tuned classification model, RadDistilBERT, with the zero-shot instruction-following capabilities of large language models such as Llama 3.1, the approach leverages the complementary strengths of both models. Our findings demonstrate the potential of this architecture to enhance efficiency, consistency, and scalability in aneurysm case identification, providing a robust foundation for automating a critical, labor-intensive component of radiological interpretation. Even while generating the initial data for fine-tuning the annotators had to spend an excruciating amount of time reading all the reports and determining aneurysm class and size. While this study focuses on binary classification only, if the classifications were subdivided, the annotations would have taken longer than expected. The fine-tuned RadDistilBERT model consistently outperformed other fine-tuned models across classification metrics (Table 3), underscoring the value of domain adaptation in medical LLMs. In parallel, the zero-shot capability of Llama 3.1 enabled flexible and interpretable extraction of aneurysm-related details, including size, even from unstructured or variably formatted reports without the need for task-specific retraining. This dual system supports a practical pipeline for clinical or research deployment in settings where annotated data is sparse or where reporting conventions may vary significantly across institutions. Beyond technical performance, the clinical relevance of this work lies in its potential to reduce the manual burden of retrospective chart reviews and enhance consistency in information extraction across diverse healthcare systems. This system can be used as a secondary screening layer to support quality assurance by flagging potentially missed cases, thereby contributing to patient safety and audit readiness. Future directions, including multi-institutional validation, multimodal integration, refined prompt engineering, and enhanced model interpretability, will further solidify the role of these AI-driven tools in the clinical domain. Ultimately, the work described in this report lays a strong foundation for the development of scalable, automated solutions for neurovascular imaging analysis, with the potential to transform IA management and improve outcomes for patients worldwide.

Acknowledgments
Generative AI tools were used to assist with language editing, grammar correction, and organization of selected revisions. They were not used for study design, data collection, diagnostic labeling, reference-standard establishment, statistical analysis, figure or table generation, or scientific interpretation. The authors reviewed and verified all AI-assisted content.

Funding
No funding was received for this work.

Conflicts of Interest
None declared.

Ethics approval
The Northwestern University Institutional Review Board (IRB) reviewed and approved this retrospective study (protocol STU00204913). Patient identifiers were removed before analysis in accordance with institutional privacy procedures and applicable HIPAA requirements.

Consent to participate
The Northwestern University Institutional Review Board (IRB) waived the requirement for written informed consent for this retrospective study.

Consent for publication
Not applicable. The manuscript does not contain identifiable individual patient information.

Data Availability
The datasets generated and analyzed during this study are not publicly available due to privacy and ethical restrictions related to clinical neuroimaging reports. Reasonable requests for de-identified data may be directed to the corresponding author and will be considered subject to institutional approval.

Authors' Contributions
AA and SF contributed equally to this work.
Conceptualization: AA, SF, MP
Data curation: AA, SF, DP, RJ, MO, NP, EW
Formal analysis: SF
Methodology: SF, AA
Supervision: MP, DC
Visualization: SF, AA
Writing – original draft: SF, AA, DP, RJ
Writing – review & editing: SF, AA, MP, DC, RJ, DP

Clinical Trial Number
Not Applicable.

References:

1. Vlak MH, Algra A, Brandenburg R, Rinkel GJ. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. Lancet Neurology. 2011;10(7):626–36.
2. Liu J, Zou X, Zhao Y, et al. Prevalence and risk factors for unruptured intracranial aneurysms in the population at high risk for aneurysm in the rural areas of Tianjin. Frontiers in Neurology. 2022;13:853054.
3. Ortiz AFH, Suriano ES, Eltawil Y, et al. Prevalence and risk factors of unruptured intracranial aneurysms in ischemic stroke patients: a global meta-analysis. Surgical Neurology International. 2023;14:222.
4. Allaw S, Khabaz K, Given TC, Montas D, Alcazar-Felix RJ, Srinath A, et al. A Review of Intracranial Aneurysm Imaging Modalities, from CT to State-of-the-Art MR. AJNR Am J Neuroradiol. 2025;46(6):1082–92.
5. Lim JH, Yoon DY, Kim ES, Jeon HJ, Lee JY, Seo YL, et al. CT Angiography, MR Angiography, and Their Combined Use for Detection of Unruptured Intracranial Aneurysms: Comparison with Digital Subtraction Angiography and 3-dimensional Rotational Angiography. Clin Neuroradiol. 2025;35(2):355–62.
6. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016;279(2):329–43.
7. Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. AJR American Journal of Roentgenology. 2017;208(4):754–60.
8. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging. 2019;49(4):939–54.
9. Li J, Lin Y, Zhao P, et al. Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (BERT) and in-domain pre-training (IDPT). BMC Medical Informatics and Decision Making. 2022;22(1):200.
10. Sugimoto K, Takeda T, Oh JH, Wada S, Konishi S, Yamahata A, et al. Extracting clinical terms from radiology reports with deep learning. J Biomed Inform. 2021;116:103729.
11. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv [Internet]. 2020. Available from: https://arxiv.org/abs/1910.01108.
12. Devlin J, Chang MW, Lee K, Toutanova K, editors. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers); 2019: Association for Computational Linguistics.
13. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv. 2019.
14. Greving JP, Wermer MJ, Brown RD, Jr., Morita A, Juvela S, Yonekura M, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. Lancet Neurol. 2014;13(1):59–66.
15. Jaiswal A, Tang L, Ghosh M, Rousseau JF, Peng Y, Ding Y. RadBERT-CL: Factually-Aware Contrastive Learning For Radiology Report Classification. Proc Mach Learn Res. 2021;158:196–208.
16. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9.
17. Pesapane F, Tantrige P, De Marco P, Carriero S, Zugni F, Nicosia L, et al. Advancements in Standardizing Radiological Reports: A Comprehensive Review. Medicina (Kaunas). 2023;59(9).
18. Suleman MU, Mursaleen M, Khalil U, Saboor A, Bilal M, Khan SA, et al. Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. Ann Med Surg (Lond). 2025;87(12):8803–11.
19. Belisle-Pipon JC. Why we need to be careful with LLMs in medicine. Front Med (Lausanne). 2024;11:1495582.
20. Roustan D, Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective With a Focus on Hallucinations. Interact J Med Res. 2025;14:e59823.

Figures & Figure Legends

Figure 1: Two Sample Radiology Reports: (a) with IA, (b) without IA

Aneurysm	No Aneurysm
1. a saccular aneurysm projects posteriorly from the right carotid terminus and measures a basetoapex of 4mm with a dome diameter of 6mm and a neck of 4mm. 2. there may be a small 2mm aneurysm at the origin of the right posterior communicating artery. there may also be mild 2mm fusiform dilation of the right mca bifurcation extending into the proximal anterior division. 3. perfusion abnormality corresponding to a small old infarct in the anterior left parietal lobe. no mr perfusion evidence of an acute infarct or penumbra. 4. no highgrade flowlimiting stenosis is otherwise identified in the major arterial vasculature identified of the head or neck.	there is no acute intracranial hemorrhage, midline shift or hydrocephalus. the gray and white matter differentiation is preserved. there are a few scattered foci of t2/flair hyperintensity in the subcortical and white matter of both frontal lobes, nonspecific. normal expected intracranial flow voids are seen. no areas of abnormal enhancement or diffusion restriction are seen. the bone marrow of the upper cervical spine is slightly heterogeneous, however, likely within normal limits for an adult male. the remaining midline craniocervical structures are within normal limits. there is minimal mucosal thickening of the frontal sinuses and ethmoid air cells, without associated airfluid level.

Figure 2: The overall proposed methodology procedure.

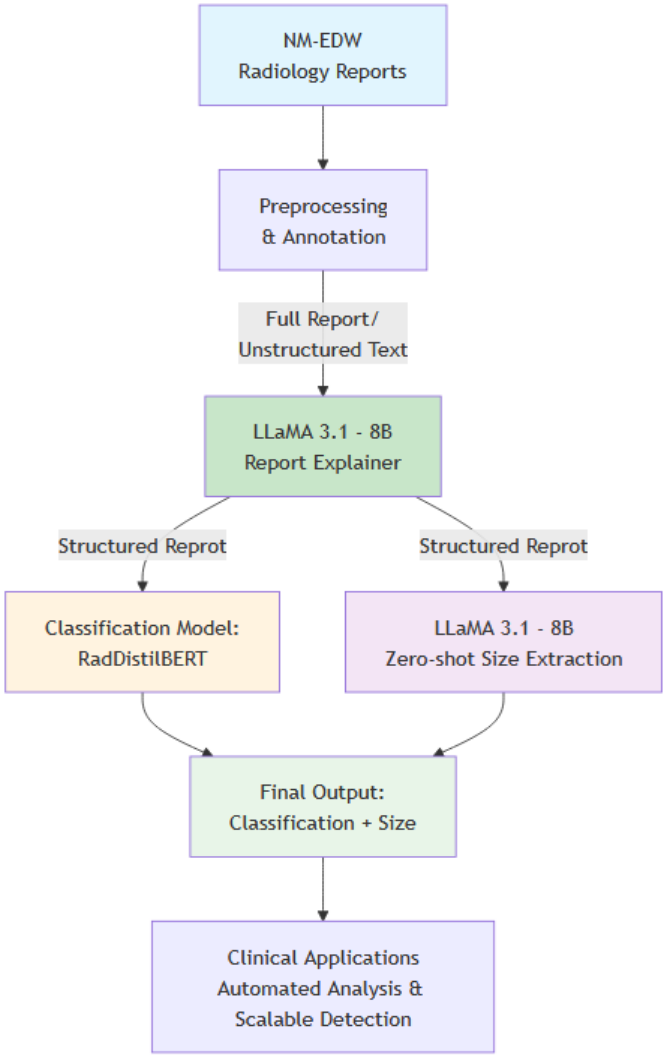


Table 1: ***Illustrative Example of Structured Information Extracted From a Radiology Report Using LLaMA 3.1***

Original report	**Extracted Information using LLaMa3.1**
PROCEDURE: CT Brain WO Contrast. **HISTORY**: History of metastatic lung carcinoma with incontinence and possible seizure. **COMPARISON**: Head CT of _____ and brain MRI of ____ **TECHNIQUE**: Contiguous axial images were obtained from the skull base to the vertex. Soft tissue, blood and bone windows reviewed. **FINDINGS**: Mild global volume loss and confluent supratentorial white matter disease is likely related to treatment effects. Postsurgical changes on a left frontotemporal craniotomy are noted and there is some encephalomalacia in anterolateral left frontal lobe on that basis. Scattered areas of dystrophic calcification are identified. No acute hemorrhage. No midline shift. No hydrocephalus. No significant change within the limits of comparing this CT to the recent MRI exam. There is a known aneurysm projecting superior/posteriorly off the proximal right superior cerebellar artery. **IMPRESSION**: No overt CT evidence of acute intracranial process. Confluent white matter lucency likely relates to post-treatment effect. A repeat MRI would be needed to exclude more subtle tumor recurrence or progression. Known proximal right SCA aneurysm.	**1. Abnormal findings**: ***a. Mild global volume loss***: **Size**: Not specified **Location**: Global (entire brain) **Changes over time**: Likely related to treatment effect **Associated descriptions**: None ***b. Confluent supratentorial white matter disease***: **Size**: Not specified **Location**: Supratentorial region **Changes over time**: Likely related to treatment effect **Associated descriptions**: None ***c. Postsurgical changes on a left frontotemporal craniotomy***: **Size**: Not specified **Location**: Left frontotemporal region **Changes over time**: Noted (postsurgical) **Associated descriptions**: Encephalomalacia in anterolateral left frontal lobe ***d. Scattered areas of dystrophic calcification***: **Size**: Not specified **Location**: Scattered throughout the brain **Changes over time**: None mentioned **Associated descriptions**: None ***e. Known aneurysm projecting superior/posteriorly off the proximal right superior cerebellar artery***: **Size**: Not specified (history of aneurysm size is mentioned, but not for final size determination) **Location**: Right superior cerebellar artery **Changes over time**: None mentioned **Associated descriptions**: None **2. Specific conditions or diagnoses:** a. Treatment effect likely causing mild global volume loss and confluent supratentorial white matter disease b. Postsurgical changes on a left frontotemporal craniotomy with encephalomalacia in anterolateral left frontal lobe c. Dystrophic calcification **3. Clinical relevance:** The patient's symptoms or the reason for the examination is not explicitly stated, but the findings suggest that the patient has undergone treatment and surgery. **4. Reasoning:** a. Based on the report, it appears that the patient has undergone significant medical interventions (surgery and treatment) which have resulted in various changes to their brain structure. b. The presence of aneurysm is noted, but its size and significance are not explicitly stated. **5. History:** Prior history of aneurysm mentioned, but not for final size determination. **6. Aneurysm:** Known aneurysm projecting superior/posteriorly off the proximal right superior cerebellar artery **Size**: Not specified (history of aneurysm size is mentioned, but not for final size determination) **Location**: Right superior cerebellar artery **Changes over time**: None mentioned **Associated descriptions**: None

Table 2: ***Prompts Used for Structured Information and Aneurysm Size Extraction Using LLaMA 3.1***

(a) Prompt for Structured Information Extraction	**(b) Prompt for Aneurysm Size Extraction**
You are a medical data extractor specialized in analyzing medical reports. Analyze the provided report and extract the following information in a structured and concise format: 1. Abnormal findings:	You are a precise medical data extractor. Analyze the following medical report and follow these steps exactly: 1. Understand whether the report discusses about any aneurysm or not.

- Identify any abnormal findings mentioned in the report. - For each finding, provide its size, location, changes over time (if any), and associated descriptions (e.g., enhancement, signal type, pathology, etc.). 2. Specific conditions or diagnoses: - Summarize any specific conditions or diagnoses mentioned in the report. - For each, provide supporting findings and their relevance to the diagnosis. 3. Clinical relevance: - Identify any findings that explain the patient's symptoms or the reason for the examination. - Provide a brief explanation linking findings to clinical context. 4. Reasoning: - Include a concise explanation of your conclusions based on the information provided in the report. 5. History: - Note any prior history of aneurysm and their size is mentioned for the patient. But this should not be considered for final size determination. 6. Aneurysm: - Identify any aneurysm based on the abnormal findings from above. - For each finding, provide its size, location, changes over time (if any), and associated descriptions (e.g., enhancement, signal type, pathology, etc.). Stick to the above format. Do not add anything extra. Do not include this instruction in the response. Re-evaluate your answer before providing your final answer. **Format the response as**: <Observation> Your structured response goes here. </Observation>	2. Search for any mention of aneurysm measurements or size. Make sure the size found in relevant to aneurysm and medical context. 3. If aneurysm measurements or sizes are found: - Note any prior history mentioned for the patient. For final size determination if the size is part of history please ignore it. The diagnostic size is only for current observation. - For current observation only follow the below steps: - Extract ONLY the exact numerical measurements (length x width x depth, or diameter, or length, or width, or depth) - Include the measurement units (mm, cm) exactly as stated - if there are multiple aneurysm measurements or sizes then provide them in numbered points. 4. If no measurements are found: - Return exactly "No Size Mentioned" 5. For any report not discussing aneurysm should mention "Report Not Relevant". IMPORTANT RULES: - Return ONLY the size/measurements - Do not include any descriptive text - Do not mention the word aneurysm - Do not include location information - Do not add any interpretation - Include exactly one set of <Size> tags **Format your response as**: <Size> numerical measurements with units OR "No Size Mentioned" </Size>

Table 3: Evaluation Result from cross validation for the classification model

Model Name	Accuracy		F1		Precision		Recall	
	Mean	**Std (±)**	**Mean**	**Std (±)**	**Mean**	**Std (±)**	**Mean**	**Std (±)**
roberta-base-squad2	0.708	0.317	0.644	0.383	0.595	0.398	0.785	0.393
RadDistilBERT (distilbert-base-uncased)	**0.964**	**0.018**	**0.917**	**0.103**	**0.940**	**0.046**	**0.907**	**0.156**
bert-base-uncased	0.932	0.050	0.875	0.140	0.919	0.064	0.857	0.203
albert-base-v2	0.317	0.316	0.311	0.354	0.464	0.427	0.590	0.481

Table 4: Evaluation Results for out of distribution data ($N = 20$) with /without structured extracted data

N = 20	**Models Class Prediction (Full Report)**	**Models Size Prediction (Full Report)**	**Models Class Prediction (Impression Section)**	**Models Size Prediction (Impression Section)**
With Generated Structured Information	**100.00%**	**66.67%**	**100.00%**	**85.00%**
Without Generated Structured Information	100.00%	61.90%	100.00%	80.95%

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixAneurysmAnnotationGuide.docx