NOVA: A Benchmark for Rare Anomaly Localization and Clinical Reasoning in Brain MRI

Cosmin I. Bercea[1,4] **Jun Li**[1,4] **Philipp Raffler**[2] **Evamaria O. Riedel**[2]
Lena Schmitzer[2] **Angela Kurz**[2] **Felix Bitzer**[2] **Paula Roßmüller**[2] **Julian Canisius**[2]
Mirjam L. Beyrle[2] **Che Liu**[6] **Wenjia Bai**[6] **Bernhard Kainz**[5,6]
Julia A. Schnabel[1,3,4,7,*] **Benedikt Wiestler**[1,2,4,*]
[1]Technical University of Munich [2]Klinikum Rechts der Isar [3]Helmholtz Munich
[4]Munich Center for Machine Learning (MCML) [5]FAU Erlangen-Nürnberg
[6]Imperial College London [7]King's College London [*]Co-senior authors
{cosmin.bercea,julia.schnabel,b.wiestler}@tum.de

Abstract

In many real-world applications, deployed models encounter inputs that differ from the data seen during training. Open-world recognition ensures that such systems remain robust as ever-emerging, previously *unknown* categories appear and must be addressed without retraining. Foundation and vision-language models are pre-trained on large and diverse datasets with the expectation of broad generalization across domains, including medical imaging. However, benchmarking these models on test sets with only a few common outlier types silently collapses the evaluation back to a closed-set problem, masking failures on rare or truly novel conditions encountered in clinical use.

We therefore present *NOVA*, a challenging, real-life *evaluation-only* benchmark of $\sim$900 brain MRI scans that span 281 rare pathologies and heterogeneous acquisition protocols. Each case includes rich clinical narratives and double-blinded expert bounding-box annotations. Together, these enable joint assessment of anomaly localisation, visual captioning, and diagnostic reasoning. Because NOVA is never used for training, it serves as an *extreme* stress-test of out-of-distribution generalisation: models must bridge a distribution gap both in sample appearance and in semantic space. Baseline results with leading vision-language models (GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL-72B) reveal substantial performance drops, with approximately a 65% gap in localisation compared to natural-image benchmarks and 40% and 20% gaps in captioning and reasoning, respectively, compared to resident radiologists. Therefore, NOVA establishes a testbed for advancing models that can detect, localize, and reason about truly unknown anomalies.

1 Introduction

Generalization under distribution shift remains a central unsolved challenge in machine learning [14, 52]. Despite advances in large-scale pretraining and transfer learning [12, 33], most models fail to reliably detect or reason about previously unseen categories or domains at test time. Anomaly detection, the task of identifying deviations from a given normative distribution, *e.g.*, samples exclusively from healthy patients, represents an extreme stress-test of out-of-distribution (OOD) generalization due to the open-ended and unpredictable nature of anomalies. While OOD generalization has been extensively studied in natural image classification [17, 34, 22], it remains underexplored in healthcare. Medical data presents extreme heterogeneity, rare event frequencies, and non-standardized acquisition protocols, making it a worst-case scenario for evaluating model robustness to distribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

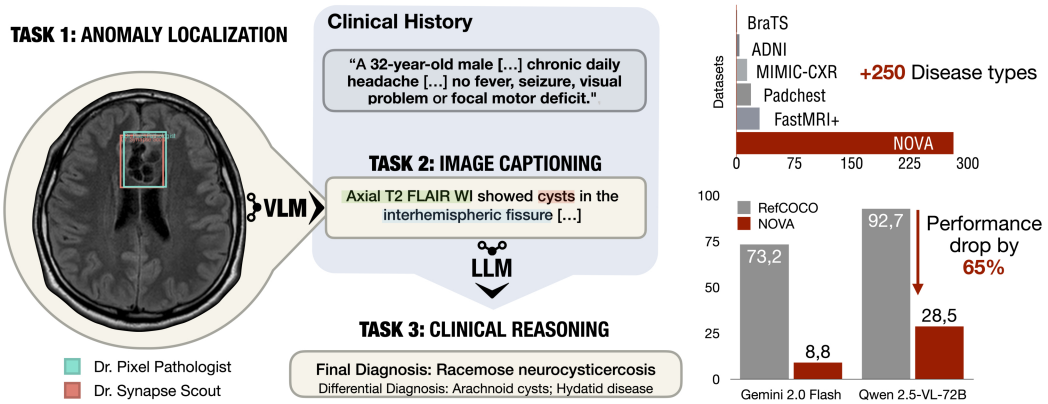

Figure 1: Overview of the NOVA benchmark. Task 1: Anomaly localization: models predict bounding boxes identifying abnormal regions in brain MRI; ground truth annotations from two independent radiologists are shown. Task 2: Image captioning: models generate a brief diagnostic description from the MRI image. Task 3: Diagnostic reasoning: models predict the final diagnosis by integrating clinical history and image findings. NOVA establishes the first benchmark designed to systematically evaluate vision-language models (VLMs) and large language models (LLMs) for rare anomaly localization, clinical description, and multimodal diagnostic reasoning in brain MRI.

shift. Detecting anomalies—potential pathologies—in imaging is often the first and most challenging step of the diagnostic process. Providing effective assistance to physicians at this stage has the potential to substantially improve clinical outcomes.

In medical imaging, unsupervised anomaly detection (UAD) methods [37, 6] are trained exclusively on healthy anatomy and identify deviations from this learned distribution as potential pathologies. While recent methods demonstrate strong performance on curated benchmarks [54, 38, 10, 43, 39, 5], they remain insufficiently reliable in the wild, particularly in high-stakes settings like clinical triage and health screening, where specificity and robustness to rare clinical presentations are essential [4, 23]. This challenge is particularly acute in magnetic resonance imaging (MRI) of the brain, where radiologists must detect subtle and diverse abnormalities across patient populations and heterogeneous imaging protocols.

The fundamental bottleneck lies in the datasets used for validation. Most existing benchmarks define anomalies through fixed categories, inducing implicit data leakage: although models are trained on healthy data, test sets remain constrained to known abnormality types. This narrows the evaluation to familiar distributions and undermines open-set detection. Datasets such as BraTS [29], ATLAS [26], and ISLES [18] were designed for segmentation and primarily capture canonical disease patterns, causing model development to converge on closed-set optimization rather than true discovery of unknown conditions.

The medical out-of-distribution analysis challenge (MOOD) [53] introduced synthetic anomalies to simulate unknown deviations. However, *real* anomalies from rare or previously unobserved diseases remain essential for clinical relevance. fastMRI+ [50] provided incremental pathology variability through bounding box annotations of brain and knee MRI scans [47], yet lacked the pathology heterogeneity and structured clinical metadata necessary to reflect clinical variability.

Detecting an abnormality alone does not satisfy clinical requirements. Radiologists must localize pathologically suspicious regions, assess severity, distinguish them from imaging artefacts, and formulate a differential diagnosis based on patient history and imaging findings. No existing dataset reflects this full diagnostic workflow, limiting prior benchmarks to binary detection and systematically failing to capture clinically meaningful information in generated text or diagnostic predictions [27]. Data sharing restrictions and poor standardization have further constrained the development of vision-language models (VLMs) in medicine.

Therefore, NOVA establishes a new benchmark for evaluating, detecting, and reasoning on unexpected abnormalities in clinical brain MRI, as illustrated in Figure 1. The dataset comprises 906 brain MRI scans spanning 281 rare and diagnostically diverse pathologies from Eurorad [13], enriched

with detailed clinical narratives. Each case is independently annotated by at least two radiologists with bounding boxes identifying suspected abnormalities. NOVA uniquely enables joint evaluation of anomaly localization, visual captioning, and diagnostic reasoning under real-world clinical heterogeneity. It is explicitly designed as an evaluation-only benchmark to serve as an extreme stress-test of OOD generalization, requiring models to bridge distribution shifts in both visual and semantic space.

We benchmark state-of-the-art vision-language models, including GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL-72B, on NOVA. Results reveal substantial performance degradation across all tasks, underscoring the urgent need for benchmarks that reflect the demands of open-world clinical reasoning.

2 Related Work

Anomaly detection, OOD detection, and novelty detection have received sustained focus in computer vision and machine learning, with advances across tasks from natural image understanding to industrial inspection [32, 25, 51, 45, 15, 36, 42]. Despite this progress, transferring these methods to medical imaging remains challenging. The concept of normality in medicine is inherently ambiguous, varying across individuals, imaging protocols, and institutions.

Clinical anomalies are often rare and highly heterogeneous, making them ill-suited for evaluation protocols that treat a selected set of predefined categories as representative out-of-distribution cases. The distinction between healthy and abnormal tissue is frequently subtle and localized, with considerable overlap between in-distribution and out-of-distribution regions within the same image. As a result, approaches that excel on constrained datasets such as MVTec-AD [7] fail systematically under the extreme clinical variability of real-world neuroimaging [19].

In medical imaging, unsupervised anomaly detection models learn the normative distribution of healthy anatomy to identify deviations as candidate pathologies [6]. Large healthy population datasets, including IXI [1], CamCAN [41], and UK Biobank [40] are valuable for normative modeling and population studies, but they are ill-suited for evaluating anomaly detection, as they lack pathological cases and localized anomaly annotations. Datasets such as ADNI [31] and OASIS [28] focus exclusively on Alzheimer's disease and neurodegeneration, providing only narrow coverage of pathologies. Similarly, condition-specific datasets, including MSLUB [24] for multiple sclerosis lesions, ATLAS [26] and ISLES [18] for stroke lesions, and BraTS [29] for brain tumors, support segmentation of predefined abnormalities but offer no framework for open-set detection or evaluation of vision-language reasoning in clinical contexts.

In parallel, large-scale vision-language datasets in medical imaging focus exclusively on chest radiographs. MIMIC-CXR [20] and PadChest [8, 9] integrate images with radiology reports for multimodal learning but are entirely disconnected from brain MRI. Hamamci et al. [16] introduced the CLM3D dataset and corresponding VLM3D challenge for developing generalist vision-language models in 3D medical imaging. However, CLM3D targets thoracic CT and focuses on common abnormality classification, report generation, and text-conditioned image synthesis, without addressing rare disease detection, anomaly localization, or open-world clinical reasoning.

Despite the critical technical and clinical need, a comprehensive neuroimaging benchmark remains absent. Brain MRI analysis presents significant technical challenges due to the wide spectrum of pathologies and their diverse appearances, ranging from localized lesions to diffuse structural alterations, coupled with inherent technical variability. Clinically, a clear need exists as most rare diseases are neurological or have neurological manifestations [35], positioning brain MRI centrally in patient care. NOVA establishes the first rigorous benchmark for systematically evaluating these capabilities under the real-world variability and diagnostic uncertainty of clinical brain MRI.

3 Dataset Description

We curated the NOVA dataset to establish an evaluation benchmark for vision-language model generalization under extreme clinical variability in brain MRI. We sourced cases from Eurorad, a peer-reviewed educational platform operating under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[1]. To comply with licensing requirements, we included only cases published after July 6, 2015. We filtered the dataset to include all cases from the "Neuroradiology"

[1] https://www.eurorad.org/node/38655

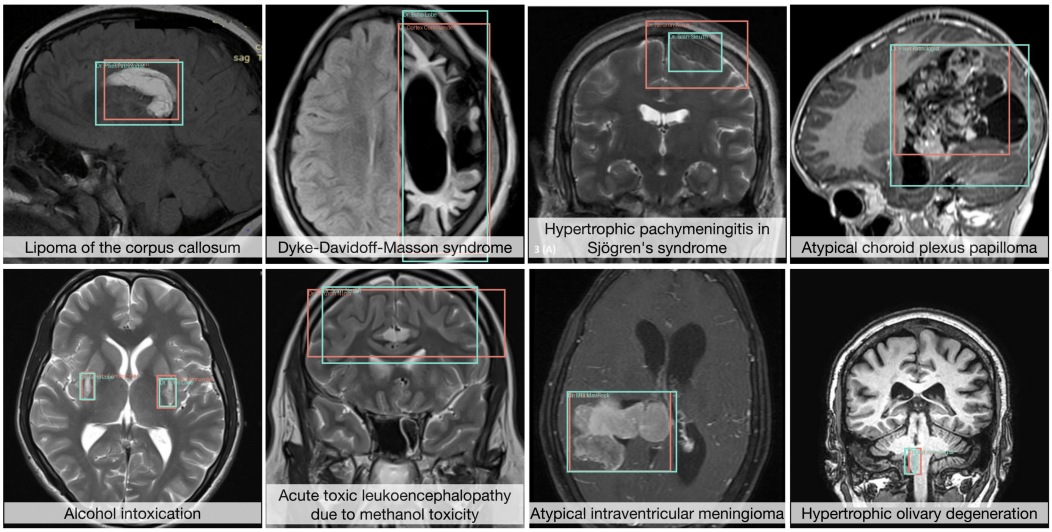

Figure 2: Representative brain MRI scans from the NOVA dataset illustrating the diversity of anatomical planes, MRI sequences, and pathological conditions. Radiologist-provided bounding box annotations are overlaid. The examples include rare congenital malformations, toxic and metabolic encephalopathies, and inflammatory or neoplastic lesions—capturing the broad radiological spectrum.

category and manually excluded non-relevant content such as CT, spine MRI, clinical photographs, and other non-MRI data. This ensured consistent imaging modality and anatomical focus.

We collected a total of 906 brain MRI scans spanning 281 unique diagnoses. We retained all images in their original form without preprocessing, cropping, or normalization to preserve the full clinical variability essential for evaluation. We preserved the naturally imbalanced long-tailed distribution of rare diseases to reflect real-world case frequencies. Representative examples illustrating the diversity of imaging planes, sequences, and pathologies are shown in Figure 2.

3.1 Dataset Composition

NOVA captures the diagnostic heterogeneity of clinical brain MRI. We included axial, sagittal, and coronal planes across standard sequences, including T1-weighted, T2-weighted, and FLAIR imaging. The distribution of anatomical planes and imaging sequences is detailed in Supplementary Tables 5 and 6. We manually grouped cases into six diagnostic categories: neoplastic, neurodegenerative, inflammatory, congenital, metabolic, and vascular pathologies. Figure 3a) shows the long-tailed distribution of diseases and the rarity of many conditions, which present unique challenges for model evaluation. The dataset's 281 unique diagnosis labels exceed the diversity of existing brain MRI benchmarks by an order of magnitude. We summarize additional statistics on patient demographics, and information of disease location, scale, and frequency in Supplementary Figures 9, 10, and 11.

3.2 Annotation Process and Quality Control

We implemented a rigorous multi-stage protocol to obtain high-quality anomaly localization annotations. Eight neuroradiology residents annotated the dataset using a custom web-based platform (Supplementary Figure 7). Ethics approval was waived by the local IRB at TUM University Hospital (IRB #2025-446-W-CB in Appendix F). Each case was independently labeled by two readers, who reviewed the full Eurorad clinical description and associated metadata to inform their annotations.

Inter-rater agreement was computed using a greedy matching algorithm that maximized intersection over union (IoU) between boxes. Figure 3b shows the distribution of inter-reader agreement, and Figure 3c presents the overall IoU distribution across annotations. Annotations with IoU > 0.3 were merged into consensus labels. While some inter-reader IoU values are modest, this reflects the inherent ambiguity of clinical neuroimaging. Lower overlap stems from diffuse pathologies with unclear boundaries, comprehensive labeling that includes incidental findings, and borderline

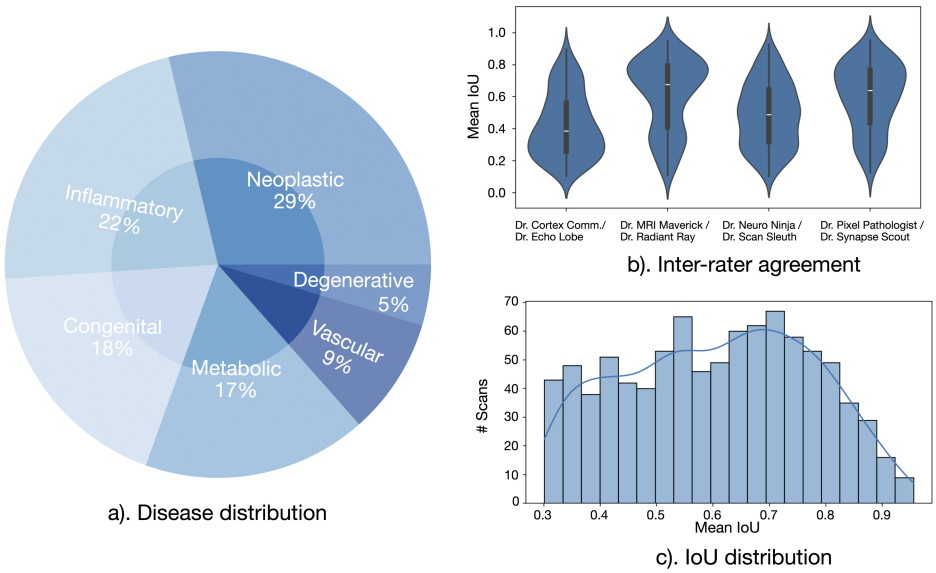

Figure 3: Dataset composition and annotation quality in NOVA. (a) Distribution of cases across six diagnostic categories. (b) Inter-rater agreement as mean intersection over union (IoU) between radiologist pairs. (c) Histogram of IoU scores across all scans.

decisions on subtle lesions. For 247 cases with persistent disagreement, a senior board-certified neuroradiologist (15+ years experience) adjudicated the final ground truth.

3.3 Data Format and Benchmark Design

We release all brain MRI scans as uniformly sized 480×480 grayscale PNG slices. We provide accompanying clinical metadata, including clinical history, patient demographics, imaging information, radiologist image captions, and bounding boxes for detected abnormalities in CSV files.

We explicitly designed NOVA as an evaluation-only dataset. Each case represents a unique diagnosis, and we do not provide predefined train, validation, or test splits. This enforces a zero-shot evaluation setting for all models, requiring them to generalize to previously unseen cases.

We publicly release NOVA on Hugging Face Datasets[2] under the same Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license as the Eurorad source. The dataset is distributed solely for non-commercial research to enable reproducible evaluation of VLMs under clinical conditions.

4 Benchmark Tasks

NOVA defines a comprehensive evaluation suite to assess the capabilities of vision-language models under realistic and clinically relevant conditions. These three tasks defined here reflect the sequential decision-making process of radiologists, progressing from anomaly localization to image description and diagnostic interpretation, enabling a realistic benchmarking.

4.1 Task 1: Anomaly Localization

This task requires models to detect and localize abnormalities within brain MRI scans, regardless of the patient's eventual diagnosis. Clinically, this is a relevant task, as most medical errors actually stem from not seeing a pathology at all [21]. Models must predict one or more bounding boxes per image corresponding to abnormal regions, using radiologist-annotated ground truth as reference. Performance is measured using mean average precision at intersection over union thresholds of 0.3 (mAP@30), 0.5 (mAP@50), and the COCO-style averaged mAP across thresholds from 0.50 to 0.95

[2] https://huggingface.co/datasets/c-i-ber/Nova

Table 1: Localization performance on NOVA. We evaluate the models with standard object detection metrics (mAP at multiple thresholds), detection accuracy (ACC50), number of true positives (TP30), number of false positives (FP30), and the false negative rate (FNR).

Model	mAP30 ↑	mAP50 ↑	mAP50-95 ↑	ACC50 ↑	TP30 ↑	FP30 ↓	FNR ↓
Gemini 2.0 Flash	20.16	7.37	1.99	8.83	227/1068	899	78.7
Qwen2-VL-72B	25.02	15.09	6.44	25.50	338/1068	1163	68.4
Qwen2.5-VL-72B	**37.66**	**24.49**	**11.23**	**28.48**	**406/1068**	**672**	**62**

in 0.05 increments (mAP@[50:95]). The benchmark also reports the number of correctly detected versus missed pathologies per case to reflect the clinical priority of minimizing false negatives. The dataset includes cases with multiple annotated abnormalities, providing a uniquely difficult evaluation setting for object detection under open-world conditions and rare disease variability.

4.2 Task 2: Image Description

This task measures the ability of models to generate clinically meaningful descriptions of brain MRI scans, an important prerequisite for making the correct diagnosis and in clinical communication. Each image is paired with an expert-generated caption describing the imaging findings. Evaluation uses case-insensitive exact keyword matching to compute precision, recall, and F1-score across the full keyword set. Modality-specific terms (such as *flair*, *axial*, *sagittal*, *t1*, *t2*, *coronal*, *dwi*, *t1w*, *t2w*, *weighted*) are evaluated separately from non-modality keywords capturing clinical content. Binary classification accuracy and F1-score for normal versus abnormal classification are also reported. Sentence-level generation quality is evaluated using BLEU-4 [30], METEOR [2], and BERT F1 [49].

4.3 Task 3: Diagnostic Reasoning

This task tests whether models can integrate clinical context and imaging observations to predict a diagnosis, arguably the "supreme discipline" in medical decision-making. Each case provides a brief clinical history and corresponding image caption as input, and the model must generate a free-text diagnostic label. Performance is reported as Top-1 accuracy (exact match with ground truth) and Top-5 accuracy (ground truth among the five most likely predictions). As model outputs are unconstrained free text, GPT-4o is used to perform semantic matching between predictions and ground truth labels. The task demands multimodal reasoning and open-ended prediction and is performed in a zero-shot setting, closely mirroring real-world clinical decision-making workflows.

5 Experiments and Results

We benchmarked large vision-language models on NOVA to systematically test their ability to generalize under extreme clinical heterogeneity. All experiments were conducted in inference-only mode. We report results for Gemini 2.0 Flash (Google DeepMind), Qwen2-VL-72B (Alibaba DAMO Academy), and Qwen2.5-VL-72B for abnormality grounding; and GPT-4o (OpenAI), Gemini 2.0 Flash, and Qwen2.5-VL-72B-Instruct for image captioning and diagnostic reasoning. As these models are proprietary and their training data is undisclosed, Eurorad cases may have been partially included. Results should thus be interpreted as an upper bound on zero-shot generalization. We encourage future evaluation of open models for a more conservative assessment.

We designed NOVA to expose generalization failures in models confronted with previously unseen, rare clinical cases. To do so, we evaluate along three critical axes of clinical reasoning: *localization, description, and diagnosis.*

5.1 Stress Test 1: Localization under Clinical Heterogeneity

The anomaly localization task revealed a strong performance degradation. While large vision-language models achieve detection scores of 73%–92% [44] on natural image benchmarks such as RefCOCO [46], performance on NOVA dropped sharply to 8.3%–28.5%. Models were evaluated using standard object detection metrics (mAP@30, mAP@50, and mAP@[50:95]), as summarized in Table 1. Despite occasional correct detections, all models exhibited poor calibration under clinical

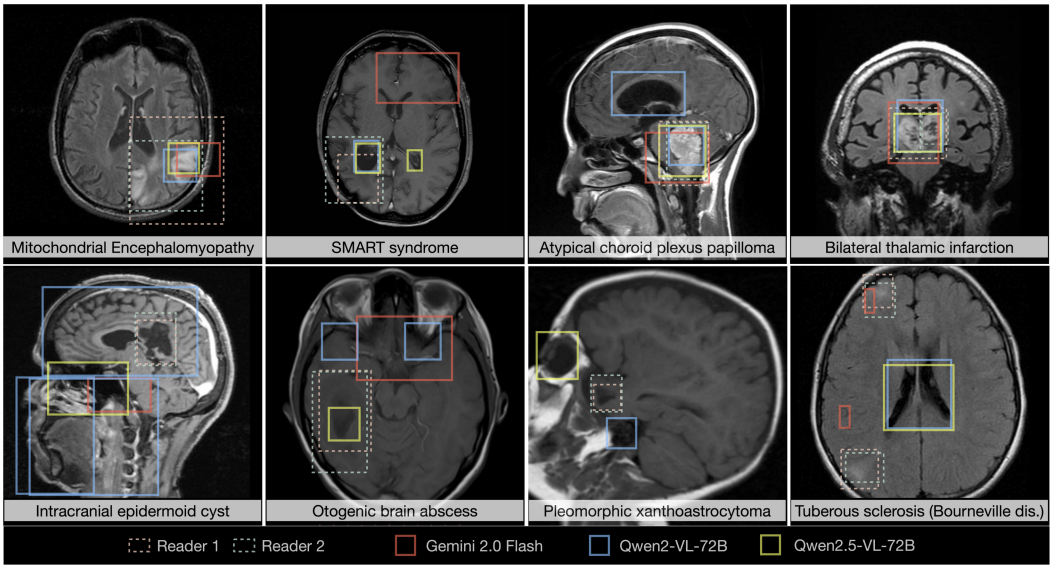

Figure 4: Examples of model predictions for anomaly grounding on NOVA. Ground truth and model-predicted bounding boxes are shown for Gemini 2.0 Flash, Qwen2.0-VL-72B, and Qwen2.5-VL-72B.

Table 2: Image Description on NOVA. Captioning quality is evaluated by Clinical Term F1, Modality Term F1, BLEU, and METEOR. Binary F1 measures binary abnormality classification performance.

Model	Clinical F1 (%)	Modality F1 (%)	BLEU-4	METEOR	BERT F1 (%)	Binary F1 (%)
Gemini 2.0 Flash	**19.8**	**59.8**	**1.83**	15.2	85.5	5.3
GPT-4o	15.7	49.3	0.92	**17.5**	84.3	**11.3**
Qwen2.5-VL-72B	13.6	45.3	1.08	17.1	84.4	2.4

distribution shift, frequently producing incomplete, misplaced, or spurious bounding boxes. For completeness, we also tested medical-domain VLMs (CheXagent [11], MAIRA-2[3], HuatuoGPT Vision [48]) on a 25-case subset. All underperformed generalist models, with CheXagent and MAIRA-2 failing to generalize (mAP@50 = 0) and HuatuoGPT Vision reaching only 8.3. We included the full results in Supplementary Table 7. Clinical inspection of representative cases (Figure 4) revealed typical failure patterns such as false-positive detection of normal anatomical structures (e.g., orbital cavity misinterpreted as lesion by Qwen2.5-VL) and failure to localize true abnormalities even under relaxed overlap thresholds. Quantitatively, even at 30% IoU criteria, fewer than half of ground truth abnormalities were detected (62% FNR), and over 600 false-positive boxes were recorded. In clinical practice, missed anomalies risk delayed or missed diagnoses, while high false-positive rates drive unnecessary specialist referrals, patient anxiety, and increased healthcare costs. NOVA sets the first extreme benchmark designed to systematically expose these critical failure modes.

5.2 Stress Test 2: Description under Semantic Shift

The second test probes whether models can generate clinically meaningful image descriptions under severe semantic shifts. Table 2 summarizes performance across Clinical Term F1, Modality Term F1, BLEU-4, METEOR, BERT F1, and binary abnormality classification accuracy. Gemini 2.0 Flash achieved the highest Clinical Term F1 (19.8%), Modality Term F1 (59.8%), and BLEU (1.83), reflecting comparatively stronger recognition of structured imaging attributes. GPT-4o outperforms in Binary F1 (11.3%) and METEOR (17.5), indicating slightly better fluency and sentence-level description. Qwen2.5-VL-72B-Instruct underperformed across all metrics. Despite the low lexical overlap captured by BLEU-4 and METEOR, models achieve high BERTScore (0.84–0.85), indicating that generated captions preserve most of the clinical meaning. To further investigate the semantic limitations observed in captioning (Table 2), we analyzed language behavior across models in Figure 5. Example cases illustrate that all models tend to produce longer but vaguer descriptions under uncertainty. The ground truth showed the highest vocabulary size (1527 unique words), reflecting

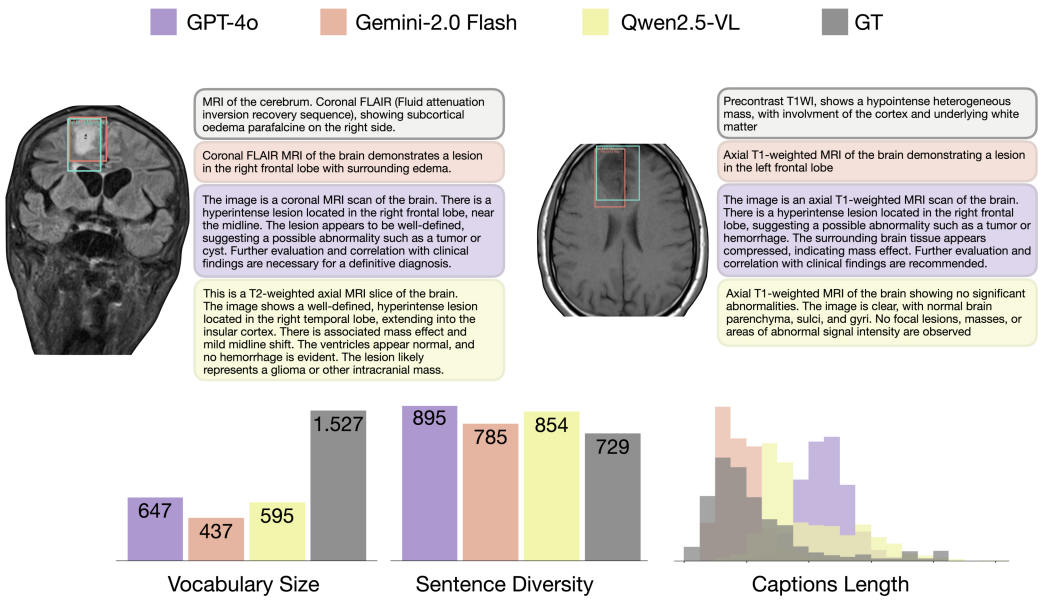

Figure 5: Task 2: Image Captioning. (Top) Example image-caption pairs showing model predictions and reference ground truth. Model outputs tend toward verbose, redundant phrasing with fewer specialized terms. (Bottom) Quantitative analysis. Left: Vocabulary size (unique words). Center: Sentence diversity (unique captions). Right: Caption length distribution (number of words per caption). Ground truth radiology reports exhibit the highest vocabulary richness but shorter, information-dense sentences. All models display severe vocabulary collapse and compensate with longer and more varied sentence constructions.

expert use of precise diagnostic terminology. VLMs exhibited drastic vocabulary compression, with GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL using only 647, 437, and 595 unique words, respectively. Interestingly, models showed comparable or slightly higher sentence diversity (895, 785, and 854 unique captions vs. 729 in the ground truth), likely due to paraphrasing and verbose redundancy. Caption length distribution revealed distinct patterns: Gemini produced captions with similar lengths to ground truth, while GPT-4o and Qwen2.5-VL consistently generated longer outputs. This analysis highlights a consistent pattern of low lexical precision and repetitive verbosity, particularly for GPT-4o and Qwen2.5-VL, confirming that current models exhibit limitations in expressing clinically meaningful descriptive detail in image captioning tasks under real-world clinical distribution shifts.

5.3 Stress Test 3: Diagnostic Reasoning under Distributional Shift

The final test evaluates models' ability to assign diagnostic labels based on combined image captions and clinical history. Performance was assessed via Top-1 and Top-5 classification accuracy (Table 3). GPT-4o achieved the highest scores (24.2% Top-1, 38.4% Top-5), with Gemini 2.0 Flash and Qwen2.5-VL-72B performing lower.

To further probe diagnostic behavior, we analyzed prediction distributions against ground truth (Figure 6). All models followed the expected Zipfian-like scaling of disease frequencies (right), demonstrating comparable rank-frequency slopes to ground truth. However, this occurred over a substantially compressed label space, with model predictions collapsing onto smaller vocabularies covering only $\sim$30% of the ground truth (Table 3). This truncation effect was also visible in the cumulative frequency curves (left), where model predictions saturated rapidly relative to ground truth.

Ground truth labels exhibited a Shannon entropy of 8.68 bits, reflecting the high uncertainty and diversity of rare disease distributions. In contrast, model outputs showed a marked entropy reduction of approximately 1 bit (Table 3), consistent with over-reliance on dominant classes and poor exploration of the long tail—an effect akin to premature entropy collapse under distributional shift.

Table 3: Diagnostic reasoning results on NOVA. Diagnostic accuracy is captured by the Top-1 and Top-5 accuracy. Coverage and entropy are extracted from diagnostic reasoning distributions.

Model	Top-1	Top-5	Cov.	Ent.
Gemini 2.0 Flash	22.1	37.4	29.4	**7.71**
GPT-4o	**24.2**	**38.4**	**31.9**	7.64
Qwen2.5-VL-72B	22.4	35.2	26.1	7.26

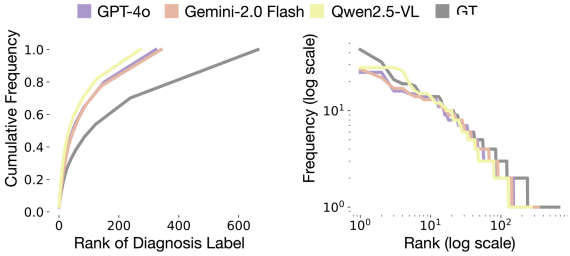

Figure 6: Distribution of diagnostic label frequencies for ground truth vs model predictions.

Table 4: Disentangling captioning and reasoning gaps. Replacing model-generated captions with human or ground-truth descriptions markedly improves diagnostic accuracy (captioning gap). Even with perfect captions, model reasoning lags expert performance (reasoning gap).

Condition	**Top-1**	**Top-5**
(a) Human captions → Human reasoning (Resident Neuroradiologists)		
Neuroradiologist 1	48.0%	76.0%
Neuroradiologist 2	52.0%	68.0%
(b) Human / ground-truth captions → AI reasoning		
Radiologist 1 captions	40.0%	60.0%
Radiologist 2 captions	40.0%	60.0%
Ground-truth captions	44.0%	52.0%
GPT-paraphrased GT captions	42.0%	66.0%
(c) AI-generated captions → AI reasoning		
GPT-4o	24.0%	38.4%
DeepSeek-R1	24.0%	44.0%

Reasoning gap
Top-1: ▼ 16%
Top-5: ▼ 21%

Captioning gap
Top-1: ▼ 40%
Top-5: ▼ 36%

5.4 Disentangling Captioning and Reasoning Gaps

The grounding performance of VLMs on NOVA drops by over 65% compared to natural-image referring-expression benchmarks, revealing a pronounced challenge in transferring generic spatial alignment to medical imaging. To understand how this failure propagates to clinical inference, we disentangle two downstream limitations: the *captioning gap*, where imperfect visual descriptions hinder reasoning, and the *reasoning gap*, where models underperform even with high-quality captions.

We compared the reasoning performance on a 25-case subset, that we also used for human baselines (a), under the following setups: (b) radiologist-written captions derived directly from images, ground-truth Eurorad captions, and GPT-paraphrased versions of the ground truth to test robustness to phrasing, (c) model-generated captions (GPT-4o). Replacing model-generated captions with human or GT captions markedly improves diagnostic accuracy (Table 4), establishing a clear captioning gap. However, even with ideal captions, models remain below neuroradiologist-level accuracy, indicating a residual reasoning gap. Using ground-truth captions, DeepSeek-R1 reaches 44% Top-1 and 52% Top-5 accuracy, approaching residents but still trailing human experts (48–52% / 68–76%). Performance remains stable under GPT-paraphrased captions (Top-1: 42% vs. 44.0%), suggesting that the model responds primarily to clinical semantics rather than memorized phrasing.

To examine whether this residual gap arises from missing rare-disease knowledge, a board-certified neuroradiologist probed GPT-4o's understanding across the same cases. The model consistently produced clinically accurate textual descriptions of hallmark findings, e.g., *Multiple intracranial meningiomas—well-circumscribed, extra-axial masses with homogeneous gadolinium enhancement and a "dural tail" sign;* These qualitative probes confirm that the rare-disease concepts and their imaging correlates are represented in the model's linguistic knowledge. The persistent reasoning deficit therefore reflects a limited ability to integrate visual evidence with this knowledge into coherent diagnostic inference, rather than an absence of medical understanding.

6 Discussion

NOVA introduces a new benchmark for evaluating anomaly detection and multimodal reasoning in clinical brain MRI. Its design offers several key advantages. First, NOVA provides one of the largest and most diverse expert-annotated collections of brain MRI scans available, covering approximately 900 scans with over 280 distinct diagnoses and rare pathological conditions. Second, the dataset uniquely integrates multimodal annotations, combining radiologist-drawn bounding boxes, expert-generated image captions, and detailed clinical histories. This comprehensive structure enables systematic evaluation of detection, description, and diagnostic reasoning within a single resource. Third, the dataset reflects real-world clinical variability by using actual patient cases rather than synthetic perturbations, creating a challenging and realistic testbed. Finally, the structured multi-reader annotation protocol with adjudication by a senior neuroradiologist ensures a high level of annotation quality and reliability.

Despite these advantages, NOVA has limitations that are important to acknowledge. The dataset is sourced from a European radiology teaching repository, which may introduce geographic or demographic biases that could affect model generalization in other healthcare systems. Additionally, NOVA provides only 2D image slices rather than full 3D volumes. While this choice may constrain certain volumetric analyses, the decision to release data in 2D format was deliberate: most standard machine learning and computer vision tools and libraries offer limited support for 3D medical imaging, which can significantly slow down experimentation and accessibility for the broader research community. Finally, NOVA is released as an *evaluation-only* benchmark, not intended for supervised model training. This design reflects the realities of rare disease imaging, where collecting sufficient labeled data for training is often infeasible and where true generalization must be tested without model adaptation. Beyond current baselines, future work could explore adapting large generalist models to medical imaging domains, injecting structured medical knowledge, and enhancing multimodal reasoning to bridge visual and clinical understanding.

Looking forward, we plan to maintain NOVA as a dynamic benchmark and to open a public leaderboard to encourage continuous community participation and advancement of the state of the art. Given the dataset's focus on rare diseases and its intended role as an inference benchmark, we do not envision extensions to fine-tuning tasks or inclusion of 3D imaging data. Instead, we anticipate that NOVA will catalyze the development of next-generation foundation models and VLMs capable of performing robust diagnostic reasoning under realistic open-set clinical conditions.

7 Conclusion

We present NOVA, the first large-scale, expert-annotated benchmark dataset for anomaly localization, clinical captioning, and diagnostic reasoning in brain MRI. NOVA provides a uniquely challenging and clinically grounded resource, combining real-world imaging variability with high-quality multimodal annotations. By releasing NOVA to the community, we aim to establish a new standard for evaluating the robustness and generalization of models for clinical anomaly detection and multimodal medical reasoning. We invite the research community to engage with NOVA and drive the development of next-generation models capable of detecting the unknown in clinical imaging.

References

[1] Ixi dataset. https://brain-development.org/ixi-dataset/. Accessed: 2023-02-15.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[3] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

[4] Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. In *Medical Imaging with Deep Learning*, 2023.

[5] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Diffusion models with implicit guidance for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–220. Springer, 2024.

[6] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Evaluating normative representation learning in generative ai for robust anomaly detection in brain imaging. *Nature Communications*, 16(1):1624, 2025.

[7] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9584–9592, 2019.

[8] Agustín Bustos, Antonio Pertusa, Jose M. Salinas, and María de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[9] Daniel C Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*, 2024.

[10] Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical Image Analysis*, 64:101713, 2020.

[11] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] European Society of Radiology (ESR). Eurorad – radiological case database, 2025. https://www.eurorad.org, accessed March 2025.

[14] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *European conference on computer vision*, pages 701–717. Springer, 2022.

[15] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10721–10740, 2023.

[16] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.

[17] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[18] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1):762, 2022.

[19] Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

[20] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019.

[21] Y. W. Kim and L. T. Mansfield. Fool me twice: Delayed diagnoses in radiology with emphasis on perpetuated errors. *AJR. American Journal of Roentgenology*, 202(3):465–470, 2014.

[22] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021.

[23] Seungjun Lee, Boryeong Jeong, Minjee Kim, Ryoungwoo Jang, Wooyul Paik, Jiseon Kang, Won Chung, Gil-Sun Hong, and Namkug Kim. Emergency triage of brain computed tomography via anomaly detection with a deep generative model. *Nature Communications*, 13:4251, 07 2022.

[24] Žiga Lesjak, Alina Galimzianova, Andrej Koren, et al. A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16(1):51–63, 2018.

[25] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[26] Sook-Lei Liew, Bethany P. Lo, ., and et al. Miarnda R. Donnelly. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data*, 9, 2022.

[27] Faisal Mahmood. A benchmarking crisis in biomedical machine learning. *Nature Medicine*, 31(4):1060–1060, 2025.

[28] Daniel S. Marcus, Aditya F. Fotenos, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010.

[29] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[31] Ronald C. Petersen, Paul S. Aisen, Laurel A. Beckett, Michael C. Donohue, Anthony C. Gamst, Danielle J. Harvey, Clifford R. Jr. Jack, William J. Jagust, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Michael W. Weiner. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, January 2010.

[32] Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

[34] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[35] Carola Reinhard, Anne-Catherine Bachoud-Lévi, Tobias Bäumer, Enrico Bertini, Alicia Brunelle, Annemieke I. Buizer, Antonio Federico, Thomas Gasser, Samuel Groeschel, Sanja Hermanns, Thomas Klockgether, Ingeborg Krägeloh-Mann, G. Bernhard Landwehrmeyer, Isabelle Leber, Alfons Macaya, Caterina Mariotti, Wassilios G. Meissner, Maria Judit Molnar, Jorik Nonnekes, Juan Dario Ortigoza Escobar, Belen Pérez Dueñas, Lori Renna Linton, Ludger Schöls, Rebecca Schuele, Marina A. J. Tijssen, Rik Vandenberghe, Anna Volkmer, Nicole I. Wolf, and Holm Graessner. The european reference network for rare neurological diseases. *Frontiers in Neurology*, Volume 11 - 2020, 2021.

[36] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022.

[37] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[38] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

[39] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489. Springer, 2022.

[40] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

[41] Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-CAN, and Richard N. Henson. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269, 2017. Data Sharing Part II.

[42] Han Wang and Yixuan Li. Bridging ood detection and generalization: A graph-theoretic view. *Advances in Neural Information Processing Systems*, 2024.

[43] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2022.

[44] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[45] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32598–32611, 2022.

[46] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[47] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

[48] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

[49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[50] Ruiyang Zhao, Burhaneddin Yaman, Yuxin Zhang, Russell Stewart, Austin Dixon, Florian Knoll, Zhengnan Huang, Yvonne W Lui, Michael S Hansen, and Matthew P Lungren. fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 9(1):152, 2022.

[51] Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 72110–72123, 2023.

[52] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[53] David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, et al. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022.

[54] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 289–297. Springer, 2019.

Acknowledgments and Disclosure of Funding

C.I.B. has received funding from the EVUK program ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria. This work was in part supported by Berdelle-Stiftung (grant TimeFlow). J.A.S. acknowledges funding from the Munich Center of Machine Learning (MCML) and the DAAD program Konrad Zuse School of Excellence in Reliable Artificial Intelligence, both sponsored by the German Federal Ministry of Science Technology and Space.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's scope and contributions. The paper introduces NOVA, a dataset designed to systematically benchmark vision-language models under rare disease distribution shift in brain MRI. The claims regarding NOVA's design, evaluation axes (localization, captioning, diagnostic reasoning), and use to expose failure modes are directly supported by the experiments and analyses. The introduction clearly motivates the problem, states the intended role of NOVA, and acknowledges limitations, including the potential inclusion of Eurorad cases in model pretraining. All claims are grounded in the presented data and follow standard dataset paper practice at NeurIPS.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses several limitations of the dataset and evaluation setup. We acknowledge that Eurorad cases may partially overlap with proprietary model pretraining data, meaning results should be interpreted as upper bounds on zero-shot generalization. We also clarify that NOVA is designed as an evaluation-only benchmark and does not address model training under rare disease shift. The dataset is restricted to brain MRI and rare neurologic pathologies, which may limit generalization to other body regions or imaging modalities. Furthermore, we discuss the challenges of clinical annotation variability and potential label noise in large-scale radiology datasets. These limitations are transparently presented and contextualized within the intended scope of NOVA as a stress-test benchmark to expose current model limitations and encourage future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results, theorems, or formal proofs. The contribution is an empirical dataset and evaluation benchmark.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses the information needed to reproduce the main experimental results. We provide a detailed description of the dataset construction pipeline, case selection, and radiologist annotation process. There are no predefined dataset splits; the dataset is designed as an evaluation-only benchmark. We release all dataset files, annotations, evaluation scripts, and detailed prompt templates used for querying each vision-language model. This ensures full reproducibility of the benchmarking experiments. While some models evaluated (e.g., GPT-4o, Gemini 2.0 Flash) are proprietary, our dataset and code release enable any future research group to replicate the evaluation on any accessible model.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed

instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset and accompanying code for evaluation were made publicly available under an open license on huggingface upon submission (https://huggingface.co/datasets/c-i-ber/Nova/). We provide comprehensive instructions for dataset access, including file structure, patient-level annotations, and radiologist bounding box labels. All code and scripts used to run the evaluation pipeline, including model query templates and evaluation metrics for anomaly localization, captioning, and diagnostic reasoning, will be provided in the supplementary material. The supplemental material describes the exact experimental setup and environment configuration to faithfully reproduce the results. Proprietary model weights (e.g., GPT-4o, Gemini 2.0 Flash) cannot be redistributed but we document how models were accessed for evaluation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides full details of the experimental setup. Since NOVA is an evaluation-only benchmark, there are no training procedures or dataset splits. We specify the prompt formats used for each model, the inference setup (including API access and model version identifiers for GPT-4o, Gemini 2.0 Flash, and Qwen2.5-VL), and the evaluation metrics used for each task (mAP, F1, BLEU, METEOR, accuracy, etc.). All task-specific evaluation details, such as IoU thresholds for detection and the diagnostic term matching logic, are described in the paper and supplemental material. This allows readers to fully interpret the results and reproduce the evaluation pipeline using our released code and data.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports quantitative evaluation results from inference-only API calls to fixed pretrained models. Since no training, random seeds, or sampling were involved, there was no variability across runs. As such, statistical significance testing or error bars were not applicable. The experimental setting is fully deterministic and exhaustively covers the benchmark data. We document this setup in the main paper and supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments were conducted using inference-only API calls to closed-source pretrained vision-language models (GPT-4o, Gemini 2.0 Flash, Qwen2.5-VL). No model training or fine-tuning was performed. As model providers do not disclose compute infrastructure or execution time per query, we do not report runtime estimates. All API interactions, prompts, and evaluation scripts are documented and will be released to ensure reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research strictly adheres to the NeurIPS Code of Ethics. The dataset was constructed entirely from publicly available, de-identified medical imaging cases released under educational or open-access use (Eurorad). No patient-identifying information is present, and no new data collection or clinical intervention was performed. Ethics approval was waived by the local IRB at TUM University Hospital (IRB #2025-446-W-CB; see Appendix F), as the radiologist's annotations were performed in an anonymized manner. All experimental evaluations were conducted using commercially available APIs without training or fine-tuning. We discuss the limitations and potential societal impacts of our work, includingthe risks of false positives or miscalibration in medical settings, and highlight that our benchmark is intended solely for research evaluation—not clinical use. Full transparency is provided regarding data provenance, evaluation methods, and access protocols.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both potential positive and negative societal impacts. On the positive side, NOVA enables more realistic evaluation of generalization and robustness in clinical AI systems, helping benchmark the performance of vision-language models in medical contexts that reflect true diagnostic complexity. This could support the development of safer and more interpretable AI for healthcare applications. On the negative side, the dataset inherits population and acquisition biases from Eurorad, which may not represent global or demographically balanced clinical populations. Additional risks include the potential misuse of benchmark results as evidence of clinical readiness, overfitting to the dataset, or use outside its intended non-clinical research scope.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: NOVA does not include generative models or scraped web content, but we recognize that medical datasets inherently carry risk of misuse. To mitigate this, we implement several safeguards. First, all data is derived from publicly released, de-identified Eurorad cases curated for educational purposes, with no patient-identifying information. Second, we release the dataset under a research-only license with clear restrictions against clinical deployment or commercial use. Third, we provide structured usage guidelines emphasizing that NOVA is an evaluation benchmark, not intended for training diagnostic systems. Finally, metadata has been reviewed to avoid inclusion of sensitive or potentially identifiable content. These steps aim to ensure responsible use and reduce misuse risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset used in this work is derived from Eurorad (https://www.eurorad.org), a publicly accessible educational platform for radiology case sharing. Each case is published by contributing radiologists under Eurorad's terms of use, which permit non-commercial use for research and education. We properly credit Eurorad in the paper and provide the original URL and citation. For evaluation models, we use only publicly accessible APIs (e.g., GPT-4o, Gemini 2.0 Flash, Qwen2.5-VL) and

respect the license terms set by each provider. All usage is restricted to non-commercial research.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduce NOVA, a new benchmark dataset for anomaly localization, visual captioning, and diagnostic reasoning in brain MRI. The dataset includes over 900 cases with rare and heterogeneous diagnoses curated from Eurorad, each containing high-resolution medical images, expert-written diagnostic reports, and bounding box annotations of anomalies provided independently by nine radiologists. We provide full documentation detailing the data structure, annotation format, label ontology, and usage guidelines. All assets are publicly availble under a non-commercial research license. A dataset card and structured metadata are provided in the public dataset.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or direct interaction with human subjects. All annotations were obtained from certified radiologists as part of their professional duties under existing institutional agreements. No crowd workers or external participants were involved, and no additional compensation was provided beyond standard institutional arrangements.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We obtained an IRB number (2025-446-W-CB) and confirmation of exemption from full IRB approval, as our study exclusively utilized publicly available anonymized datasets. In addition, our annotation protocol was performed internally by in-house radiologists under anonymized conditions, ensuring that no identifiable human subject data were involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does not introduce any novel LLM-based methods nor use LLMs as core methodological components. All evaluations are conducted via zero-shot inference using existing publicly available LLM APIs (e.g., GPT-4o, Gemini 2.0, Qwen2.5-VL), which are the *subject* of the benchmark rather than tools used for scientific innovation or methodology development.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Appendix

This supplementary material provides detailed insights into the NOVA benchmark. Section A outlines the annotation protocol, including our custom web interface, rater instructions, adjudication strategy, and the weighted consensus merging based on expert agreement. Section B presents summary statistics capturing the demographic diversity of patients and the spatial characteristics of the annotated abnormalities, including bounding box distributions and heatmaps. Section C compares medical VLMs against generalist VLMs. Section D introduces the three main tasks defined in our benchmark: abnormality detection, radiological image captioning, and clinical reasoning through differential diagnosis. For each task, we detail the prompting formats and model-specific configurations. We further provide two standardized evaluation protocols to assess the factual consistency and diagnostic correctness of model outputs in zero-shot settings. Section E includes concrete prompting examples to support reproducibility. Section F includes our IRB document.

A Annotation Protocol and Interface

To ensure reliable ground truth annotations for NOVA, we designed a multi-stage annotation pipeline in collaboration with experienced neuroradiologists. Each image was independently annotated by two medical experts using a custom web interface developed for this project (Figure 7). Annotators were presented with the case description, including clinical history and radiological findings, alongside the MRI image. They could interactively draw, adjust, and delete bounding boxes.

Consensus merging and disagreement resolution. Eight neuroradiology residents participated in the annotation process. Each image was reviewed by a consistent pair of annotators drawn randomly from this pool. Annotators were instructed to mark all visually and clinically relevant abnormalities, excluding normal anatomical variations and imaging artifacts. To construct a consensus set, we first computed the intersection-over-union (IoU) for all bounding box pairs. When annotations from the two readers did not overlap sufficiently (IoU ≤ 0.3), the image was flagged for adjudication. A board-certified senior neuroradiologist reviewed these cases and provided the final reference bounding boxes. Examples of annotator disagreement and expert adjudication are illustrated in Figure 8. This set of 188 images served as the basis for estimating each annotator's agreement with the expert.

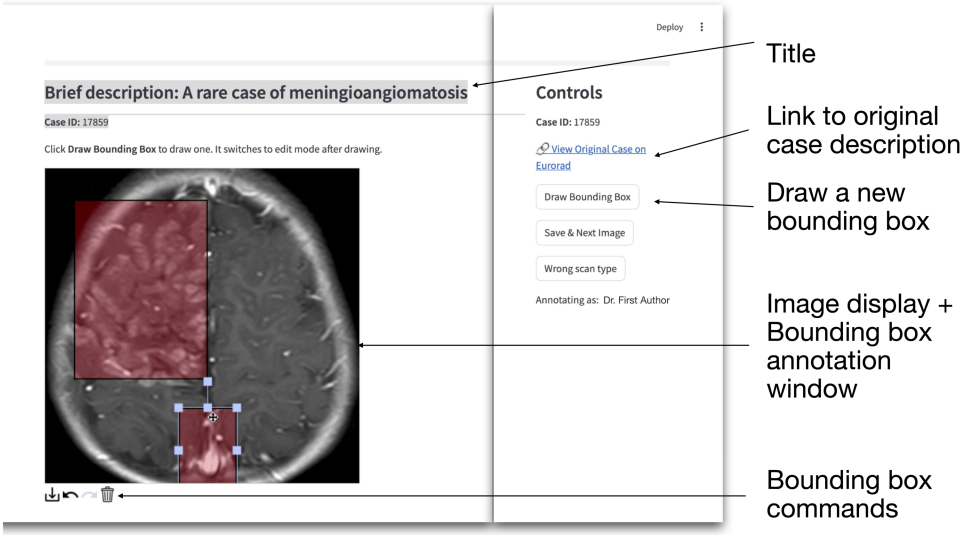

Figure 7: Each case includes a brief clinical description and a link to the full Eurorad entry. Annotators mark pathologically relevant regions using a custom bounding box tool. Controls are optimized for clinical workflows, enabling rapid annotation and review.

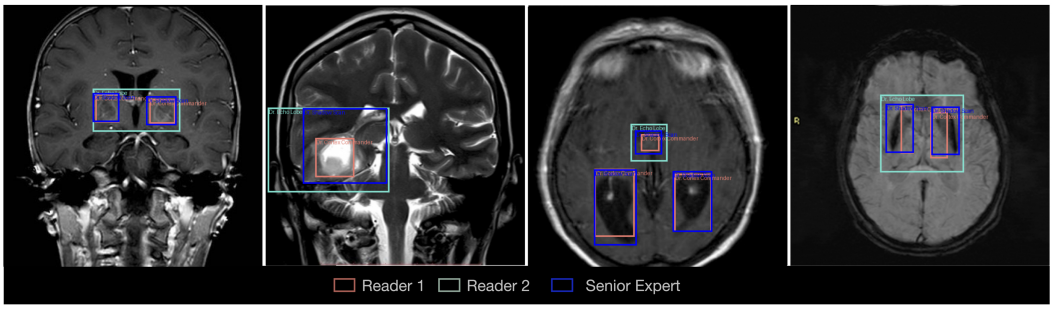

Figure 8: **Examples of annotation disagreement.** Each image shows bounding boxes from two independent annotators (Reader 1 in salmon and Reader 2 in teal) and the adjudicated expert bounding box (blue). These cases illustrate scenarios where annotators either disagreed on lesion boundaries or identified different pathological structures. In such cases, a board-certified neuroradiologist reviewed the image and merged or revised annotations to produce the final consensus labels.

Table 5: Distribution of anatomical planes in NOVA.

View	Axial	Sagittal	Coronal	Unknown
Count	548	64	140	154

For each annotator r, we computed the average intersection-over-union (IoU) between their annotations and the expert-approved boxes across all adjudicated images they participated in. Let I_r denote this mean agreement score for reader r.

For the remaining images where readers produced overlapping boxes (IoU > 0.3), we merged these into a single consensus box. The coordinates of the merged box $\mathbf{b}_{\text{merged}}$ were computed using a weighted average of the two boxes:

$$\mathbf{b}_{\text{merged}} = w_A \cdot \mathbf{b}_A + w_B \cdot \mathbf{b}_B, \quad \text{with} \quad w_A = \frac{I_A}{I_A + I_B}, \quad w_B = \frac{I_B}{I_A + I_B}$$

where I_A and I_B are the expert agreement scores for annotators A and B, and $\mathbf{b}_A$, $\mathbf{b}_B$ are their respective bounding boxes.

This approach ensured that annotators with a stronger history of agreement with the expert contributed more to the final consensus. It allowed us to systematically leverage expert-reviewed cases to calibrate reader reliability, even when direct adjudication was not performed.

Annotation reliability. The resulting annotation set combines double-blinded readings with targeted expert oversight. While some inter-reader variability reflects the inherent subjectivity in clinical interpretation, the adjudication procedure mitigates systematic noise and ensures high-quality labels suitable for benchmarking robust detection systems.

B Dataset Composition and Annotation Statistics

The NOVA dataset encompasses a wide spectrum of demographic and spatial variability, reflecting the diversity of real-world clinical neuroimaging. In this section, we provide supporting statistics to contextualize the challenges posed by the benchmark.

Acquisition variability. NOVA preserves the heterogeneity of real-world clinical brain MRI: scans originate from multiple sites and protocols and include standard anatomical planes (axial, coronal, sagittal) across T1-weighted, T2-weighted, FLAIR, and other sequences without harmonization. As scanner vendor and field strength information were unavailable for Eurorad cases, acquisition metadata were extracted from image captions and summarized in Tables 5, and 6.

Patient demographics. Figure 9 presents the distribution of patient sex and age across all included cases. The dataset spans a broad age range, from pediatric to geriatric populations, with a relatively

Table 6: Distribution of imaging sequences in NOVA.

Sequence	FLAIR	T2w	T1w	DWI	ADC	SWI	GRE	PD	Unknown
Count	242	226	223	41	16	15	2	1	140

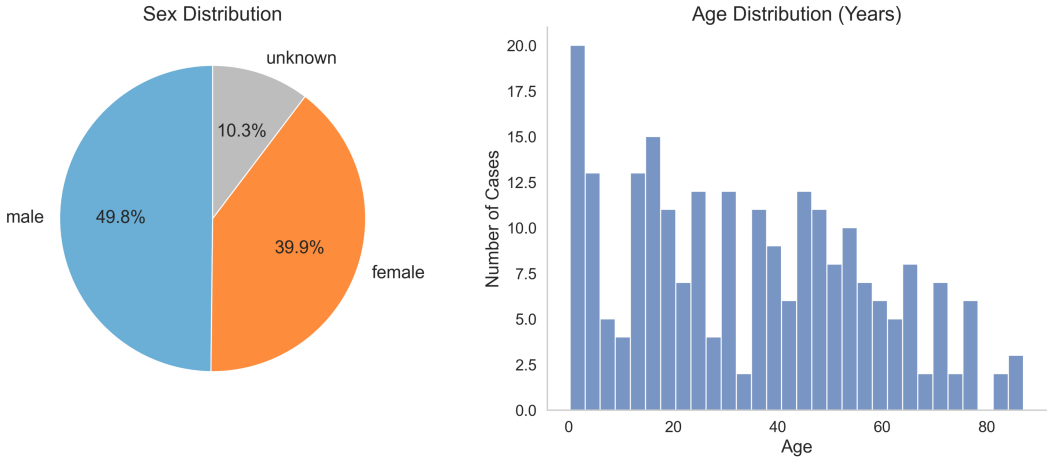

Figure 9: **Demographics.** Left: Sex distribution of cases, with a nearly balanced male-to-female ratio and a subset of cases with unknown sex. Right: Histogram of patient ages showing broad coverage across pediatric, adult, and elderly populations.

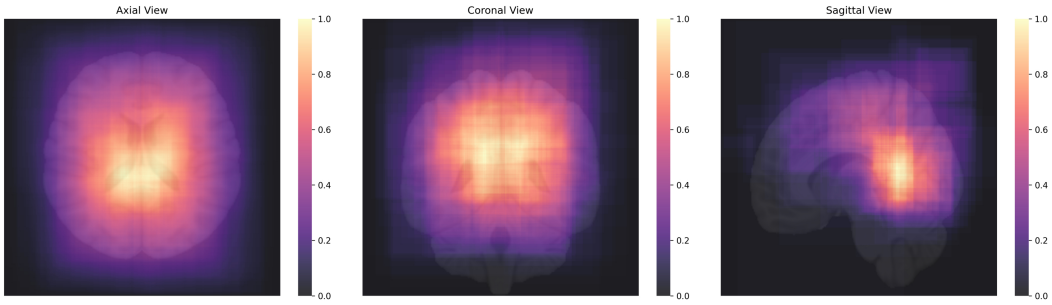

Figure 10: Heatmaps of bounding box locations aggregated across axial, coronal, and saggital views.

balanced sex distribution (min: 4 months; max: 87 years: mean: 34 years and 6 months). This heterogeneity emphasizes the need for models that generalize across anatomical, developmental, and demographic variations.

Spatial distribution of annotations. Figure 10 visualizes the anatomical spread of bounding boxes across axial, sagittal, and coronal planes. The heatmaps reflect the diversity of pathological presentations in the dataset, including cortical, subcortical, ventricular, brainstem, and cerebellar anomalies. This spatial variability introduces significant challenges for localization models, which must be robust to changes in context and anatomical orientation.

Bounding box properties. Figure 11 summarizes key properties of the annotated bounding boxes. The top panel reports the log-area distribution, indicating a wide range of lesion sizes—from small focal abnormalities to extensive pathology. The bottom panels show the number of boxes per image and a scatterplot of width versus height. Notably, a large fraction of cases contain multiple distinct findings, while many pathologies are highly non-square or irregularly shaped.

Implications. The demographic breadth and spatial diversity observed in NOVA mirror the complexity of clinical imaging workflows. These statistics underscore the difficulty of the anomaly

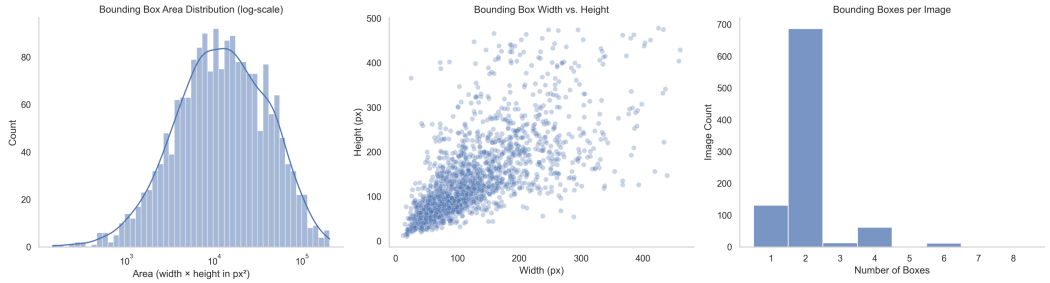

Figure 11: **Bounding box statistics.** Left: Log-area distribution of annotated bounding boxes. Center: Scatterplot of width versus height. Right: Histogram of the number of bounding boxes per image.

Table 7: Performance of generalist and medical VLMs on Task 1 (Anomaly Localization), 25-case NOVA subset.

Model(s)	mAP@50	mAP@50:95	RoDeO / Loc.	RoDeO / Shape
General Models				
Gemini 2.0 Flash	16.28	4.54	**53.06**	**34.45**
Qwen 2-VL-72B	11.11	4.12	42.45	23.82
Qwen 2.5-VL-72B	**18.97**	**8.02**	51.30	27.41
Medical Models				
CheXagent / MAIRA-2	0.00	0.00	0.00	0.00
HuatuoGPT Vision	8.33	1.94	49.25	32.90

localization task and highlight the importance of structured evaluation settings that go beyond synthetically simplified benchmarks.

C Evaluation of Generalist vs. Medical VLMs on NOVA (Task 1)

We benchmarked three medical vision–language models (VLMs)—CheXagent, MAIRA-2, and HuatuoGPT Vision in Table 7. We evaluated the results on a 25-case NOVA subset for *Task 1: anomaly localization*, using the same evaluation protocol as in the main experiments. The chest-X-ray–tuned models (CheXagent, MAIRA-2) did not generalize to brain MRI (mAP@50 = 0), while HuatuoGPT Vision achieved mAP@50 = 8.3. Generalist VLMs such as Gemini 2.0 Flash (16.3) and Qwen 2.5-VL-72B (19.0) remained substantially stronger. These results highlight the limited transferability of current modality-specific medical VLMs to brain MRI.

D Prompting and Evaluation Details of LLMs

This section details the prompting strategies, evaluation metrics, and model-specific configurations used in our benchmark for assessing clinical reasoning and factual consistency of large language and vision-language models on brain MRI tasks.

We design a comprehensive benchmark comprising three clinically grounded tasks to evaluate the capabilities of advanced language and vision-language models in neuroradiological image understanding. We show details in Section D.1:

- **Detection** — identifying and localizing abnormalities via bounding boxes.
- **Captioning** — generating structured radiological descriptions from MRI scans.
- **Reasoning** — performing differential diagnosis based on imaging findings and clinical history.

In addition, we define two evaluation procedures to assess the factual consistency and correctness of generated outputs (Section D.2). All models are evaluated in a zero-shot setting with fixed parameters

(temperature = 0.1, max output length = 2048 tokens) and a unified system prompt: *"You are a medical expert."*. The three clinical tasks are performed on four advanced models: GPT-4o, Gemini 2.0 Flash, Qwen 2.5-VL 72B, and Qwen 2.0-VL 72B. For evaluation tasks, we employ GPT-4o to conduct output assessment and consistency verification.

Note on DeepSeek-R1. As DeepSeek-R1 does not support direct visual input, we evaluated it using externally provided image descriptions. Specifically, we tested two setups: one using ground-truth (GT) captions and one using captions generated by GPT-4o. With GT captions, DeepSeek-R1 achieved 52.3% Top-1 and 67.9% Top-5 accuracy. When prompted with GPT-4o-generated captions, performance was 25.9% Top-1 and 41.6% Top-5. While these results highlight the potential of text-only diagnostic reasoning, they are not directly comparable to the other tested models due to differences in inputs.

D.1 Clinical Tasks

This component includes three tasks, each targeting a specific dimension of diagnostic reasoning:

(1) Abnormality Grounding. Given an MRI image, the model identifies abnormalities with bounding box coordinates and corresponding labels. Due to differences in coordinate conventions, we use model-specific prompt formats. For Qwen-series models, boxes are expressed as [x1, y1, x2, y2]; for Gemini models, we use [ymax, xmin, xmax, ymin]. Prompt templates and parsing logic are tailored accordingly to ensure compatibility across models.

Abnormality Grounding—Qwen Series:

Template 1: Abnormality Grounding Prompt

Return bounding boxes of any abnormal areas as JSON format.
If the image does not have the target, return the string: `"no target"`.
If detected, return a list of 2D bounding boxes around the target regions in the following JSON format:

```
[
  {"bbox_2d": [x1, y1, x2, y2], "label": "label"},
  ...
]
```

where `x1, y1` and `x2, y2` are the coordinates of the top-left and bottom-right corners of the bounding box, and `label` is the abnormality type.

Abnormality Grounding—Gemini:

Template 2: Abnormality Grounding Prompt

Return bounding boxes of any abnormal areas as JSON format.
If the image does not have the target, return the string: `"no target"`.
If detected, return a list of 2D bounding boxes around the target regions in the following JSON format:

```
[
  {"bbox_2d": [ymax, xmin, xmax, ymin], "label": "label"},
  ...
]
```

where `ymax, xmin, xmax, ymin` represent the coordinates of the bounding box corners, and `label` is the abnormality type.

(2) Medical Image Description. The model generates structured medical image findings directly from MRI scans. Prompts guide the model to describe the imaging modality, slice orientation, lesion location, and key visual abnormalities.

Template 3: Medical Image Description Prompt

System Prompt:
You are a highly skilled radiologist AI assistant. Your task is to analyze medical images with precision and generate accurate, concise diagnostic descriptions suitable for clinical use. Always prioritize clarity, accuracy, and domain-specific terminology in your responses.

Please carefully examine the provided medical image and perform a comprehensive, in-depth analysis. Generate a clear, concise description focusing on the **imaging modality**, **slice orientation**, **lesion location**, and any **notable abnormalities** observed.

Format to Follow:
- Answer:
`[Only output the final concise description result.]`

(3) Differential Diagnosis. The model receives a patient's clinical history and imaging findings as input. It outputs a list of five candidate diagnoses: one primary diagnosis and four plausible alternatives. The format is standardized to support automatic top-1 and top-5 accuracy evaluation. For a detailed example, see Sec. E.0.1.

Template 4: Differential Diagnosis Prompt

Please provide the most likely diagnosis along with four other possible differential diagnoses based on the following clinical history and MRI findings. Your output should be structured in JSON format.

Clinical History:
`"Clinical_History"`

MRI Findings:
`"MRI_Findings"`

Format to Follow:
`json`

```
{
  "most_likely_diagnosis": "Diagnosis name here",
  "other_possible_diagnoses": [
    "Diagnosis 1 here",
    "Diagnosis 2 here",
    "Diagnosis 3 here",
    "Diagnosis 4 here"
  ]
}
```

D.2 Evaluation

We design two evaluation tasks to quantitatively assess the quality of model outputs:

(1) Image Description Evaluation. This prompt extracts key clinical findings from model-generated MRI image descriptions and compares them to ground truth annotations. It emphasizes relevant anatomical, pathological, and imaging details, standardizing terms for direct comparison. Consistency is measured by agreement on the presence or absence of abnormalities, enabling accurate diagnostic evaluation. For a detailed example, see Sec. E.0.2.

Template 5: Image Description Evaluation Prompt

You are given two radiology reports: Ground Truth (GT) and Predicted (Pred). Your task is to extract and standardize medically important keywords from both reports.

Task: Extract keywords related to the following categories:

- Anatomical structures: e.g., brain regions, body parts.
- Imaging characteristics: e.g., hyperintensity, low density, enhancement, mass-like, signal changes.
- Disease or pathological findings: e.g., leukoencephalopathy, infarct, tumor.
- Negated findings: any finding explicitly stated as absent or negative, such as "no hemorrhage", "no mass" — keep the negation in the keyword.
- Imaging sequence and plane: e.g., T1, T2, FLAIR, DWI, sagittal, axial, coronal.

Standardization Rules:

- Normalize synonymous or semantically similar expressions into a single canonical form.
- Normalize anatomical mentions related to disease into their broader anatomical structures when appropriate.
- Ensure that after normalization, all terms that refer to the same concept are exactly string-equal, to support direct set-based comparison (e.g., for intersection/union using string matching).
- Prefer higher-level or broader terms when multiple expressions refer to variations of the same anatomical area (e.g., "inferior pointing of the ventricles", "ventricles slightly enlarged", and "ventricular dilation" should all be normalized to "ventricles").
- The goal is to eliminate variation in expression and granularity, so that conceptually equivalent phrases normalize to the same string.

Consistency

- GT and Pred are labeled as "normal" or "abnormal" based on their findings.
- Is_Consistent is true if both GT and Pred are either "normal" or both "abnormal".
- Is_Consistent is false if one is "normal" and the other is "abnormal".

Input:
GT = `"GT_INPUT"`
Pred = `"PRED_INPUT"`
Output Format (JSON):

```
{
  "Raw_Keywords": {
    "GT": ["keyword1", "keyword2", "..."],
    "Pred": ["keyword1", "keyword2", "..."]
  },
  "Standardized_Keywords": {
    "GT": ["standardized_keyword1", "standardized_keyword2", "..."],
    "Pred": ["standardized_keyword1", "standardized_keyword2", "..."]
  },
  "Consistency": {
    "GT": "normal" | "abnormal",
    "Pred": "normal" | "abnormal",
    "Is_Consistent": true | false
  }
}
```

Only return valid JSON with no extra text.

(2) Diagnosis Result Evaluation. This prompt evaluates predicted diagnoses against ground truth labels using top-1 and top-5 accuracy. It focuses on the core diagnostic entity, allowing synonyms and terminology variations, while ignoring differences in specificity or etiology unless the diagnosis is fundamentally different. For a detailed example, see Sec. E.0.3.

Template 6: Medical Diagnosis Evaluation

You are a professional medical diagnosis evaluation system. You will receive two inputs:

- **Ground Truth Diagnosis (GT)**: A single confirmed diagnosis.
- **Predicted Diagnosis (Pred)**: One most likely diagnosis and four additional possible diagnosis candidates.

Evaluation Rules

- Focus only on the **core diagnosis**, regardless of etiology or cause.
- Allow for synonyms and variations in medical terminology.
- If the same diagnostic entity (imaging pattern, pathological finding, or clinical condition) is present in the predictions, consider it correct.
- Do not penalize for differences in specificity or cause (e.g., idiopathic vs secondary), unless the disease is fundamentally different.

Input:
GT: `"GT_Diagnosis"`
Pred: `"Pred_Diagnosis"`
Output Format:
Return only JSON in the following structure:

```
{
  "Top_1": "Correct" | "Wrong",
  "Reason_for_Top1": "<your explanation>",
  "Top_5": "Correct" | "Wrong",
  "Reason_for_Top5": "<your explanation>"
}
```

Only return valid JSON with no extra text.

E Examples of Different prompts

E.0.1 Example of Differential Diagnosis Prompt

Example 1: Differential Diagnosis Example

Please provide the most likely diagnosis along with the other four possible diagnoses based on the following clinical history and MRI findings from the patient. The output should be in JSON format.

Clinical History:
"A 6-year-old boy came for MRI with complaints of delayed development, hypotonia, seizures. Birth history was normal and he was born to non-consanguineous parents. His younger sibling was normal. On clinical examination, the patient had multiple hypopigmented and hyperpigmented patches on limbs, back and chest."

MRI Findings:
"Slice 1: The image is an axial T2-weighted MRI of the brain. It shows hyperintense lesions in the periventricular white matter, suggestive of demyelination. The lesions are located adjacent to the lateral ventricles, which is characteristic of multiple sclerosis.
Slice 2: The image is a sagittal MRI scan of the brain. It shows a well-defined mass in the posterior fossa, likely affecting the cerebellum. There is no obvious midline shift or hydrocephalus. Further evaluation and correlation with clinical findings are recommended for diagnosis.
Slice 3: The image is an axial MRI scan of the brain. It shows a T1-weighted sequence. There are multiple small hyperintense lesions located in the periventricular white matter, which may suggest demyelinating disease or chronic small vessel ischemic changes.
Slice 4: The image is an axial FLAIR MRI of the brain. It shows hyperintense lesions in the periventricular white matter, which may indicate demyelination or other white matter pathology.
Slice 5: The image is an axial FLAIR MRI scan of the brain. There are hyperintense lesions visible in the periventricular white matter, which may suggest demyelination or other pathological processes."

Format to Follow:
```
json

{
  "most_likely_diagnosis": "Diagnosis name here",
  "other_possible_diagnoses": [
    "Diagnosis 1 here",
    "Diagnosis 2 here",
    "Diagnosis 3 here",
    "Diagnosis 4 here"
  ]
}
```

E.0.2 Example of Image Description Evaluation Prompt

Example 2: Image Description Evaluation Example

You are given two radiology reports: Ground Truth (GT) and Predicted (Pred). Your task is to extract and standardize medically important keywords from both reports.

Task: Extract keywords related to the following categories:

- Anatomical structures: e.g., brain regions, body parts.
- Imaging characteristics: e.g., hyperintensity, low density, enhancement, mass-like, signal changes.
- Disease or pathological findings: e.g., leukoencephalopathy, infarct, tumor.
- Negated findings: any finding explicitly stated as absent or negative, such as "no hemorrhage", "no mass" — keep the negation in the keyword.
- Imaging sequence and plane: e.g., T1, T2, FLAIR, DWI, sagittal, axial, coronal.

Standardization Rules:

- Normalize synonymous or semantically similar expressions into a single canonical form.
- Normalize anatomical mentions related to disease into their broader anatomical structures when appropriate.
- Ensure that after normalization, all terms that refer to the same concept are exactly string-equal, to support direct set-based comparison (e.g., for intersection/union using string matching).
- Prefer higher-level or broader terms when multiple expressions refer to variations of the same anatomical area (e.g., "inferior pointing of the ventricles", "ventricles slightly enlarged", and "ventricular dilation" should all be normalized to "ventricles").
- The goal is to eliminate variation in expression and granularity, so that conceptually equivalent phrases normalize to the same string.

Consistency

- GT and Pred are labeled as "normal" or "abnormal" based on their findings.
- Is_Consistent is true if both GT and Pred are either "normal" or both "abnormal".
- Is_Consistent is false if one is "normal" and the other is "abnormal".

Input:

- GT = "Coronal T1W with GADO: peripherical enhancement on post-contrast image."
- Pred = "Coronal T1-weighted MRI of the brain demonstrating multiple enhancing lesions, suggestive of metastatic disease."

Output Format (JSON):

```
{
  "Raw_Keywords": {
    "GT": ["keyword1", "keyword2", "..."],
    "Pred": ["keyword1", "keyword2", "..."]
  },
  "Standardized_Keywords": {
    "GT": ["standardized_keyword1", "standardized_keyword2", "..."],
    "Pred": ["standardized_keyword1", "standardized_keyword2", "..."]
  },
  "Consistency": {
    "GT": "normal" | "abnormal",
    "Pred": "normal" | "abnormal",
    "Is_Consistent": true | false
  }
}
```

Only return valid JSON with no extra text.

E.0.3 Example of Medical Diagnosis Evaluation Prompt

Example 3: Medical Diagnosis Evaluation Example

You are a professional medical diagnosis evaluation system. You will receive two inputs:

1. **Ground Truth Diagnosis (GT)**: A single confirmed diagnosis.
2. **Predicted Diagnosis (Pred)**: One most likely diagnosis and four additional possible diagnosis candidates.

Evaluation Rules

- Focus only on the **core diagnosis**, regardless of etiology or cause.
- Allow for synonyms and variations in medical terminology.
- If the same diagnostic entity (imaging pattern, pathological finding, or clinical condition) is present in the predictions, consider it correct.
- Do not penalize for differences in specificity or cause (e.g., idiopathic vs secondary), unless the disease is fundamentally different.

Input:
GT: `"Septo - optic dysplasia"`

Pred:

```
{
  "most_likely_diagnosis": "Craniopharyngioma",
  "other_possible_diagnoses": [
    "Optic Pathway Glioma",
    "Arachnoid Cyst",
    "Hydrocephalus",
    "Neurofibromatosis Type 1"
  ]
}
```

Output Format:
Return only JSON in the following structure:

```
{
  "Top_1": "Correct" | "Wrong",
  "Reason_for_Top1": "<your explanation>",
  "Top_5": "Correct" | "Wrong",
  "Reason_for_Top5": "<your explanation>"
}
```

Only return valid JSON with no extra text.

F IRB Documentation

Technische Universität München | Ethikkommission

Our sign: 2025-446-W-CB

TUM Universitäsklinikum Klinikum rechts der Isar
Institut für Diagnostische und Interventionelle Neuroradiologie
Herrn Prof. Dr. med. Benedikt Wiestler
Ismaninger Str. 22
81675 München
Germany

Munich, 07.08.2025

Consultation according to § 15 of the professional code of conduct for physicians in Bavaria
Confirmation of exemption

Title of study:	NOVA: A Benchmark for Anomaly Localization and Clinical Reasoning in Brain MRI
Applicant:	Prof. Dr. med. Benedikt Wiestler

Dear Prof. Dr. med. Benedikt Wiestler,

the Ethics Committee has reviewed your application dated 06.08.2025 on the basis of the documents submitted.

The Ethics Committee hereby confirms that professional advice pursuant to Section 15 of the Professional Code of Conduct for Physicians in Bavaria is not required for the submitted research project.

Yours sincerely

Prof. Dr. Georg Schmidt
Chairman of the Ethics Committee
Technical University Munich

The correspondence contains only a name and is valid without a signature.

Technische Universität München
Ethikkommission

Prof. Dr. Georg Schmidt
Vorsitzender

Prof. Dr. Kurt Ulm
Stellvertretender Vorsitzender

Prof. Dr. Christian Peschel
Stellvertretender Vorsitzender

Grillparzerstraße 16
D-81675 München

ethikkommission@mri.tum.de
www.ek-med-muenchen.de

Telefon:
+49 89 4140-7737

Fax:
+49 89 4140-4199