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

Anonymous Author(s)
Affiliation
Address
email

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 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. Finally, Section D includes concrete prompting examples to support reproducibility.

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 1). Annotators were

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

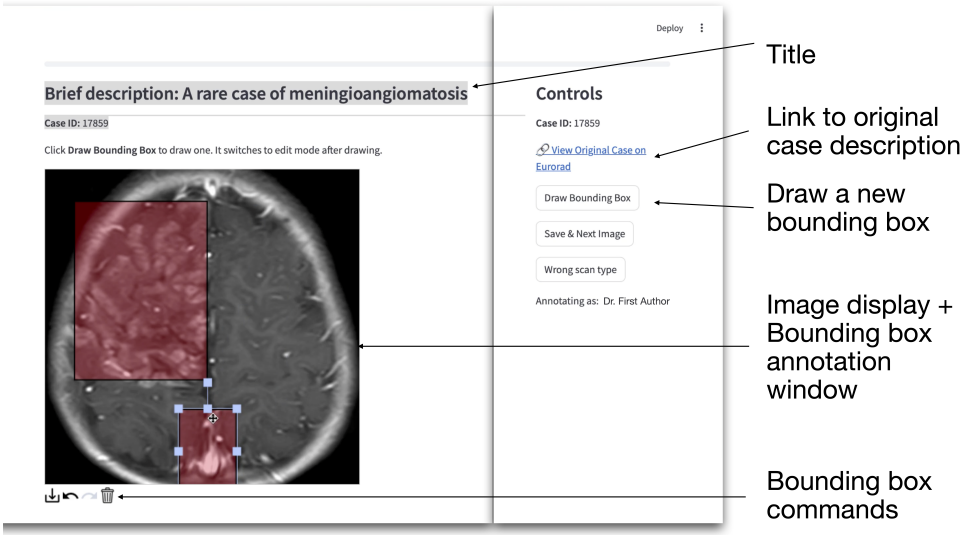

Figure 1: Each case includes a brief clinical description and a link to the full Eurorad entry. Annotators mark pathologically relevant regions using a custom bounding box tool. Controls are optimized for clinical workflows, enabling rapid annotation and review.

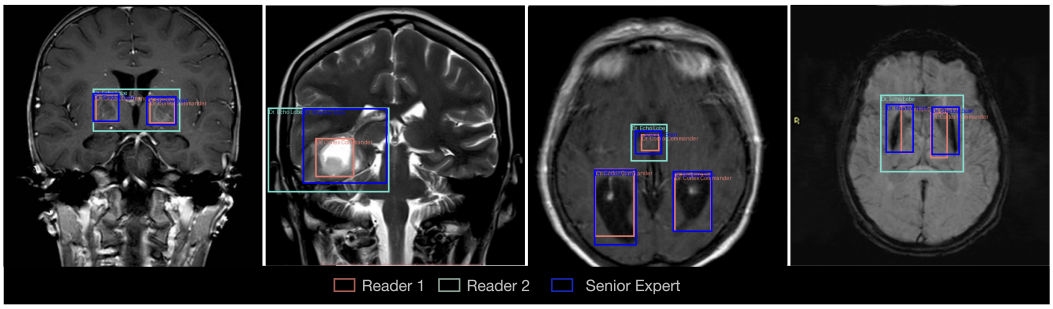

Figure 2: **Examples of annotation disagreement.** Each image shows bounding boxes from two independent annotators (Reader 1 in salmon and Reader 2 in teal) and the adjudicated expert bounding box (blue). These cases illustrate scenarios where annotators either disagreed on lesion boundaries or identified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

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 2. This set of 188 images served as the basis for estimating each annotator's agreement with the expert.

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.

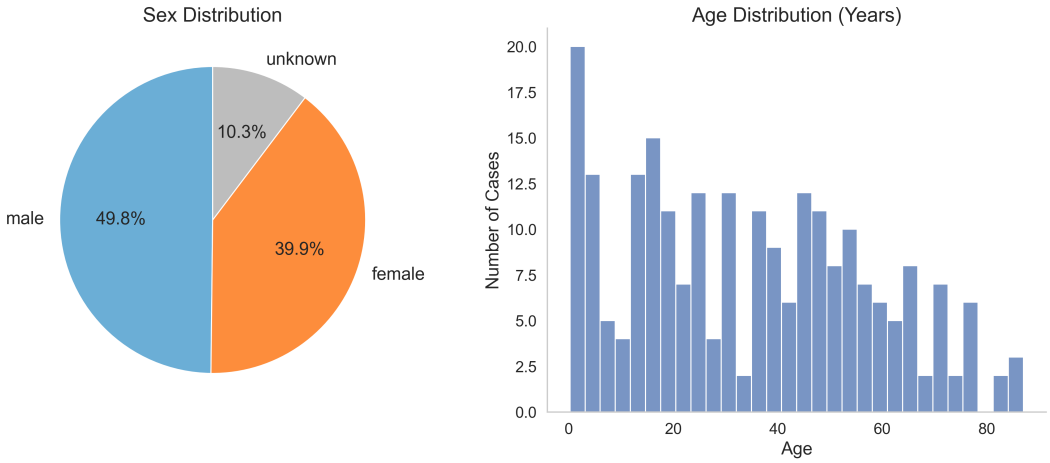

Figure 3: **Demographics.** Left: Sex distribution of cases, with a nearly balanced male-to-female ratio and a subset of cases with unknown sex. Right: Histogram of patient ages showing broad coverage across pediatric, adult, and elderly populations.

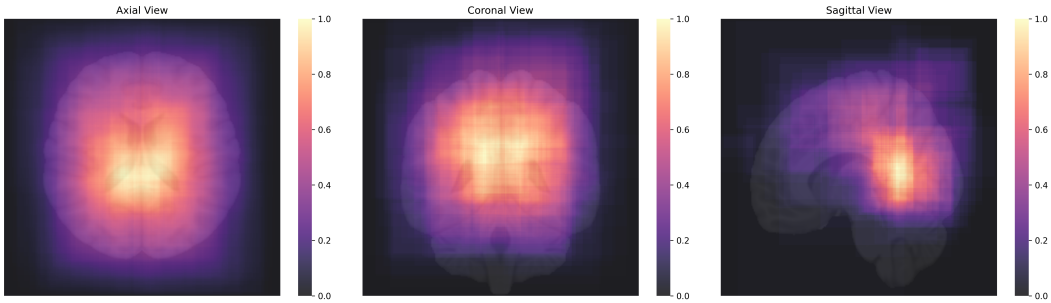

Figure 4: Heatmaps of bounding box locations aggregated across axial, coronal, and saggital views.

Patient demographics. Figure 3 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 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 4 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 5 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 localization task and highlight the importance of structured evaluation settings that go beyond synthetically simplified benchmarks.

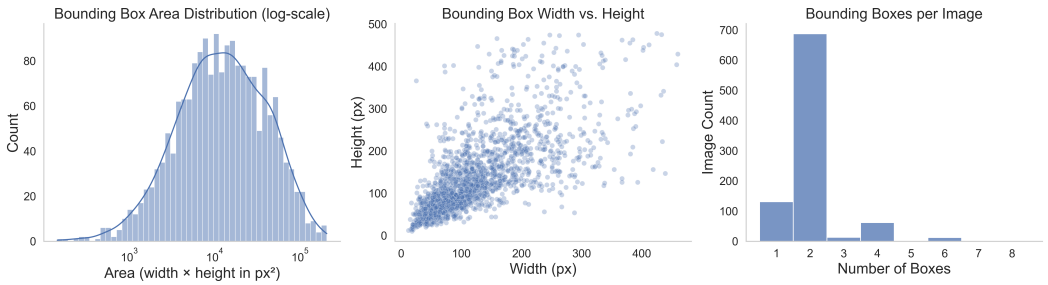

Figure 5: **Bounding box statistics.** Left: Log-area distribution of annotated bounding boxes. Center: Scatterplot of width versus height. Right: Histogram of the number of bounding boxes per image.

C 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 C.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 C.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.

C.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. D.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"
  ]
}
```

C.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. D.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. D.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.

D Examples of Different prompts

D.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"
  ]
}
```

D.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.

D.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.