Visual and Domain Knowledge for Professional-level Graph-of-Thought Medical Reasoning

Rina Bao [1] **Shilong Dong** [2] **Zhenfang Chen** [3] **Sheng He** [1] **Patricia Ellen Grant** [1] **Yangming Ou** [1]

Abstract

Medical Visual Question Answering (MVQA) requires AI models to answer questions related to medical images, offering significant potential to assist medical professionals in evaluating and diagnosing diseases, thereby improving early interventions. However, existing MVQA datasets primarily focus on basic questions regarding visual perception and pattern recognition, without addressing the more complex questions that are critical in clinical diagnosis and decision-making. This paper introduces a new benchmark designed for professional-level medical reasoning, simulating the decision-making process. We achieve this by collecting MRI and clinical data related to Hypoxic-Ischemic Encephalopathy, enriched with expert annotations and insights. Building on this data, we generate clinical question-answer pairs and MRI interpretations to enable comprehensive diagnosis, interpretation, and prediction of neurocognitive outcomes. Our evaluation of current large vision-language models (LVLMs) shows limited performance on this benchmark, highlighting both the challenges of the task and the importance of this benchmark for advancing medical AI. Furthermore, we propose a novel "Clinical Graph of Thoughts" model, which integrates domain-specific medical knowledge and clinical reasoning processes with the interpretive abilities of LVLMs. The model demonstrates promising results, achieving around 15% absolute gain on the most important neurocognitive outcome task, while the benchmark still reveals substantial opportunities for further research innovation. Project page: https://github.com/i3-research/HIE-Reasoning

[1]Boston Children's Hospital and Harvard Medical School, Boston, USA [2]New York University [3]MIT-IBM Watson AI Lab. Correspondence to: Zhenfang Chen <hiereasoning@gmail.com>, Yangming Ou <yangming.ou@childrens.harvard.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

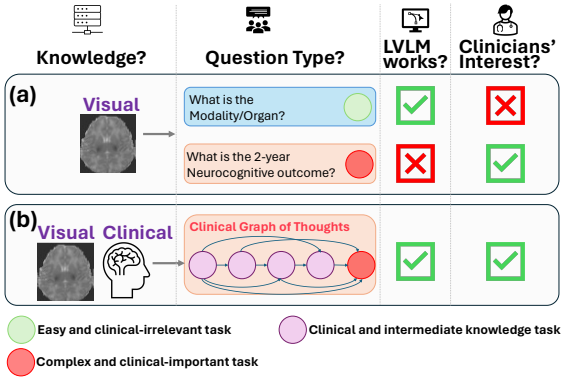


Figure 1. Differences between existing MVQA (a) and the proposed dataset (b). Compared to existing methods that focus on simple, clinically irrelevant tasks (green circle), our approach targets more complex, clinically important tasks that align with clinicians' interests—tasks that traditional LVLMs fail to address. To tackle this challenge, we propose *Clinical Graph of Thought Model*, a model that decomposes these complex, clinically significant tasks (red circle) into a graph of simpler clinical and intermediate knowledge tasks (purple circle), leveraging both visual and clinical knowledge.

1. Introduction

Large Vision-Language Models (LVLMs) have achieved significant success in areas like natural image and video analysis (Bordes et al., 2024; OpenAI, 2023; 2024) and natural language processing (NLP) (Ouyang et al., 2022; Touvron et al., 2023). These models leverage advancements in neural networks, large-scale datasets, and computational resources, enabling capabilities such as automated image captioning and complex visual question answering. However, in the medical field, LVLMs are still in their early stages of development (Hu et al., 2024; He et al., 2023), with research progressing but not yet matching the success seen in non-medical domains. A key challenge lies in the limited availability of suitable medical data, hindering their broader application.

Many medical datasets for LVLMs (Alkhaldi et al., 2024; Li et al., 2024; Moor et al., 2023; Saab et al., 2024) are organized in a visual question-answering format, which provides natural inputs for LVLMs and offers a flexible way

to test their capabilities across different aspects of medical data. However, most of these datasets are labeled by general annotators. As shown in Fig. 1, they focus on general question-answer pairs (e.g., *"Q: What is the organ? A: Brain"*), primarily assessing basic pattern recognition and object classification. It remains uncertain how effectively these datasets can aid AI models in understanding complex MRI tasks that clinical experts prioritize, such as *"What is the predicted two-year neurocognitive outcome for this individual?"* and enhancing clinical workflows (see Figure 3 (B)). This paper introduces a benchmark designed to challenge LVLMs' performance in professional-level medical reasoning, focusing on questions relevant to medical professionals to support diagnossis and prognosis.

Developing such a professional-level medical reasoning benchmark poses challenges. Medical data is often scarce and subject to strict sharing and usage limitations. Additionally, labeling this data is resource-intensive and costly, requiring medical professionals like radiologists and neonatologists to carefully examine MRI images, consult with patient families, and compose diagnostic reports. To build this benchmark, we compiled a decade's worth of MRI images and clinical expert-verified interpretations related to Hypoxic-Ischemic Encephalopathy (HIE, a neonatal brain dysfunction that occurs in 1-5/1000 neonates) from 133 individuals, in collaboration with experts in pediatric neuroradiology, neonatal MRI interpretation, neonatology, and neonatal neurology. This interdisciplinary team brings critical expertise in managing neonates with HIE, resulting in a high-quality benchmark closely aligned with real-world neonatal care and long-term neurodevelopmental outcomes.

Using this raw annotated data, we generated question-answer pairs along with a comprehensive MRI interpretation summary that reflects expert reasoning, enabling us to evaluate LVLMs' capacity to comprehend these questions, identify correct answers, make accurate decisions, and predict future neurocognitive outcomes. We defined the questions with input from medical experts and obtained the answers by parsing clinical reports.

After acquiring raw MRI reports from experts, we formulated tasks that address the priorities of medical professionals, as opposed to general pattern recognition questions in existing benchmarks. We modeled clinical decision-making workflows and designed six tasks, as shown in Fig. 2. Each task was meticulously validated by an expert neonatal radiologist or neonatologist, depending on the task, with over 30 years of experience, ensuring clinical relevance, significance, and accuracy for model training. We prompted large language models to generate answers by parsing medical reports and manually verified the responses for accuracy.

This approach establishes a benchmark that mirrors essential components of professional-level MRI assessment as shown in Fig. 2. We defined six tasks: *Lesion Grading*, *Lesion Anatomy*, *Lesion in Rare Regions*, *Neurocognitive Outcome Prediction*, and *MRI Interpretation Summary*. Each task targets distinct aspects of clinical significance. Specifically, **Lesion Grading** assesses HIE lesion severity; **Lesion Anatomy** identifies the anatomical locations of lesions; **Lesion in Rare Locations** determines whether injuries are present in typical or atypical regions; **MRI Injury Score** generates an overall brain injury severity score, widely used as a biomarker for predicting adverse two-year outcomes and is implemented in many HIE clinical trials worldwide (Laptook et al., 2017b; Shankaran et al., 2017b); **Neurocognitive Outcome Prediction** forecasts two-year neurocognitive outcomes, crucial since 30%-50% of HIE patients experience adverse outcomes (Graham et al., 2008; Lee et al., 2013; Weiss et al., 2019); and **MRI Interpretation Summary** provides brief, structured MRI summaries from neonatal radiologists and neonatologists' views, highlighting key findings on neonatal brain injury. Totally, our benchmark comprises 749 professional question-answer pairs and 133 MRI interpretation summaries derived from unique MRI images of 133 individuals.

We evaluated a range of state-of-the-art general and medical LVLMs on our benchmark, finding while these models perform well on general datasets focusing on visual perception and pattern recognition, they demonstrate limitations on our benchmark, which requires both visual perception and specialized medical knowledge for reasoning.

To address this gap, we propose the Clinical Graph of Thought Model (CGoT), which emulates the diagnostic process through clinical knowledge-guided graph-of-thought prompting. This approach not only enhances model performance but also improves transparency. Additionally, CGoT incorporates domain-specific clinical knowledge as input both visually and textually to strengthen the predictive power of LVLMs. Experimental results demonstrate a significant improvement, with performance gains exceeding over 15% across tasks, although further progress is needed.

In summary, the key contributions of this paper are: (1) the creation of the novel HIE-Reasoning benchmark, which replicates clinical decision-making workflows to evaluate LVLMs in professional-level medical reasoning, representing the first medical reasoning benchmark that combines clinical visual perception with professional-level medical knowledge; (2) a comprehensive evaluation of state-of-the-art general and medical LVLMs, revealing their limitations in medical domain knowledge; and (3) the introduction of CGoT, a clinically guided model that mimics the clinical decision-making process, integrating medical expertise with LVLMs to enhance decision-making support.

Table 1. Comparison of Medical VQA Datasets. HIE-Reasoning uniquely integrates clinician-labeled, profession-level knowledge and requires AI models to predict future outcomes, crucial for clinical early interventions.

Dataset	General Medical Knowledge	Disease-Specific Knowledge	Future Prediction	Data Source	Clinical Workflow Related
VQA-Med (Ben Abacha et al., 2019)	✓	Weak	×	Medical Database	Low
OmniMedVQA (Hu et al., 2024)	✓	Weak	×	Medical Database	Low
VQA-RAD (Lau et al., 2018)	✓	Weak	×	Medical Students & Fellows	Medium
PathVQA (He et al., 2020)	✓	Weak	×	Textbooks	Low
SLAKE (Liu et al., 2021)	✓	Weak	×	Medical Database	Low
HIE-Reasoning	✓	Strong	✓	Clinical Report	High

2. Related Work

Large Vision Language Models for VQA. LLMs provide a promising performance in natural language processing and it has been extended to multi-modality LVLMs, integrating the textural information from texts and visual information from images. Most famous foundation models start to support the texts and images, such as GPT4o (OpenAI, 2024) and Gemini (Saab et al., 2024) which accept as input any combination of text, audio, image, and video and generate any combination of text, audio, and image outputs. These models are trained on general vision-language tasks that may or may not include medical data. Medical vision-language tasks require strong domain knowledge and professional skills. Thus, efforts have been made to integrate medical specific knowledge into foundation models by finetuning the general LVLMs, such as the Med-Flamingo (Moor et al., 2023), LLava-Med (Li et al., 2024), and MiniGPT-Med (Alkhaldi et al., 2024). Med-Flamingo is a multimodal model adapted to the medical domain by pre-training on paired and interleaved medical image-text data from publications and textbooks (Moor et al., 2023). LLava-Med extends multimodal instruction-tuning to the biomedical domain as a biomedical multimodal conversational assistant (Li et al., 2024). MiniGPT-Med (Alkhaldi et al., 2024) is a vision-language model derived from large-scale language models and tailored for medical applications which shows superior performance in VQA benchmarks on medical imaging. In this paper, we evaluate models on medical disease by asking complex and clinical-important questions that require specialized knowledge from medical experts.

Benchmark Datasets for VQA in Medical Images. Benchmark datasets are important to evaluate the performance of AI models for visual question answering and a few medical benchmarks have been developed, such as VQA-med (Ben Abacha et al., 2019), OmiMed-VQA (Hu et al., 2024), SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018) and Path3000 (He et al., 2020). The VQA-med (Ben Abacha et al., 2019) is designed to answer four categories of clinical questions: Modality, Plane, Organ System, and Abnormality. OmiMed-VQA (Hu et al., 2024) provides five different question types: Modality Recognition, Anatomy Identification, Disease Diagnosis, Lesion grading, and Other Biological Attributes. SLAKE (Liu et al., 2021) contains only two types of questions about organ and disease on medical images. VQA-RAD (Lau et al., 2018) focuses on visual attributes of medical images, such as modality, organ, color, size of objects, counting, etc. PathVQA (He et al., 2020) supports specific questions on pathology images.

These datasets are designed to answer the question from the image content without requiring deep professional medical knowledge, domain-specific inference, or clinical-workflow-related knowledge. Questions about general information on medical images such as the modality, organ, and disease information are easy and clinical-workflow-irrelevant (see Fig. 1(a) and Table 1), which are not clinicians-interested ones. Other types of questions, such as the complex and clinical-important task (see Fig. 1(b) and Table 1) are the core questions for clinicians but failed for the traditional LVLMs. To fill this gap, we proposed a template benchmark, named HIE-Reasoning, focusing on answering the clinicians-interested questions using both general medical knowledge and strong disease-specific knowledge for future outcome prediction, which cannot be performed on other datasets (Table 1).

Prompt Engineering in Large Language Models. LLMs face limitations in complex reasoning tasks that require specialized knowledge (Sun et al., 2023). These limitations can be addressed using prompting techniques, such as think-of-graph (ToG) (Sun et al., 2023) and chain-of-thought (CoT) (Wei et al., 2022), which encourage LLMs to generate step-by-step solutions for complex tasks. ToG treats the LLM as an agent to interactively explore related entities and relations on knowledge graphs (Sun et al., 2023), while CoT breaks down multi-step problems into intermediate steps. These prompt engineering techniques are primarily used for VQA on text inputs. In this paper, we extend this approach by introducing a clinical graph-of-thought prompting method for medical VQA, answering complex and clinical-important tasks by decomposing it into a graph of thoughts with easy clinical and intermediate knowledge tasks (see Fig. 1 and Fig. 3).

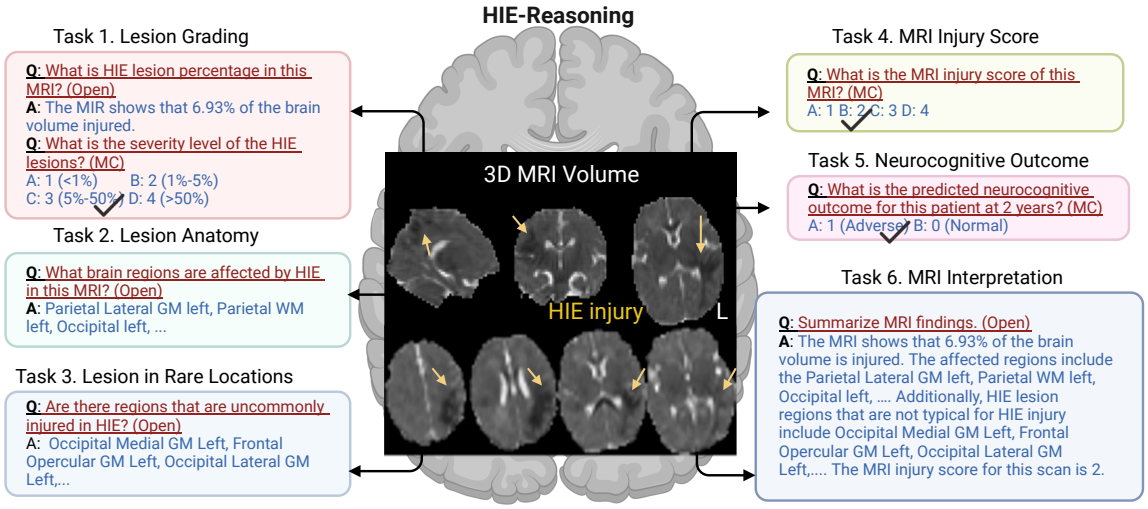


Figure 2. HIE-Reasoning Dataset and Task Overview. This figure illustrates the HIE-Reasoning dataset structure, comprising six tasks designed to assess reasoning capabilities for HIE MRI interpretation and outcome prediction using MRI data. This dataset supports the development of reasoning models by providing both open-ended (open) and multiple-choice (MC) questions, encouraging comprehensive understanding of HIE MRI and crucial aspects of MRI towards prognosis.

3. HIE-Reasoning Dataset

We define a series of tasks, illustrated in Fig. 2, for LVLMs to perform professional-level clinical reasoning.

Task 1. Lesion Grading. This task quantifies brain injury by estimating the percentage of brain volume affected by HIE lesions and assessing the lesion severity extents. This task outputs lesion volume percentage and severity. Accordingly, we defined two specific tasks as in Fig. 2: one to evaluate lesion percentage and another to categorize lesion severity into four levels: level 0 ($< 1\%$ brain injury), level 1 (1%-5%), level 2 (5%-50%), and level 3 ($> 50\%$). Manual lesion annotations served as the ground truth (Bao et al., 2025b) and were used to generate the ground-truth answers.

Task 2. Lesion Anatomy. This task identifies specific brain regions affected by lesions. The brain is divided into 62 regions of interest (ROIs) based on standard anatomical references (Doshi et al., 2016; Morton et al., 2020), and the task outputs the lesioned regions among these ROIs. Such task is critical for accurately predicting functional impairments, assessing injury severity, and informing prognosis in conditions like HIE. For example, lesions in the basal ganglia or thalamus are strongly associated with adverse neurocognitive outcomes.

Task 3: Lesions in Rare Locations. This task identifies lesions caused by HIE and categorizes affected regions as common or uncommon, helping to determine if additional attention is needed for the patient. Specifically, (Bao et al., 2025b) provides a lesion atlas generated by aggregating lesion masks from each patient, resulting in statistical lesion maps across the HIE cohort. This atlas highlights brain regions most commonly affected by HIE, aligning with clinical knowledge on areas frequently impacted by HIE injuries (Shankaran et al., 2012; Weiss et al., 2019). A 22.5% threshold, recommended by an experienced radiologist, was applied to generate commonly affected regions. The brain regions most frequently impacted by HIE include the basal ganglia, internal capsules, thalamus, perirolandic cortex/subcortical white matter, temporal lobes, cerebral white matter, brainstem, and vermis. For each case, primary injury regions (i.e., commonly affected areas) are identified, with uncommon injury sites noted when affected regions fall outside the typical injury profile.

Task 4. MRI Injury Score. This task outputs an overall injury score of MRI, providing a standardized measure of injury severity to guide treatment and predict outcomes. Expert scoring of neonatal MRI is currently employed in clinical trials to predict 18- to 22-month outcomes (Laptook et al., 2017b; Shankaran et al., 2017b). The National Institute of Child Health and Human Development (NICHD) Neonatal Research Network (NRN) injury scoring system (Shankaran et al., 2012; 2015) is the most widely used MRI injury score for assessing brain injury severity in HIE. To facilitate machine learning modeling, we consolidated the original 6-level score into 4 levels, instructed by radiologist, as certain distinctions within the 6 levels were clinically ambiguous and difficult to quantify for machine learning purposes.

Task 5. Two-Year Neurocognitive Outcome. This task predicts the patient's neurocognitive outcome at two years, aiding clinicians in anticipating long-term impacts and planning

appropriate interventions. Specifically, this task outputs a binary label where the two-year neurocognitive outcome is categorized as either normal (0) or adverse (1), based on clinical criteria and NRN recommendations (Laptook et al., 2017b; Shankaran et al., 2017b). An outcome is classified as adverse if any of the following conditions are met: a Bayley-III cognitive score below 85 in any domain, a GMFCS level between 2 and 5, blindness, hearing impairment, or if the patient is deceased in two years. Otherwise, the outcome is considered normal (Laptook et al., 2017b; Shankaran et al., 2017b; Bao & Ou, 2024).

Task 6. MRI Interpretation Summary. The initial Tasks 1 to 4 cover the key aspects of MRI interpretation: estimating lesion volume, identifying lesion locations, detecting atypical lesion locations, and rating injury severity. These tasks are structured based on a neonatal MRI summary template recommended by radiologists, enabling the generation of comprehensive MRI interpretations for patients.

Dataset Details. The HIE-Reasoning is the first publicly available HIE dataset that integrates MRIs, clinical information, neurocognitive outcomes, and includes question-answer (QA) pairs along with comprehensive MRI interpretation summaries. It was retrospectively collected from Massachusetts General Hospital (MGH) between 2001 and 2018. The dataset includes high-quality MRIs acquired within the first 0-14 days after birth from neonates with HIE, totaling 133 MRI scans. Expert lesion annotations are available for all cases, provided by one clinical fellow with over 3 years of experience and three additional neuroradiologists with over 5, 5, and 30 years of experience, respectively. Neurocognitive outcome data was retrospectively gathered from follow-ups conducted by more than four neonatologists and neurologists, assessing outcomes at least 18-22 months post-birth. In total, the dataset comprises 749 question-answer pairs and 133 MRI comprehensive interpretation summary.

Justification for Small Sample Size. While this sample size may appear modest for contemporary machine learning applications, several factors underscore its significance. First, HIE diagnosis and treatment demand specialized expertise available only at select tertiary care centers, making large-scale data collection inherently challenging. Second, this benchmark emerges from decades of clinical collaboration across multiple specialties' efforts, establishing its unique value in neonatal care research. The dataset's comprehensive nature, depth of clinical knowledge, combining diverse clinical indicators and expert annotations, compensates for its size limitations and provides a foundational resource for advancing medical AI in this critical domain.

4. Proposed Clinical Graph of Thought Model

A straightforward approach to addressing the HIE-Reasoning task is to directly input MRI images into an LVLM to generate a predicted outcome as shown in Fig. 3 (A). However, without integrated clinical reasoning and domain-specific knowledge, the LVLM's performance is often no better than random chance due to its lack of essential interpretative capabilities as shown in our experimental analysis in Sec 5. To overcome this limitation, we propose Clinical Graph of Thought Model (CGoT), which incorporates clinical knowledge into the LVLM, guiding it through a clinician-like diagnostic process, as depicted in Fig. 3 (B)-(C). This framework yields more reliable neurocognitive outcome predictions. This novel approach demonstrates how structured clinical reasoning, when coupled with advanced AI, can elevate medical image interpretation to new levels of precision and reliability.

As detailed in Fig. 3 (C), CGoT systematically incorporates insights from previous clinical evaluations, mimicking real-world diagnosis workflows. The framework works with two innovations, (1) clinical graph-structured prompting for medical reasoning and (2) visual and professional-level clinical knowledge as input.

4.1. Clinical Graph of Thought for Reasoning

In practice, a neonatologist would interpret a radiologist's report to assess brain injury severity. At the core of CGoT is a structured "reasoning graph of thought" that mirrors clinicians' step-by-step clinical diagnosis and decision-making process. Similarly, CGoT leverages the knowledge gained from each preceding task to guide the LVLM, enabling it to address complex questions by progressively refining its understanding in a stepwise manner.

CGoT consists of six key tasks that replicate the domain-specific expertise of medical specialists. These tasks include lesion grading (informed by radiologists and neonatologists), anatomical localization (radiologist-driven), injury scoring (led by neonatologists), and two-year outcome prediction (integrating insights from radiologists, neonatologists, and neurologists). Each task builds upon the previous one, creating a reasoning pipeline that aligns with the sequential clinical evaluation process. Finally, the outputs of these tasks are integrated and structured to form the MRI interpretation summary by MRI summary template. Fig. 3 (B) depicts CGoT 's reasoning graph, which integrates critical diagnostic elements and emulates the specialized expertise of clinical roles, forming an interactive, layered diagnostic approach. For example, in the lesion anatomy task, CGoT, after receiving clinical guidelines, enables the LVLM to emulate a radiologist's role, identifying regions affected by HIE with statements like, "I can identify HIE-injured brain regions in this patient."

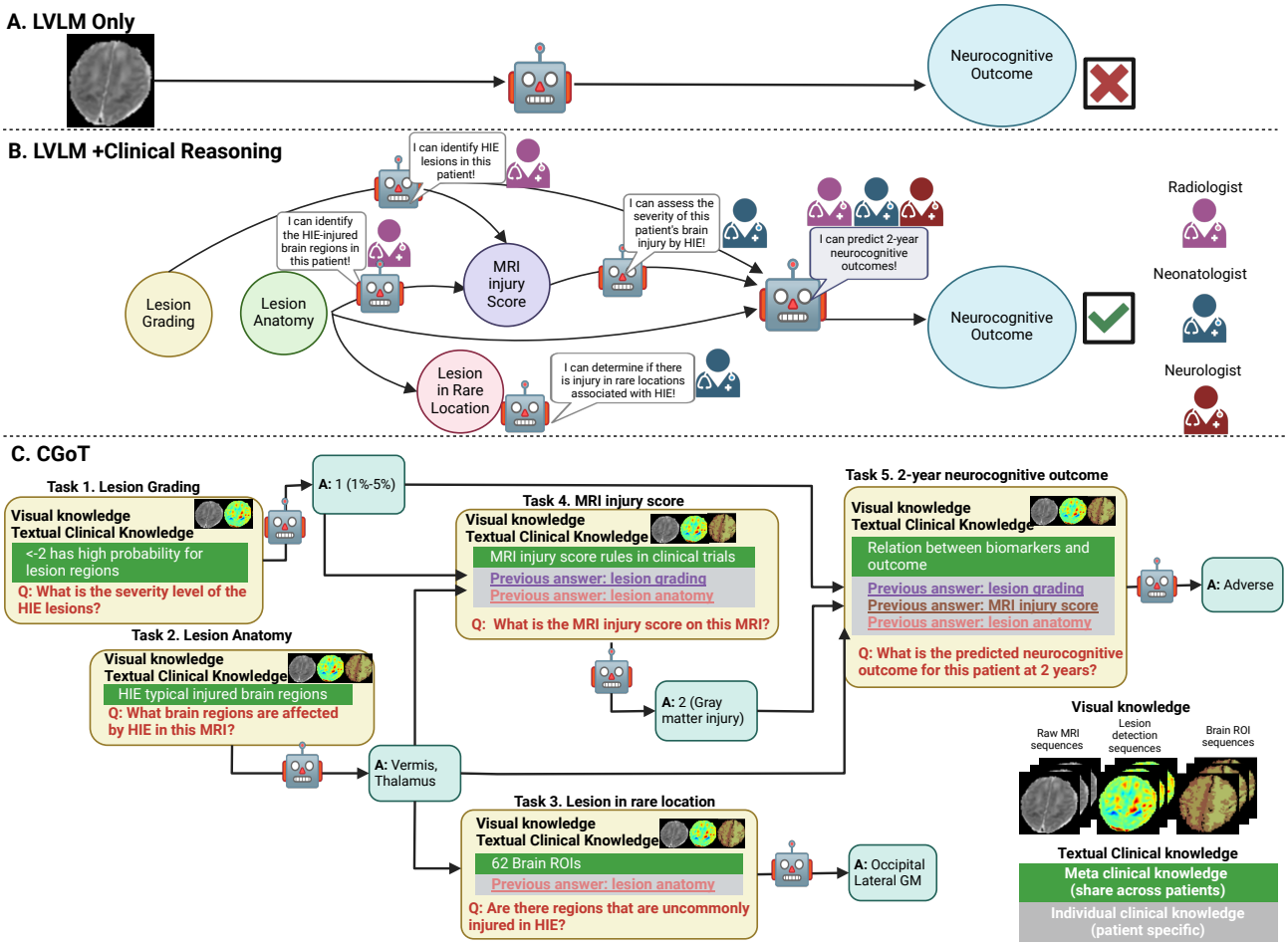


Figure 3. (A) The LVLM-only approach, which directly inputs MRIs for outcome prediction, lacks structured clinical reasoning and thus performs poorly in outcome accuracy. (B) The proposed CGoT incorporates clinical reasoning steps that mimic a clinician's diagnostic workflow. (C) CGoT organizes tasks into a sequential, clinician-inspired diagnostic flow, making more reliable neurocognitive outcome predictions by progressively integrating visual and textual clinical knowledge across interrelated diagnostic steps.

4.2. Clinical Knowledge Representation

CGoT relies on two primary knowledge sources for each task: visual knowledge and textual clinical knowledge. Visual knowledge includes raw MRI data and derived visual features, such as lesion probability and brain anatomical region information. Textual clinical knowledge is divided into essential clinical contextual knowledge for answering the current question and individual clinical knowledge, which is obtained from the previous task's output. This approach allows CGoT to sequentially build understanding and tackle each task's question by leveraging preceding answers, thereby simulating a clinician's reasoning process. Below we dive into the knowledge detail.

4.2.1. Visual Knowledge

There are three types of visual knowledge: raw MRI, lesion probability, and patient brain anatomy.

Raw MRI. Apparent diffusion coefficient (ADC) maps serve as the primary imaging modality for identifying HIE abnormalities by radiologists (Douglas-Escobar & Weiss, 2015; Wei et al., 2019; Liauw et al., 2009) and are used as visual knowledge input for the LVLM.

Lesion Probability. Neuroradiologists detect acute brain injury from HIE by identifying regions with abnormally low ADC values, which indicate reduced water diffusion due to ischemic necrosis (Weiss et al., 2019). However, distinguishing these regions from normal variation is challenging due to variability in ADC values across different brain regions (Ou et al., 2017; Sotardi et al., 2021). To address this, Bao *et al.* (Bao et al., 2025b) introduced Z_{ADC} maps to normalize ADC values across brain voxels. By applying a threshold of -2 on Z_{ADC}, a lesion prediction map is generated, offering performance comparable to more complex deep learning methods for lesion detection (Bao et al., 2025a;b). Details

of this information are provided in the appendix.

Brain Anatomy. Knowledge of patient-specific brain anatomy is essential for accurate interpretation. Brain regions are identified using a standard set of predefined 62 ROIs. To enable the LVLM to understand brain anatomy, a normalized brain structure map with these 62 ROIs in standard brain space was transformed to each patient's individual space using the DRAMMS tool (Ou et al., 2011). This patient-specific anatomy map provides the LVLM with the same anatomical knowledge a radiologist would use, enhancing its ability to interpret brain region localization during tasks. Details of the ROIs in the standard atlas space are provided in the appendix.

4.2.2. Textual Clinical Knowledge

Answering expert-level questions in HIE diagnosis requires advanced clinical knowledge and patient-specific insights. We define two types of textual clinical knowledge: meta clinical knowledge and individual clinical knowledge.

Meta Clinical Knowledge. Meta knowledge includes general disease-related insights, such as brain anatomy, lesion distributions, relationships between MRI biomarkers and outcomes, and knowledge from radiology, neonatology, and neurology. This shared knowledge across patients provides essential context for HIE diagnostic tasks. A summary of the relevant meta knowledge is provided, with further details available in the appendix.

Individual Clinical Knowledge. Individual knowledge is patient-specific and essential for personalized diagnostic tasks. Derived dynamically through the task's reasoning process, it allows the model to incorporate personalized clinical insights. As shown in Figure 3, meta and individual knowledge components are used selectively, ensuring tailored responses for each task. Additional details about task-specific knowledge are in the appendix.

5. Experiments

We conduct comprehensive experiments to demonstrate the effectiveness of the proposed benchmark, HIE-Reasoning, and model, Clinical Graph of Thought Model.

5.1. Implementation Details

Model Details. We begin by performing zero-shot evaluations on six representative large vision-language models (LVLMs): three general-purpose LVLMs (Gemini-1.5-Flash (Team, 2024), GPT4o-Mini (OpenAI, 2024), and GPT4o (OpenAI, 2024)) and three medical LVLMs (MiniGPT4-Med (Alkhaldi et al., 2024), LLava-Med (Li et al., 2024), and Med-Flamingo (Moor et al., 2023)). All settings and hyperparameters are configured according to the specifications of the released versions. We provide more model details and all clinical prompts in the appendix.

Evaluation Metrics. We design task-specific evaluation metrics in HIE-Reasoning to comprehensively assess model performance as summarized in Table 2. For *Lesion Grading* task, we use accuracy and Mean Absolute Error (MAE) to assess models' prediction performance of severity levels and lesion percentages, respectively. The F1 Score is applied to measure the classification and retrieval quality of model-predicted lesion regions compared to ground truth in tasks such as *Lesion Anatomy Identification* and *Lesion in Rare Locations*. To evaluate predictions for *MRI Injury Scores*, we use accuracy, while for the *two-year Outcome* task, we utilize the average inter-class accuracy (defined as the mean accuracy across all output categories) to address imbalanced and biased ground truth label distribution. Finally, the ROUGE-L Score (Lin, 2004), capturing content overlap and fluency by comparing generated answers to expert references, commonly in medical summarization tasks with LLM (Tang et al., 2023), is employed to measure the quality of the *MRI interpretation summary* task.

5.2. Result Analysis

Performance of Different Baseline LVLMs. As shown in Table 2, when original MRI image slices and task descriptions are directly input into the LVLMs, their performance is either comparable to or worse than random guessing. This indicates a lack of sufficient clinical knowledge to provide accurate answers. Med-Flamingo fails to address hallucination issues in the tasks Lesion Anatomy and Lesion in Rare Locations, generating repetitive and meaningless responses; consequently, its performance is not reported for these tasks. Additionally, GPT-4o and GPT-4o-mini fail to answer certain questions due to their conservative alignment, which rejects responses to questions with high uncertainty. For a comprehensive evaluation, we aggregate their performance metrics as GPT4o-Series. These results highlight the fact that, despite being state-of-the-art models, these LVLMs do not inherently acquire or retain professional-level medical knowledge during training. This reinforces the importance of our proposed HIE-Reasoning for evaluating LVLMs' capabilities in professional-level medical reasoning.

Effectiveness of CGoT. We also compare our CGoT with the corresponding baselines in Table 2. The results demonstrate that CGoT significantly outperforms the baselines, particularly in predicting the critical two-year outcome, which is the primary task of HIE-Reasoning and supports medical experts in HIE prognosis. In addition to the two-year outcome, we observe a substantial performance gap between CGoT and the naive direct-output baselines across tasks such as *Lesion Grading*, *Lesion Anatomy*, and *MRI Injury Score*. This underscores the value of incorporating additional clinical knowledge into LVLMs. Overall, these findings highlight the effectiveness of CGoT in mimicking the diagnostic and clinical decision making process, thereby enhancing professional medical reasoning.

Table 2. Performance comparison of various models on HIE-Reasoning benchmark. Performance comparison on HIE-Reasoning benchmark reveals that directly feeding the MRI images and task descriptions into LVLMs yields poor performance. In contrast, our CGoT demonstrates enhanced reasoning capabilities, significantly aiding medical decision-making.

Model	Lesion Grading		Lesion Anatomy	Lesion in Rare Locations	MRI Injury Score	Neurocognitive Outcome	Interpretation Summary
	Acc (↑)	MAE (↓)	F1 Score (↑)	F1 Score (↑)	Acc (↑)	Inter Class Acc (↑)	ROUGE-L (↑)
GPT4o-Series	33.83%	0.1085	14.81%	11.47%	30.08%	52.16%	37.03%
Gemini-1.5-Flash	24.81%	0.1977	30.17%	22.53%	30.83%	56.60%	41.47%
MiniGPT4-Med-7B	12.03%	0.3785	8.00%	4.62%	34.59%	53.95%	41.09%
LLava-Med-7B	30.83%	0.1468	24.28%	22.52%	11.28%	49.50%	42.05%
Med-Flamingo-7B	18.05%	0.4651	–	–	21.05%	50.00%	36.90%
CGoT-GPT4o-Series	56.25%	0.0715	34.25%	33.04%	**51.13%**	61.11%	44.14%
CGoT-Gemini-1.5-Flash	**62.41%**	**0.0703**	**43.57%**	**41.47%**	49.62%	**71.73%**	**53.68%**

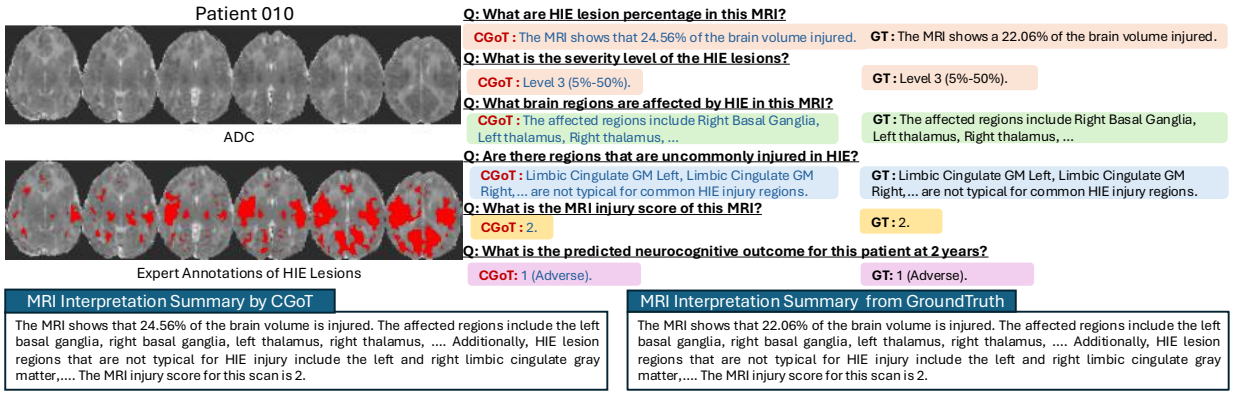


Figure 4. Typical qualitative result of CGoT. CGoT can generate clinically relevant intermediate outputs at each diagnostic step, such as lesion regions and MRI injury scores, which directly inform the final two-year neurocognitive outcome prediction and provide key biomarkers in natural language format.

Table 3. Ablation of Clinical knowledge and graph of thought. It shows that models incorporating both clinical knowledge and graph of reasoning can improve prediction accuracy.

Clinical Knowledge	Graph of Thoughts	Neurocognitive Outcome Prediction
✓	✓	**71.73%**
✓	×	52.43% (↓ 19.30%)
×	✓	51.89% (↓ 19.84%)
×	×	54.44% (↓ 17.29%)

Qualitative Result of CGoT. The graph-of-thoughts reasoning approach in CGoT offers greater transparency and provides interpretable intermediate steps leading to the final two-year outcome prediction. A qualitative example is shown in Fig. 4. CGoT can progressively generate clinically relevant information at each step of the diagnostic process (e.g., *"the affected regions include the right basal ganglia and left thalamus"* and *"MRI injury score is 2"* in Fig. 4). These intermediate outputs are directly relevant to the clinical question (*e.g., What is the two-year neurocognitive outcome for this patient?*) and provide key biomarkers in natural language.

5.3. Ablation Study

In this subsection, we conduct a series of ablation studies to assess the contribution of each component in CGoT and address the following research questions: **Q1**: Does the inclusion of clinical knowledge in CGoT improve performance? **Q2**: Does the clinical graph of reasoning in CGoT enhance answer accuracy? **Q3**: Are all the *reasoning tasks* for neurocognitive outcome prediction in CGoT essential for achieving optimal performance? **Q4**: Is CGoT robust to minor inaccuracies in its intermediate clinical tasks?

As shown in Table 3, models without clinical knowledge as input and without a graph of reasoning perform worse. The combination of both clinical knowledge and the reasoning graph yields the highest prediction accuracy, underscoring the importance of a structured, clinically-informed approach in neurocognitive outcome prediction. The results also demonstrate that the clinical graph of reasoning in CGoT significantly improves answer accuracy. The largest accuracy drop occurs when the reasoning graph is absent, highlighting its critical role in the model's reasoning process

Table 4. Ablation of CGoT with various combinations of reasoning tasks. The results show that including all tasks, particularly the MRI Injury Score, is crucial for optimal performance.

Lesion Grading	Lesion Anatomy	MRI Injury Score	Neurocognitive Outcome Prediction
✓	✓	✓	**71.73%**
✓	✓	×	50.94% (↓ 20.79%)
✓	×	✓	70.12% (↓ 1.61%)
×	✓	✓	61.11% (↓ 10.62%)

(answering **Q1** and **Q2**).

As shown in Table 4, the results indicate that including all reasoning tasks in CGoT is necessary for achieving optimal performance. The model's accuracy drops significantly when any task, especially the MRI Injury Score, is excluded. This was expected, as the MRI Injury Score is a key MRI biomarker for predicting neurocognitive outcomes at two years in many clinical trials worldwide, making it a crucial piece of clinical knowledge for outcome prediction. Furthermore, this suggests that a holistic approach, incorporating Lesion Grading, Lesion Anatomy, and the MRI Injury Score, is essential for high-quality neurocognitive predictions. As a medical AI model, it is clear that each task provides unique and complementary information, and omitting any one of these tasks disrupts the comprehensive reasoning process required for reliable outcome prediction (answering **Q3**).

Table 5. CGoT robustness to perturbations in intermediate predictions.

Perturbation Ratio (%)	Neurocognitive Outcome Prediction (%)
0	71.73
10	67.83
20	66.22
30	62.72

Although CGoT relies on prerequisite reasoning tasks, it remains robust to minor inaccuracies in these intermediate steps. As shown in the second row of Table 4 (without MRI Injury Score: 50.94%; with MRI Injury Score: 71.73%), removing the MRI Injury Score from the reasoning chain results in a substantial drop in outcome prediction accuracy, highlighting the critical role of prerequisite tasks in the model's performance. As demonstrated in Table 5, CGoT maintains robustness when faced with small errors in intermediate clinical reasoning tasks. Specifically, when ± 1-level perturbations are applied to the predicted MRI Injury Score in 10%–30% of test cases, the model exhibits only a gradual decline in performance. This indicates that CGoT is resilient to the types of variability and uncertainty commonly encountered in real-world clinical settings. Such robustness is essential for ensuring reliable predictions under real conditions, further supporting the practical utility of our clinically grounded, stepwise reasoning framework CGoT (answering **Q4**).

6. Conclusion

We present a novel benchmark and framework for advancing Medical VQA, focusing on professional-level medical reasoning for neonatal brain injury. We introduce HIE-Reasoning, the first benchmark to combine clinical visual perception with specialized medical knowledge, specifically for the prediction of neurocognitive outcomes in HIE. We also evaluate state-of-the-art LVLMs, revealing their limitations in handling complex medical data. To address this gap, we propose CGoT, a clinical knowledge-guided model that enhances diagnostic accuracy by mimicking clinical reasoning processes. Our experiments demonstrate a significant performance improvement, underscoring the potential for further innovation in medical AI for professional reasoning.

Impact Statement

This paper introduces a benchmark for evaluating LVLMs in professional-level medical reasoning, with a focus on neonatal Hypoxic-Ischemic Encephalopathy diagnosis and outcome prediction. The broader impact of this work lies in its potential to improve AI-driven clinical decision support, thereby assisting radiologists and neonatologists in diagnosing and prognosticating neonatal brain injury more efficiently and effectively. By structuring complex medical reasoning tasks through clinical knowledge-guided graph-of-thought prompting, our approach enhances interpretability and reliability, which are critical for real-world adoption in healthcare. However, as with any AI-driven medical tool, ethical considerations, such as potential biases in training data, the risk of over-reliance on automated predictions, and the necessity of human oversight, must be carefully addressed. Future research should focus on model generalizability across diverse populations, sites and clinical settings to ensure equitable benefits.

Acknowledgement

This work was funded, in part, by the Harvard Medical School and Boston Children's Hospital. This work was also funded by Thrasher Research Fund Early Career Awards 02402, NIH R21NS121735, R61NS126792, and R03HD104891.

References

Alkhaldi, A., Alnajim, R., Alabdullatef, L., Alyahya, R., Chen, J., Zhu, D., Alsinan, A., and Elhoseiny, M. Minigpt-med: Large language model as a general interface for radiology diagnosis. *arXiv preprint arXiv:2407.04106*,

2024.

Azzopardi, D. V., Strohm, B., Edwards, A. D., Dyet, L., Halliday, H. L., Juszczak, E., Kapellou, O., Levene, M., Marlow, N., Porter, E., et al. Moderate hypothermia to treat perinatal asphyxial encephalopathy. *New England Journal of Medicine*, 361(14):1349–1358, 2009.

Bao, R. and Ou, Y. Boston neonatal brain injury data for hypoxic ischemic encephalopathy (bonbid-hie): Ii. 2-year neurocognitive outcome and nicu outcome. *arXiv preprint arXiv:2411.03456*, 2024.

Bao, R., Grant, E., Kirkpatrick, A., Wachs, J., and Ou, Y. *AI for Brain Lesion Detection and Trauma Video Action Recognition: First BONBID-HIE Lesion Segmentation Challenge and First Trauma Thompson Challenge, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 16 and 12, 2023, Proceedings*, volume 14567. Springer Nature, 2025a.

Bao, R., Song, Y., Bates, S. V., Weiss, R. J., Foster, A. N., Jaimes, C., Sotardi, S., Zhang, Y., Hirschtick, R. L., Grant, P. E., et al. Boston neonatal brain injury data for hypoxic ischemic encephalopathy (bonbid-hie): I. mri and lesion labeling. *Scientific Data*, 12(1):53, 2025b.

Ben Abacha, A., Hasan, S. A., Datla, V. V., Demner-Fushman, D., and Müller, H. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September, 2019.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

Doshi, J., Erus, G., Ou, Y., Resnick, S. M., Gur, R. C., Gur, R. E., Satterthwaite, T. D., Furth, S., Davatzikos, C., Initiative, A. N., et al. Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *Neuroimage*, 127:186–195, 2016.

Douglas-Escobar, M. and Weiss, M. D. Hypoxic-ischemic encephalopathy: a review for the clinician. *JAMA Pediatrics*, 169(4):397–403, 2015.

Edwards, A. D., Brocklehurst, P., Gunn, A. J., Halliday, H., Juszczak, E., Levene, M., Strohm, B., Thoresen, M., Whitelaw, A., and Azzopardi, D. Neurological outcomes at 18 months of age after moderate hypothermia for perinatal hypoxic ischaemic encephalopathy: synthesis and meta-analysis of trial data. *British Medical Journal*, 340, 2010.

Graham, E. M., Ruis, K. A., Hartman, A. L., Northington, F. J., and Fox, H. E. A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy. *American Journal of Obstetrics and Gynecology*, 199(6):587–595, 2008.

He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P. E., and Ou, Y. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets. *arXiv preprint arXiv:2304.09324*, 2023.

He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., and Luo, P. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Laptook, A. R., Shankaran, S., Tyson, J. E., Munoz, B., Bell, E. F., Goldberg, R. N., Parikh, N. A., Ambalavanan, N., Pedroza, C., Pappas, A., et al. Effect of therapeutic hypothermia initiated after 6 hours of age on death or disability among newborns with hypoxic-ischemic encephalopathy: a randomized clinical trial. *JAMA*, 318 (16):1550–1560, 2017a.

Laptook, A. R., Shankaran, S., Tyson, J. E., Munoz, B., Bell, E. F., Goldberg, R. N., Parikh, N. A., Ambalavanan, N., Pedroza, C., Pappas, A., et al. Effect of therapeutic hypothermia initiated after 6 hours of age on death or disability among newborns with hypoxic-ischemic encephalopathy: a randomized clinical trial. *JAMA*, 318 (16):1550–1560, 2017b.

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Lee, A. C., Kozuki, N., Blencowe, H., Vos, T., Bahalim, A., Darmstadt, G. L., Niermeyer, S., Ellis, M., Robertson, N. J., Cousens, S., et al. Intrapartum-related neonatal encephalopathy incidence and impairment at regional and global levels for 2010 with trends from 1990. *Pediatric Research*, 74(1):50–72, 2013.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Liauw, L., van Wezel-Meijler, G., Veen, S., Van Buchem, M., and van der Grond, J. Do apparent diffusion coefficient measurements predict outcome in children with

neonatal hypoxic-ischemic encephalopathy? *American Journal of Neuroradiology*, 30(2):264–270, 2009.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL https://api.semanticscholar.org/CorpusID:964287.

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P., and Rajpurkar, P. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Morton, S. U., Vyas, R., Gagoski, B., Vu, C., Litt, J., Larsen, R. J., Kuchan, M. J., Lasekan, J. B., Sutton, B. P., Grant, P. E., et al. Maternal dietary intake of omega-3 fatty acids correlates positively with regional brain volumes in 1-month-old term infants. *Cerebral Cortex*, 30(4): 2057–2069, 2020.

Murphy, K., van der Aa, N. E., Negro, S., Groenendaal, F., de Vries, L. S., Viergever, M. A., Boylan, G. B., Benders, M. J., and Išgum, I. Automatic quantification of ischemic injury on diffusion-weighted mri of neonatal hypoxic ischemic encephalopathy. *NeuroImage: Clinical*, 14:222–232, 2017.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenAI, R. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.

Ou, Y., Sotiras, A., Paragios, N., and Davatzikos, C. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4):622–639, 2011.

Ou, Y., Zöllei, L., Retzepi, K., Castro, V., Bates, S. V., Pieper, S., Andriole, K. P., Murphy, S. N., Gollub, R. L., and Grant, P. E. Using clinically acquired mri to construct age-specific adc atlases: Quantifying spatiotemporal adc changes from birth to 6-year old. *Human Brain Mapping*, 38(6):3052–3068, 2017.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pinto, A. L., Ou, Y., Sahin, M., and Grant, P. E. Quantitative apparent diffusion coefficient mapping may predict seizure onset in children with sturge-weber syndrome. *Pediatric Neurology*, 84:32–38, 2018.

Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Shankaran, S., Laptook, A. R., Ehrenkranz, R. A., Tyson, J. E., McDonald, S. A., Donovan, E. F., Fanaroff, A. A., Poole, W. K., Wright, L. L., Higgins, R. D., et al. Whole-body hypothermia for neonates with hypoxic–ischemic encephalopathy. *New England Journal of Medicine*, 353 (15):1574–1584, 2005.

Shankaran, S., Barnes, P. D., Hintz, S. R., Laptook, A. R., Zaterka-Baxter, K. M., McDonald, S. A., Ehrenkranz, R. A., Walsh, M. C., Tyson, J. E., Donovan, E. F., et al. Brain injury following trial of hypothermia for neonatal hypoxic–ischaemic encephalopathy. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 97(6):F398–F404, 2012.

Shankaran, S., McDonald, S. A., Laptook, A. R., Hintz, S. R., Barnes, P. D., Das, A., Pappas, A., Higgins, R. D., Ehrenkranz, R. A., Goldberg, R. N., et al. Neonatal magnetic resonance imaging pattern of brain injury as a biomarker of childhood outcomes following a trial of hypothermia for neonatal hypoxic-ischemic encephalopathy. *The Journal of pediatrics*, 167(5):987–993, 2015.

Shankaran, S., Laptook, A. R., Pappas, A., McDonald, S. A., Das, A., Tyson, J. E., Poindexter, B. B., Schibler, K., Bell, E. F., Heyne, R. J., et al. Effect of depth and duration of cooling on death or disability at age 18 months among neonates with hypoxic-ischemic encephalopathy: a randomized clinical trial. *JAMA*, 318(1):57–67, 2017a.

Shankaran, S., Laptook, A. R., Pappas, A., McDonald, S. A., Das, A., Tyson, J. E., Poindexter, B. B., Schibler, K., Bell, E. F., Heyne, R. J., et al. Effect of depth and duration of cooling on death or disability at age 18 months among neonates with hypoxic-ischemic encephalopathy: a randomized clinical trial. *JAMA*, 318(1):57–67, 2017b.

Sotardi, S., Gollub, R. L., Bates, S. V., Weiss, R., Murphy, S. N., Grant, P. E., and Ou, Y. Voxelwise and regional brain apparent diffusion coefficient changes on mri from birth to 6 years of age. *Radiology*, 298(2):415, 2021.

Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., Ni, L. M., Shum, H.-Y., and Guo, J. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, 2024. *URL https://arxiv. org/abs/2307*, 7697, 2023.

Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., et al. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158, 2023.

Team, G. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wei, R., Wang, C., He, F., Hong, L., Zhang, J., Bao, W., Meng, F., and Luo, B. Prediction of poor outcome after hypoxic-ischemic brain injury by diffusion-weighted imaging: A systematic review and meta-analysis. *Plos One*, 14(12):e0226295, 2019.

Weiss, R. J., Bates, S. V., Song, Y., Zhang, Y., Herzberg, E. M., Chen, Y.-C., Gong, M., Chien, I., Zhang, L., Murphy, S. N., et al. Mining multi-site clinical data to develop machine learning mri biomarkers: application to neonatal hypoxic ischemic encephalopathy. *Journal of Translational Medicine*, 17(1):1–16, 2019.

Appendix

In this appendix, we back up our claims by supplementing the main paper with details and examples. We provide our dataset structure in Section A, visual knowledge details in Section B, Clinical Graph of Thought Model (CGoT) details in Section C, model details in Section D and more ablation studies in Section E, respectively.

A. HIE-Reasoning Benchmark Structure

A.1. Benchmark Data Structure

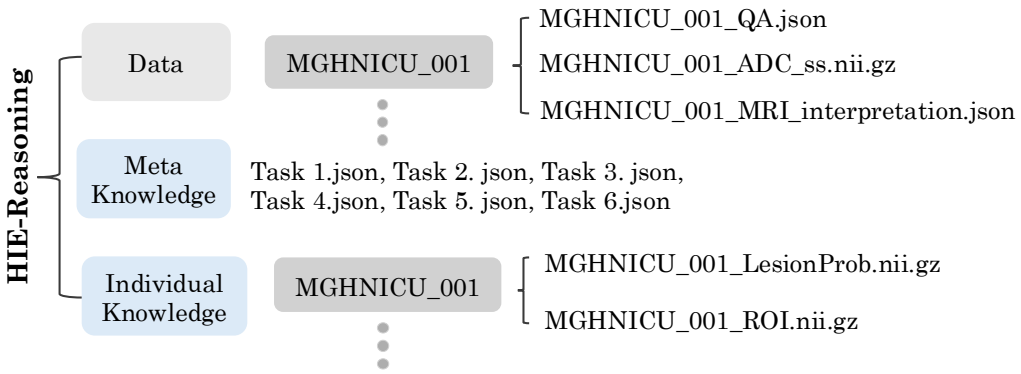


Figure 5. HIE-Reasoning dataset structure. The HIE-Reasoning dataset is organized into three components: Data, Meta Knowledge, and Individual Knowledge.

This work was approved by Institutional Review Boards (IRBs) at Boston Children's Hospital and IRB at Massachusetts General Hospital. Figure 5 illustrates the hierarchical data structure for the HIE-Reasoning framework, which is designed to integrate multimodal data and domain-specific reasoning for Hypoxic-Ischemic Encephalopathy (HIE) research. The structure is organized into three primary components: **Data**, **Meta Knowledge**, and **Individual Knowledge**. Alongside the Data components, the meta knowledge and individual knowledge in this dataset provide expert knowledge to complement this dataset, facilitating further research and fostering advancements in this domain.

The **Data** component comprises patient-level files, such as `MGHNICU_001`, which encapsulate core data assets, including raw MRI files (e.g., `MGHNICU_001_ADC_ss.nii.gz` for apparent diffusion coefficient maps) and task-specific reasoning files. These include our proposed Question-Answering pairs (e.g., `MGHNICU_001_QA.json`) and MRI interpretation (`MGHNICU_001_MRI_interpretation.json`). The **Meta Knowledge** component contains task-specific meta knowledge files (`Task1.json` through `Task6.json`, as shown in Table 7), enabling structured reasoning processes for all six tasks. Lastly, the **Individual Knowledge** component includes patient-specific files like `MGHNICU_001_LesionProb.nii .gz` for probabilistic lesion mapping and `MGHNICU_001_ROI.nii.gz` for brain anatomy labeling. This structured data organization supports comprehensive reasoning workflows, bridging raw data, domain-specific meta knowledge, and task-driven insights to enable automated analysis in HIE MRI interpretation and outcome prediction tasks.

A.2. Benchmark Distribution

Figure 6 illustrates the label distributions for three key aspects: lesion grading, MRI injury scores, and neurocognitive outcomes. Table 6 provides a detailed breakdown of six critical reasoning tasks (e.g., lesion grading, anatomical assessment, and rare location identification) in the HIE-Reasoning benchmark. This benchmark emphasizes a comprehensive evaluation of lesion severity, anatomical analysis, injury scoring, and neurocognitive outcome prediction, advancing automated reasoning for clinical HIE diagnosis and prognosis.

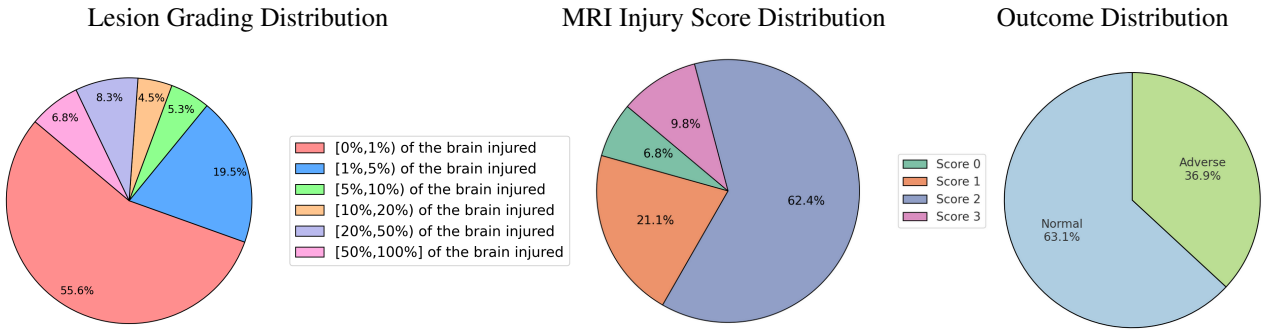


Figure 6. Distribution of benchmark labels.

Table 6. The HIE-Reasoning benchmark includes six critical reasoning tasks related to neonatal HIE in MRI analysis. It consists of six question-answer (QA) pairs focusing on specific reasoning aspects.

Tasks	QA Pairs
Q1: Lesion Grading	133
Q2: Lesion Anatomy	133
Q3: Lesion in Rare Locations	133
Q5: MRI Injury Score	133
Q5: Neurocognitive Outcome Prediction	84
Q6: Interpretation Summary	133
Total MRI: 133	749

B. Visual Knowledge

B.1. Raw MRI: ADC

In the clinical flow of HIE, apparent diffusion coefficient (ADC) maps are pivotal imaging modalities, supplemented by structural MRIs, for detecting HIE-related abnormalities (Douglas-Escobar & Weiss, 2015; Wei et al., 2019; Liauw et al., 2009). Neuroradiologists rely on ADC maps to identify acute brain injuries associated with HIE by locating regions with abnormally low ADC values. These low values indicate restricted water diffusion, a hallmark of ischemic necrosis caused by hypoxic-ischemic insult within the first week after birth (Weiss et al., 2019). In alignment with these clinical insights, HIE-Reasoning includes patient-specific ADC maps as part of the MRI data, as illustrated in Figure 7 (A).

B.2. Derived Lesion Probability

Distinguishing abnormally low ADC values from normal regional variations remains challenging, even for experienced neuroradiologists, as normal ADC values vary across brain regions (Ou et al., 2017; Sotardi et al., 2021). To address this issue, (Bao et al., 2025b) introduced Z_{ADC} maps, which normalize ADC values by quantifying how many standard deviations a patient's ADC value at a given voxel deviates from the mean normal ADC value at that anatomical location(Pinto et al., 2018). These Z_{ADC} maps, residing in the patient's raw ADC image space, enable consistent comparisons across brain regions.

As highlighted by (Bao et al., 2025b), Z_{ADC} values are effective for segmenting HIE-related lesions, with thresholding-based segmentation achieving a Dice score comparable to machine learning algorithms (Murphy et al., 2017). Applying a threshold of -2, the most straightforward value for Z_{ADC} maps, yielded the highest Dice score (0.54 ± 0.28). Accordingly, we used this threshold to generate lesion prediction masks for each patient, as shown in Figure 7 (B) (Z_{ADC}) and Figure 7 (C) (lesion mask from thresholded Z_{ADC}, and lesions are colored by red).

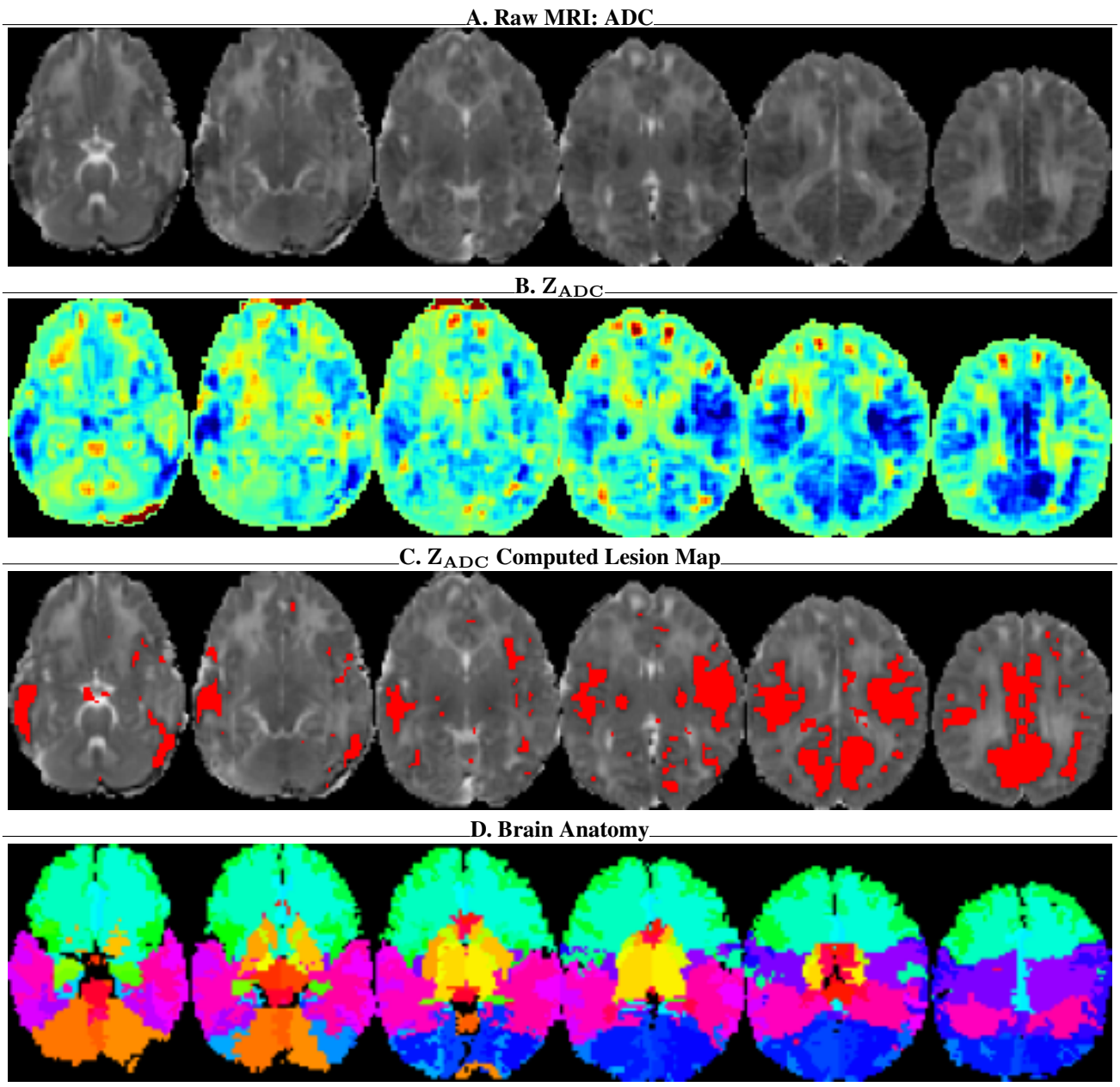


Figure 7. Example of visual knowledge from patient 10. (A) Raw ADC maps from MRI, representing apparent diffusion coefficient values used to identify hypoxic-ischemic injury. (B) Z_{ADC} maps, which normalize ADC values to quantify deviations from regional means in healthy neonatal ADCs, highlighting abnormalities. (C) Lesion prediction map generated by applying a threshold of -2 to the Z_{ADC} map, with lesions visualized in red. (D) Patient-specific brain anatomy map, showcasing ROIs mapped from standard brain anatomical space to individual patient space. Different colors represent different brain ROI. Together, these visual knowledge facilitate tasks in HIE-Reasoning.

B.3. Brain Anatomy

The brain is divided into 62 regions of interest (ROIs) based on standard anatomical references (Doshi et al., 2016; Morton et al., 2020), which are crucial for accurate interpretation of patient-specific MRI interpretation. Figure 7 (D) illustrates the brain ROIs mapped into patient space, generated by transforming the standard brain anatomy ROIs into the patient-specific space using the DRAMMS tool (Ou et al., 2011). Figure 8 illustrates the predefined 62 ROIs of standard brain structure space used in clinical practice. Although the generated ROIs may not achieve 100% accuracy, they represent the most

reliable brain anatomy that can be produced with maximal information and minimal effort.

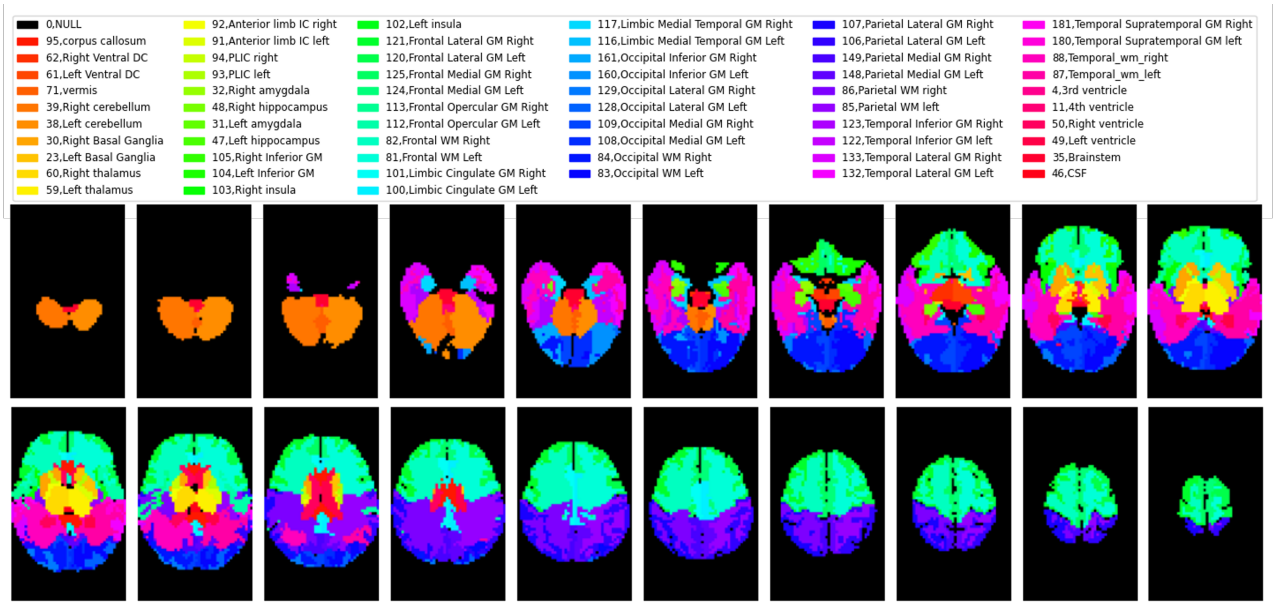


Figure 8. Brain Anatomy. Standard brain structure map is divided into 62 regions of interest (ROIs) based on predefined anatomical references. Each ROI is assigned a unique label and color for visualization, encompassing major anatomical areas such as gray matter (GM), white matter (WM), ventricles, and basal ganglia. These ROIs provide a detailed knowledge for region-specific analysis and lesion localization in HIE-Reasoning tasks.

C. CGoT Knowledge Details

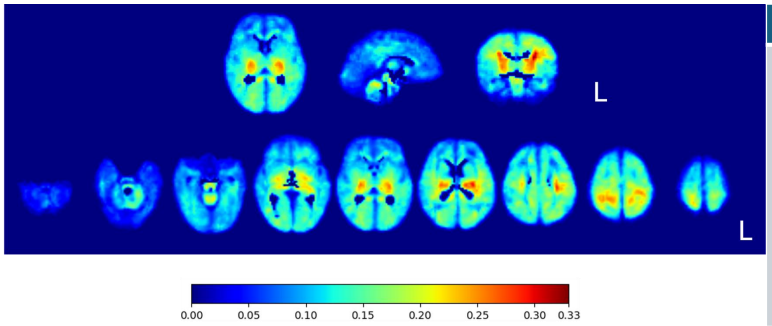
In this section, we provide a detailed explanation of how professional-level clinical knowledge is used to guide the model in correctly addressing each node task, ultimately forming the CGoT pipeline.

Table 7. Clinical knowledge proposed in CGoT for HIE-Reasoning tasks is summarized in the table. *Meta Knowledge* encompasses essential clinical knowledge for HIE clinical reasoning, while *Individual Knowledge* refers to clinical knowledge tailored to each individual patient for each task.

Task	**Meta Knowledge**	**Individual Knowledge**
Lesion Grading	Lesion definition Lesion grading levels	Lesion detection
Lesion Anatomy	Brain ROI regions	Lesion detection Lesion anatomy
Lesion in Rare Location	HIE commonly affected regions	Lesion anatomy
MRI Injury Score	Definition of MRI injury score	Lesion grading Lesion anatomy
Neurocognitive Outcome	Relationship between MRI biomarkers and outcome	Lesion grading Lesion anatomy MRI injury score
MRI Interpretation	MRI interpretation template	Lesion grading Lesion anatomy Lesion in rare locations MRI injury score

Table 7 summarizes the various types of clinical knowledge utilized across the proposed in CGoT in HIE-Reasoning tasks. These include lesion definitions, lesion grading levels, lesion detection, brain ROI regions, and lesion anatomy (as illustrated in the previous section). In the following sections, we further illustrate the brain regions commonly affected by HIE, the definition of the MRI injury score, and the relationship between MRI biomarkers and neurocognitive outcomes.

C.1. HIE Commonly Affected Brain Regions



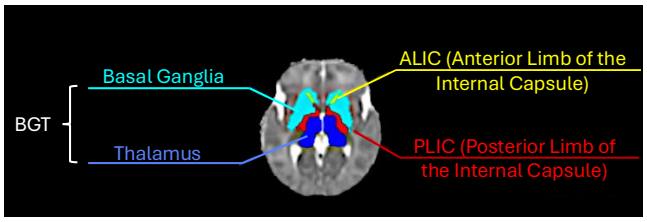
HIE commonly affected brain regions
Left Basal Ganglia, Right Basal Ganglia
Left Thalamus, Right thalamus
Brainstem, Vermis, Corpus Callosum
Frontal WM Left, Frontal WM Right
Parietal WM Left, Parietal WM Right,
Temporal WM Left, Temporal WM Right
Anterior Limb IC Left, Anterior Limb IC Right
PLIC Left, PLIC Right
Parietal Lateral GM Left, Parietal Lateral GM Right
Parietal Medial GM Left, Parietal Medial GM Right

Figure 9. The left panel displays the lesion frequency atlas for HIE, illustrating the statistical distribution of lesions across brain regions based on cohort data. Higher frequencies are represented by warmer colors (yellow to red), indicating areas most commonly affected by HIE-related brain injury. The right panel lists the high-probability brain regions associated with HIE, including basal ganglia, thalamus, brainstem, corpus callosum, anterior limb of the internal capsule (IC), and etc. These regions were identified based on clinical thresholds and expert validation.

To address Task 4, which focuses on lesions in rare locations, it is essential to first identify the regions commonly affected by HIE. According to the study (Bao et al., 2025b), the lesion atlas provided by (Bao et al., 2025b) (left panel) visualizes the distribution of HIE lesions. Based on recommendations from a neonatal radiologist, a threshold of 22.5% was applied to identify high-probability regions of HIE lesions. The resulting list of regions (right panel) was further validated by a national radiologist, confirming these areas as high-probability brain injury regions associated with HIE.

C.2. MRI Injury Score

A. Key Regions in the NICHD-NRN scoring criteria



B. NRN MRI scoring criteria

Score	MRI Finding
0	No injury detected
1A	Minimal cerebral injury without BGT, ALIC, PLIC or WS injury
1B	More extensive cerebral injury without BGT, APLC, PLIC, or WS injury
2A	Any BGT, ALIC, PLIC, or WS injury without other cerebral injury
2B	Any BGT, ALIC, PLIC or WS injury with other cerebral injury
3	Cerebral hemisphere devastation

Figure 10. NRN MRI biomarker system. (A) Key regions in the NICHD-NRN scoring criteria, highlighting the basal ganglia, thalamus (BGT), anterior limb of the internal capsule (ALIC), and posterior limb of the internal capsule (PLIC) as critical areas for assessing HIE-related brain injury. (B) The NRN MRI scoring criteria used for severity grading, ranging from 0 (no injury detected) to 3 (cerebral hemisphere devastation). Intermediate scores (1A, 1B, 2A, 2B) describe varying levels of cerebral and subcortical injuries involving BGT, ALIC, PLIC, or watershed (WS) regions, with increasing severity correlating with adverse neurocognitive outcomes.

To address Task 5 of predicting neurocognitive outcomes, it is crucial to incorporate MRI biomarkers from clinical trials for HIE. The Neonatal Research Network (NRN) score, developed by the National Institute of Child Health and Human Development (NICHD), is the most widely used method in clinical trials to assess the severity of brain injury and predict 2-year neurocognitive outcomes in infants with HIE (Shankaran et al., 2015; 2012). This scoring system relies on neuroradiological experts evaluating the severity of brain injuries based on neonatal brain MRI. As shown in Figure 10, the

NRN MRI score is based on the anatomic locations and extent of HIE lesions visualized on brain MRIs. The score ranges across six levels (0, 1a, 1b, 2a, 2b, and 3), where 0 represents the absence of HIE injury, and 3 indicates the most extensive injury within the brain. In our task, we simplify the scoring by grouping 1a and 1b into a single score of 1 and 2a and 2b into a single score of 2, as recommended by neonatal radiologists. This adjustment accounts for the ambiguity and lack of distinct quantitative boundaries between these subcategories and relies on clinical expert judgment.

C.3. MRI Biomarkers for Neurocognitive Outcome

Predicting 2-year neurocognitive outcomes in HIE remains a significant clinical challenge (Bao & Ou, 2024; Weiss et al., 2019; Bao et al., 2025b; Laptook et al., 2017a; Shankaran et al., 2017a). Medical experts continue to explore more accurate MRI biomarkers to improve outcome prediction (Shankaran et al., 2005; Edwards et al., 2010; Azzopardi et al., 2009; Laptook et al., 2017a; Shankaran et al., 2017a). Current clinical consensus indicates that the higher the MRI injury score, the more severe the brain injury, with the MRI injury score serving as a biomarker strongly associated with adverse 2-year neurocognitive outcomes. A score of 3 is nearly 100% predictive of adverse outcomes. Additionally, injuries to critical regions such as the vermis, cerebellum, brainstem, and hippocampus are linked to adverse neurocognitive outcomes. Patients with lower MRI injury scores or fewer lesion regions are likely to have normal neurocognitive outcomes after 2 years. In this context, the answers provided by Task 1 (lesion grading), Task 2 (lesion anatomy), and Task 4 (MRI injury score) align closely with the consensus knowledge and deliver essential information to aid in neurocognitive outcome prediction.

C.4. Knowledge Prompts

Knowledge Example for Task 1. Lesion Grading

Meta Knowledge:
[Input Description] Suppose you are an expert in detecting Neonatal Brain Injury for Hypoxic Ischemic Encephalopathy, and you are allowed to use any necessary information on the Internet to answer questions. I will give you a set of MRI scanning slices of neonatal brains, these slices are marked with corresponding slice labels, like "Slice 10" and "Slice 11". The label means the slice depth of this slice, for example, "Slice 11" is in the middle layer between "Slice 10" and "Slice 12".

[Lesion Definition] I will give you a pair of images (actually two images) marked with the same "Slice xx" label. The one with the title "Slice x" is the original ADC value of MRI scanning, while the one with the title "ZADC Slice x" is the ZADC value visualization of the gray-scale scan processed by a lesion detection algorithm, where the highly possible abnormal (lesion) region pixel is marked with red (indicating their ZADC values are less than -2). If there is no red pixel, there should be no lesion in this MRI. You should make comprehensive judgments based on both your domain knowledge and the ZADC visualization.

[Lesion Grading Levels] The lesion percentage is defined as the area with ZADC value less than -2 divided by the area with ADC value greater than 0. This will help you better answer the following questions. You need to judge the lesion level of the brain MRI slices by the following rules:
If the lesion region percentage ≤ 0.01, answer with `"level1, <lesion region percentage>"`,
If $0.01 \leq$ lesion region percentage ≤ 0.05, answer with `"level2, <lesion region percentage>"`,
If $0.05 <$ lesion region percentage ≤ 0.5, answer with `"level3, <lesion region percentage>"`,
If $0.5 <$ lesion region percentage ≤ 1.0, answer with `"level4, <lesion region percentage>"`.
The output format should be like `"level4, 0.7344"`.

Individual Knowledge:
[Visual Input] Refer to Figure 7 (A) and (C).

Question & Answer:
[Multiple Choices] What is the severity level of the HIE lesions?
[Open Ending] What is HIE lesion percentage in this MRI?
[Task 1 Answer] level 3, 0.2206.

Knowledge Example for Task 2. Lesion Anatomy

Meta Knowledge:
[Input Description] Suppose you are an expert in detecting Neonatal Brain Injury for Hypoxic Ischemic Encephalopathy, and you are allowed to use any necessary information on the Internet to answer questions. I will now provide you with a series of MRI scanning slices and some pre-processed slices as visual input. The titles of these images include slice sequence labels, such as"Slice 10" and "Slice 11". These labels indicate the depth of the slice; for example, "Slice 11" represents the layer between "Slice 10" and "Slice 12". These images can be grouped into sets of three based on the same slice label. For instance, the images titled "Slice 10", "ZADC Slice 10", and "ROI Slice 10" form one group, indicating that they all correspond to the same scanning depth. These three images, based on their titles, represent the following:
"Slice 10": This is the original MRI scanning ADC value at depth 10, visualized in grayscale.
"ZADC Slice 10": This is the gray-scale visualization of the ZADC values processed by a lesion detection algorithm. In this image, pixels with a high probability of being abnormal (lesions) are marked in red, indicating that their ZADC values are less than -2. If no red pixels are present, it suggests that this MRI scan contains no lesions. It is also possible that the individual has no lesions but has a few areas marked in red.
"ROI Slice 10": This represents different ROI areas of the brain appearing at this scanning depth, with each area highlighted in a different color. The color-to-ROI mapping is provided in the legend on the right side of the image. Note that the ROI regions appearing in slices of different depths are not exactly the same, as only the ROI regions present at a particular depth are displayed. However, for the same cross-slice ROI region, the color used remains consistent across slices.

[Brain ROI List] ID and Region Name Relationship:
95 corpus callosum, 62 Right Ventral DC, 61 Left Ventral DC, 71 vermis, 39 Right cerebellum,
38 Left cerebellum, 30 Right Basal Ganglia, 23 Left Basal Ganglia, 60 Right thalamus, 59 Left thalamus,
92 Anterior limb IC right, 91 Anterior limb IC left, 94 PLIC right, 93 PLIC left, 32 Right amygdala,
48 Right hippocampus, 31 Left amygdala, 47 Left hippocampus, 105 Right Inferior GM, 104 Left Inferior GM,
103 Right insula, 102 Left insula, 121 Frontal Lateral GM Right, 120 Frontal Lateral GM Left,
125 Frontal Medial GM Right, 124 Frontal Medial GM Left, 113 Frontal Opercular GM Right,
112 Frontal Opercular GM Left, 82 Frontal WM Right, 81 Frontal WM Left, 101 Limbic Cingulate GM Right,
100 Limbic Cingulate GM Left, 117 Limbic Medial Temporal GM Right, 116 Limbic Medial Temporal GM Left, 161 Occipital Inferior GM Right, 160 Occipital Inferior GM Left, 129 Occipital Lateral GM Right,
128 Occipital Lateral GM Left, 109 Occipital Medial GM Right, 108 Occipital Medial GM Left,
84 Occipital WM Right, 83 Occipital WM Left, 107 Parietal Lateral GM Right, 106 Parietal Lateral GM Left,
149 Parietal Medial GM Right, 148 Parietal Medial GM Left, 86 Parietal WM right, 85 Parietal WM left,
123 Temporal Inferior GM Right, 122 Temporal Inferior GM left, 133 Temporal Lateral GM Right,
132 Temporal Lateral GM Left, 181 Temporal Supratemporal GM Right, 180 Temporal Supratemporal GM left, 88 Temporal_wm_right, 87 Temporal_wm_left, 4 3rd ventricle, 11 4th ventricle, 50 Right ventricle,
49 Left ventricle, 35 Brainstem, 46 CSF.
You need to choose the names of the ROIs from the above 62 ROI list that contain lesions in this case and output them along with their IDs in the format like: 4_3rd ventricle, 123_Temporal Inferior GM Right, 84_Occipital WM Right, 116_Limbic Medial Temporal GM Left.

Individual Knowledge:
[Visual Input] Refer to Figure 7 (A), (C) and (D).

Question & Answer:
[Question] What brain regions are affected by HIE in this MRI?
[Task 2 Answer] The affected regions include Right Basal Ganglia, Left thalamus, Right thalamus, ...

Knowledge Example for Task 3. Lesion in rare Locations

Meta Knowledge:
[HIE Common Injured Regions] Refer to Figure 9.

Individual Knowledge:
[Task 2 Answer] Your previous answer for the lesion anatomy task is Right Basal Ganglia, Left thalamus, Right thalamus, ...

Question & Answer:
[Open Question] Are there regions that are uncommonly injured in HIE?
[Task 3 Answer] Limbic Cingulate GM Left, Limbic Cingulate GM Right,... are not typical for common HIE injury regions.

Knowledge Example for Task 4. MRI Injury Score

Meta Knowledge:
[Input Description] Refer to [Input Description] in knowledge for Task 2.

[Lesion Definition] Refer to [Lesion Definition] in knowledge for Task 1.

[Lesion Grading Levels] Refer to [Lesion Grading Levels] in knowledge for Task 1.

[Brain ROI List] Refer to [Brain ROI List] in knowledge for Task 2.

[Definition of MRI Injury Score] We have introduced a new diagnostic metric called MRI Injury Score. This metric consists of four levels: Score 0, Score 1, Score 2, and Score 3. Each score level is determined by the injury regions within the ROIs in a given case and the severity of the injury in certain regions.

Score 0: Defined as no injury detected in this case.

Score 1: Defined as either the following a) or b) situation occurs:

a). Minimal cerebral injury without BGT region, ALIC region PLIC region, or detected WS (watershed) injury.

b). More extensive cerebral injury without BGT region, ALIC region PLIC region, or detected WS (watershed) injury.

NOTE: BGT region (including left_BGT and right_BGT), ALIC region (including left_ALIC and right_ALIC), PLIC region (including left_PLIC and right_PLIC).

Score 2: Defined as either the following a) or b) situation occurs:

a). Any BGT region, ALIC region, PLIC region, or WS injury detected without other cerebral injury.

b). Any BGT region, ALIC region, PLIC region, or WS injury detected with other cerebral injury.

NOTE: BGT region (including left_BGT and right_BGT), ALIC region (including left_ALIC and right_ALIC), PLIC region (including left_PLIC and right_PLIC).

Score 3: Defined as cerebral hemisphere devastation.

Individual Knowledge:
[Visual Input] Refer to Figure 7 (A), (C) and (D).
[Task 1 Answer] Your previous answer for the lesion grading task is level 3, 0.2206.
[Task 2 Answer] Your previous answer for the lesion anatomy task is Right Basal Ganglia, Left thalamus, Right thalamus, ...

Question & Answer:
[Question] What is the MRI injury score of this MRI?
[Task 4 Answer] Score 2.

Knowledge Example for Task 5. Neurocognitive Outcome

Meta Knowledge:
[Input Description] Refer to [Input Description] in knowledge for Task 2.
[Lesion Definition] Refer to [Lesion Definition] in knowledge for Task 1.
[Lesion Grading Levels] Refer to [Lesion Grading Levels] in knowledge for Task 1.
[Brain ROI List] Refer to [Brain ROI List] in knowledge for Task 2.
[Definition of MRI Injury Score] Refer to [Definition of MRI Injury Score] in knowledge for Task 4.
[Relationship between MRI Injury Biomarkers and Outcome] The higher the MRI injury score, the more severe the brain injury. The MRI injury score is a biomarker strongly associated with adverse HIE 2-year neurocognitive outcomes. If the score is 3, then towards 100% adverse outcome. Injuries to the vermis, cerebellum, brainstem, and hippocampus are also related to adverse neurocognitive outcomes. If an individual is a current patient, it doesn't necessarily mean he/she will still be a patient in 2 years. Patients with lower MRI injury scores, or fewer lesion regions are likely to become normal in 2 years.

Individual Knowledge:
[Visual Input] Refer to Figure 7 (A), (C) and (D).
[Task 1 Answer] Your previous answer for the lesion grading task is level 3, 0.2206.
[Task 2 Answer] Your previous answer for the lesion anatomy task is Right Basal Ganglia, Left thalamus, Right thalamus, ...
[Task 4 Answer] Your previous answer for the MRI injury score task is Score 2.

Question & Answer:
[Question] What is the predicted neurocognitive outcome for this patient at 2 years?
[Task 5 Answer] 1 (Adverse).

Knowledge Example for Task 6. MRI Interpretation

Meta Knowledge:
[MRI Interpretation Template]
The MRI shows `<Task 1 Answer>` of the brain volume injured.
The affected regions include `<Task 2 Answer>`.
For uncommon areas, `<Task 3 Answer>` are not typical for common HIE injury regions.
The MRI injury score for this MRI is `<Task 4 Answer>`.

Individual Knowledge:
[Task 1 Answer] 22.06%
[Task 2 Answer] Right Basal Ganglia, Left thalamus, Right thalamus, ...
[Task 3 Answer] Limbic Cingulate GM Left, Limbic Cingulate GM Right, ...
[Task 4 Answer] Score 2

Caption:
The MRI shows 22.06% of the brain volume injured.
The affected regions include Right Basal Ganglia, Left thalamus, Right thalamus, ...
For uncommon areas, Limbic Cingulate GM Left, Limbic Cingulate GM Right, ... are not typical for common HIE injury regions.
The MRI injury score for this MRI is Score 2.

D. Model Details

In this section, we provide additional details about the baselines used in the experimental section. The baselines are directly accessed using public APIs or implemented using publicly available source code.

D.1. General-Purpose LVLMs

As detailed in Section C, for the three general-purpose LVLMs, we arranged selected MRI slices into a sequence as visual input, paired with specifically designed prompts, and integrated input from previous steps where applicable to form the input context for the proposed CGoT pipeline. The Gemini-1.5-Flash (Team, 2024) responded effectively to our prompts, whereas GPT4o-Mini (OpenAI, 2024) and GPT4o (OpenAI, 2024) declined to answer certain questions with high uncertainty. GPT4o faild to respond to all five tasks introduced in Figure 2 of the main paper when directly feeding the MRI images and task descriptions. Consequently, we used the quantitative results from GPT4o-Mini to represent the performance of GPT4o-Series models on these tasks. Similarly, after adapting the GPT4o model into CGoT-GPT4o, it still refused to answer lesion grading, lesion anatomy, and lesion-in-rare-locations tasks, but it successfully addressed the MRI injury score and Neurocognitive Outcome tasks based on the answers of previous tasks generated by CGoT-GPT4o-Mini. Thus, we reported CGoT-GPT4o-Mini's evaluation results for the first three tasks and CGoT-GPT4o's performance for the remaining tasks. This fusion approach provided a comprehensive and effective evaluation of the capabilities of the CGoT-GPT4o-Series models.

D.2. Medical-Purpose LVLMs

For the three medical LVLMs—MiniGPT4-Med (Alkhaldi et al., 2024)[1], LLava-Med (Li et al., 2024)[2], and Med-Flamingo (Moor et al., 2023)[3]—we selected implementations directly from the respective papers where these models were implemented or introduced. To ensure a fair comparison, we used the 7B parameter versions of each model, as this configuration provides a balanced trade-off between model performance and computational resource requirements. In line with their instruction-finetuning approach, we combined all selected MRI slices into a single image, annotated with titles indicating their scanning depth, and paired them with task briefings and evaluation questions. However, as demonstrated in the results, these models performed only at or even below the random-guess level, with Med-Flamingo even failing to generate valid outputs for certain tasks.

Since the models are primarily trained on 2D general-purpose or medical images, while brain MRI scanning data inherently represents a 3D structural format where slices at different depths are interrelated, we propose to unroll 3D volumes into sequential 2D slices in order, serving as the visual input for the LVLMs. However, for the three mentioned medical LVLMs, since their fine-tuning or instruction-tuning data mainly consists of single image-text pairs, they fail to provide satisfactory zero-shot responses when queried about 2D-slice sequence input. Therefore, we arrange the series of slices sequentially as subplots within a unified image and annotate subplots with corresponding scanning depth information to ensure input compatibility with these models.

E. More Ablation Studies

Table 8. Ablation study on three types of visual inputs.

			Lesion Grading		**Lesion Anatomy**	**Lesion in Rare Locations**	**MRI Injury Score**	**Neurocognitive Outcome**	**Interpretation Summary**
Raw ADC	ZADC	Brain Anatomy	Acc (↑)	MAE (↓)	F1 Score (↑)	F1 Score (↑)	Acc (↑)	Inter Class Acc (↑)	ROUGE-L (↑)
✓	✓	✓	62.41%	0.0703	43.57%	41.47%	49.62%	71.73%	53.68%
	✓	✓	58.64%	0.0849	42.78%	41.22%	47.37%	68.11%	51.07%
✓		✓	46.62%	0.1152	26.96%	25.10%	30.83%	60.44%	34.55%
✓	✓		62.41%	0.0703	39.90%	37.98%	39.85%	64.05%	47.97%

[1]Code available at: https://github.com/OpenGVLab/Multi-Modality-Arena
[2]Code available at: https://github.com/microsoft/LLaVA-Med
[3]Code available at: https://github.com/snap-stanford/med-flamingo

Table 9. Ablation study on different thresholds of Z_{ADC} .

Z_{ADC}	Lesion Grading		Lesion Anatomy	Lesion in Rare Locations	MRI Injury Score	Neurocognitive Outcome	Interpretation Summary
	Acc (↑)	MAE (↓)	F1 Score (↑)	F1 Score (↑)	Acc (↑)	Inter Class Acc (↑)	ROUGE-L (↑)
$Z_{ADC} < -1.8$	58.64%	0.0782	43.04%	41.42%	47.14%	68.23%	51.42%
$Z_{ADC} < -2.2$	60.90%	0.0867	42.58%	38.19%	37.59%	70.93%	50.67%
$Z_{ADC} < -2.0$	62.41%	0.0703	43.57%	41.47%	49.62%	71.73%	53.68%

E.1. Effectiveness of visual knowledge

Each visual knowledge type (ADC, Z_{ADC}, brain anatomy) has complementary roles. Raw ADC provides signal details, Z_{ADC} identifies abnormal regions, and brain anatomy supply anatomical priors. Removing any component degrades performance, but removing Z_{ADC} has the crucial effect, as it provides the probability of abnormal brain regions—crucial for MRI injury interpretation.

E.2. Robustness to varying Z_{ADC} threshold values

Justification of Z_{ADC} thresholding on –2: The threshold of –2 aligns with clinical understanding and prior studies for HIE abnormal regions probabilities (Weiss et al., 2019; Bao et al., 2025b), indicating ADC values below two standard deviations from the normal atlas—often interpreted as abnormally reduced diffusion in neonatal HIE ADC maps—and serves as a marker for potential brain injury regions.

Across Z_{ADC} threshold variations, CGoT outperforms baselines, demonstrating robust and effective performance. The drop at $Z_{ADC} < -2.2$ in MRI injury stems from an overly strict threshold that misses mild injuries. NRN includes mild cases (0 and 1), which are often excluded by this threshold, leading to missed low-grade injuries. $Z_{ADC} < -2$ better captures these signals, yielding optimal performance.