Title Page
Structural Chain of Thoughts for Radiology Education

Akash Awasthi
PhD Student
Department of Electrical and Computer Engineering
University of Houston
akashcseklu123@gmail.com

Brandon Chung
Undergrad Student
Department of Computer Science
University of Houston
bvchung@CougarNet.UH.EDU

Anh Vu Mai
PhD Student
Department of Electrical and Computer Engineering
University of Houston
mvu9@cougarnet.uh.edu

Saba Khan
PhD Student
Department of Computer Science
University of Houston
skhan225@CougarNet.UH.EDU

Ngan Le, Ph.D.
Assistant Professor
Department of Computer Science & Computer Engineering
University of Arkansas
thile@uark.edu

Zhigang Deng, Ph.D.
Moores Professor of Computer Science
Department of Computer Science
University of Houston
zdeng4@central.uh.edu

Rishi Agrawal, MD
Associate Professor
Department of Thoracic Imaging, Division of Diagnostic Imaging
The University of Texas MD Anderson Cancer Center, Houston, TX
RAgrawal1@mdanderson.org

Carol C. Wu, MD
Professor
Department of Thoracic Imaging, Division of Diagnostic Imaging
The University of Texas MD Anderson Cancer Center, Houston, TX
CCWu1@mdanderson.org

Hien Van Nguyen, Ph.D.
Associate Professor
Department of Electrical and Computer Engineering
University of Houston
hvnguy35@central.uh.edu

Corresponding Author:
Akash Awasthi, PhD Student
Department of Electrical and Computer Engineering
University of Houston
Room No- N368, Cullen College of Engineering Building-1
4222 Martin Luther King Blvd, Houston, TX 77204
akashcseklu123@gmail.com

Structural Chain of thoughts for Radiology Education

Abstract

Radiology education requires trainees to develop both perceptual and interpretive expertise, yet the feedback required to develop these skills remain scarce due to the demanding schedules of experienced radiologists. This lack of personalized guidance makes it difficult for learners to understand not just what errors they made, but also the reason why those errors occurred and how to refine their reasoning skills. Although Large Language Models (LLMs) and Large Multimodal Models (LMMs) have shown promise in radiology applications, they struggle with fine-grained multimodal reasoning. Specifically, these models struggle in detecting subtle cross-modal patterns, such as variations in gaze behavior and diagnostic decisions. These small yet critical differences in how experts and novices allocate visual attention can reveal underlying perceptual gaps, which are often overlooked by current AI-driven approaches. To address these limitations, we introduce Structural Chain of Thoughts (SCoT)—a novel framework that enhances AI sensitivity to nuanced multimodal differences by structuring gaze data and diagnostic reasoning into a thought graph. By leveraging a structural prior, SCoT systematically identifies key perceptual and interpretive discrepancies, allowing models to provide targeted, context-aware feedback. This structured approach not only highlights missed findings but also explains the reasoning behind perceptual errors, turning them into learning opportunities. Applied within radiology education, SCoT bridges the gap between expert and novice performance, offering a scalable solution for AI-driven diagnostic training. We further contribute a simulated dataset of perceptual errors, facilitating future research into multimodal reasoning and educational AI in medical imaging. The code and data will be available here: GitHub Repository

Keywords: Perceptual Error, Large Multimodal Models (LMMs), Large Language Models (LLMs), Zero-Shot (ZS), Few-Shot (FS), Chain of Thought (CoT), Structural Chain of Thoughts (SCoT)

1. Introduction

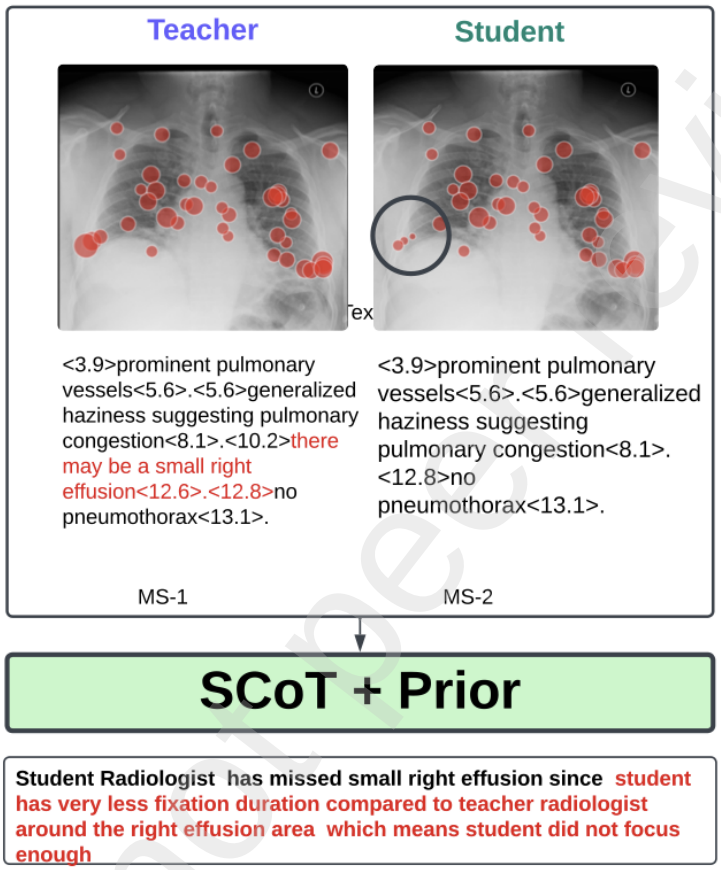

Figure 1: Comparison of multimodal signals (gaze + report transcription) between student and teacher radiologists. Column 1 displays the teacher's gaze and report transcription data, while Column 2 shows the student's data. Despite similar gaze patterns, the student misses a diagnostic finding in the right lung base due to a brief fixation duration. These subtle differences in eye gaze data highlight the challenges the model faces in detecting the reasoning behind the missed diagnosis.

Radiology education is a highly specialized field that requires learners to develop both perceptual and interpretive expertise [1, 2, 3]. However, one of the primary challenges in radiology education is the limited availability of expert feedback due to time constraints [4]. Radiologists often have demanding clinical responsibilities, leaving little time for personalized teaching and mentorship [5, 6]. This creates a gap in the learning process, where students

may struggle to receive targeted guidance on how to improve their diagnostic skills [7, 8].

With advancements in artificial intelligence, particularly Large Language Models (LLMs) and Large Multimodal Models (LMMs), there is growing interest in leveraging these technologies to support radiology education [9, 10, 11]. LLMs and LMMs have already been explored for tasks such as report generation [12, 13], image interpretation [14, 15], and clinical decision support [16, 17]. Recently, researchers have begun investigating their potential in educational settings, including curriculum development, structured report assessment, and new training methodologies [18, 19, 20, 21]. However, a critical gap remains; these models have not been developed to provide personalized feedback on image interpretation or to identify and explain perceptual errors in chest X-ray (CXR) analysis.

Perceptual errors in radiology are closely tied to the radiologist's eye-gaze patterns, which reveals how visual attention is allocated during image interpretation [22, 23]. These errors often occur due to three reasons [22]: (1) a student may fail to fixate on the abnormality at all, meaning they never searched for it—similar to the "satisfaction of search" effect, where once one abnormality is found, further searching is neglected; (2) they may fixate on the abnormal region but the duration is too short, suggesting they were in the area of interest but did not process the abnormality sufficiently; or (3) they may follow a reasonable gaze pattern but still miss the diagnosis due to lack of experience or knowledge. These subtle lapses in attention allocation can lead to diagnostic mistakes, yet no AI-driven solution currently explains why these errors occur. By analyzing multimodal signals—such as eye-tracking data and diagnostic reports—LLMs and LMMs could offer valuable insights into how expert and novice radiologists approach clinical cases, enabling targeted feedback to address perceptual gaps.

However, leveraging LLMs and LMMs for personalized feedback in radiology presents significant challenges, particularly in their limited ability to perform complex multimodal reasoning and compare subtle spatial and temporal variations in multimodal signals [24, 25]. Current prompting strategies often struggle to capture nuanced differences in multimodal data, which limits their effectiveness in providing meaningful feedback. For instance, as illustrated in Figure 1, a student radiologist overlooks a small effusion in the right lung. While their gaze patterns closely resemble those of an expert, subtle discrepancies, such as shorter fixation durations on the abnormal region, can be critical. Detecting such fine-grained variations requires structured

reasoning mechanisms that go beyond simple pattern recognition[26].

The complexity of comparing multimodal signals, such as radiology report transcriptions and gaze patterns of student and expert radiologists, lies in the intricate spatial and temporal interactions between these modalities. Our findings reveal that LLMs and LMMs, under current prompting strategies, often fail to detect these subtle variations. This limitation reduces their sensitivity to perceptual and interpretive gaps, particularly in tasks requiring detailed comparisons, such as assessing how an expert and a novice approach the same diagnostic task. In radiology education, the ability to analyze and compare eye-tracking patterns and diagnostic interpretations between experienced radiologists and trainees could enable more precise, targeted feedback. Such feedback would go beyond broad recommendations, ultimately improving learning outcomes [27]. This highlights the need for advanced reasoning capabilities in LLMs and LMMs to bridge these gaps and enhance their utility in radiology education.

To address these limitations, we propose a novel framework, Structural Chain of Thoughts (SCoT), designed to enhance the sensitivity of LLMs and LMMs in analyzing multimodal signals. SCoT introduces a structured comparison mechanism using a "thought graph" that formalizes interactions between different modalities, such as gaze data and textual reports. This framework enables LLMs and LMMs to systematically identify fine-grained discrepancies in diagnostic reasoning, thus facilitating targeted and personalized feedback for learners.

Our methodology begins by organizing multimodal data—such as eye-tracking patterns and corresponding diagnostic interpretations—into a unified structural representation. By computing structural differences within this representation, we generate a "structural prior" that directs the model's attention to critical variations. This improves the model's ability to provide personalized feedback, allowing precise identification of perceptual and interpretive gaps in a learner's diagnostic process.

Our contributions are as follows:

- **Personalized Perceptual Error Feedback:** We present the Structural Chain of Thoughts (SCoT) framework, tailored for radiology education, to analyze eye-gaze and diagnostic report data from both expert and student radiologists. SCoT provides a structured approach to identify and interpret perceptual errors, offering precise, context-aware feedback. This feedback highlights specific gaps in attention and inter-

pretation, helping students improve their diagnostic skills by focusing on areas where their perceptual processes differ from those of expert radiologists.

- **Methodological Contribution:** We introduce SCoT as a novel methodology for structuring multimodal data in radiology education. By enhancing model sensitivity to subtle differences in complex spatio-temporal signals, SCoT enables detailed comparisons across modalities, such as eye-gaze data and radiological reports. This methodology facilitates nuanced, context-driven reasoning, allowing for the detection of fine-grained variations in diagnostic processes and ultimately improving educational outcomes.

- **Data Contribution:** We release a new simulated dataset containing various perceptual errors in radiology, which supports ongoing research into diagnostic accuracy and the development of advanced error correction systems.

2. Related Work

2.1. Radiology Education

Radiology education requires learners to develop both technical knowledge and perceptual expertise to accurately interpret medical images [1, 2, 3]. Effective training is often hindered by multiple challenges. Senior physicians have limited time for mentorship due to clinical responsibilities, reducing opportunities for direct trainee supervision [28]. The absence of standardized training frameworks across hospitals results in inconsistent skill development [4, 29]. Additionally, variations in faculty teaching proficiency contribute to disparities in residency education quality, making it difficult to ensure a uniform learning experience [4, 29]. The traditional medical education system is often fragmented and inefficient. It relies on uniform, teacher-centered instruction, outdated textbooks, and resource-intensive programs, leading to inconsistencies in educational quality [29, 30]. Studies have also highlighted the lack of expert feedback and personalized training in radiology education [5, 6]. Without tailored guidance, trainees struggle to develop strong diagnostic reasoning skills [7, 8].

Aligned with Deliberate Practice Theory, effective training in radiology hinges on structured practice, clearly defined goals, and consistent feedback

[31, 32]. In particular, visual attention and gaze allocation are crucial to improving diagnostic performance [33, 34]. Recent advancements in Eye Movement Modeling Examples (EMMEs) have highlighted the value of incorporating expert gaze patterns to help learners refine their visual attention [35]. Studies show that leveraging eye-tracking data significantly enhances anatomical identification [33] and aids in decision-making during dynamic visual tasks [34].

Despite these advances, existing computer-aided education systems primarily focus on automated assessments rather than interactive, personalized learning experiences [36, 37]. Prior work often lacks integration with multimodal data, such as eye-gaze patterns and diagnostic reports, which are crucial for understanding reasoning processes [14, 15]. Consequently, there remains a need for intelligent, adaptive systems that provide targeted feedback to bridge the gap between self-directed learning and expert mentorship in radiology training.

2.2. Role of Artificial Intelligence in Radiology Education

Artificial intelligence, particularly LLMs and LMMs, has significantly advanced various applications in radiology. Traditionally, these models have been explored extensively for tasks related to clinical decision support systems, including radiology report generation [12, 13], image interpretation [14, 15], and visual question answering (VQA) [16, 17] to enhance diagnostic accuracy and workflow efficiency. In recent years, however, there has been growing interest in leveraging LLMs and LMMs for radiology education [20, 9, 10, 11]. Studies have explored their role in enriching curriculum development, supporting teaching and learning, and facilitating learner assessment [19, 20]. These models can generate explanatory content, offer automated assessments to improve student engagement, and create medical case studies with associated diagnoses and treatments [21, 19].

Despite these improvements, many existing AI-driven educational tools in radiology do not incorporate personalized feedback mechanisms to identify perceptual errors [22, 23]. This occurs largely due to their limited capacity for complex fine-grained reasoning and their inability to effectively compare multimodal signals with subtle spatial and temporal variations [38, 39]. This gap highlights the need for AI-driven teaching systems that not only assist in knowledge dissemination but also adapt to individual learning patterns, ensuring a more effective and tailored educational experience.

2.3. Chain-of-Thought (CoT) and Multimodal Reasoning

A key prompting and reasoning technique in LLMs, Chain of Thought (CoT), breaks complex problems into smaller steps, improving performance in tasks such as arithmetic, logic, and common sense reasoning [40]. CoT can be applied in Zero-Shot settings, where models reason without prior examples [41], or Few-Shot Scenarios, where a limited number of examples guide the reasoning process [42, 43]. However, standard CoT can be inconsistent, resulting in the development of improved techniques such as Self-Consistency CoT (CoT-SC) [44], which selects the most reliable reasoning path, and self-verification methods [45, 46, 47] that incorporate confidence measures.

Numerous studies have explored optimizing CoT to improve model performance for complex tasks such as in-context learning [48], where models retrieve relevant prompts, and Least-to-Most Prompting [49], which breaks problems into smaller subproblems. Advanced structures further refine CoT. For instance, Tree-of-Thoughts (ToT) [50] organizes reasoning as a tree, enabling multiple solution paths, Graph-of-Thoughts (GoT) [51] maps dependencies between steps, while Skeleton-of-Thought (SoT) [52] processes multiple reasoning streams in parallel. CoT techniques continue to evolve, expanding LLM capabilities through structured problem-solving.

Extending CoT to Multimodal tasks, researchers have adapted its reasoning mechanisms for image- and video-based applications using zero-shot [53], few-shot [54], and self-consistency prompting [44]. Notable approaches include VidIL [55] for video, DDCoT [56] for images, and CCoT [57], which enhances visual reasoning with scene graphs. Additionally, DCoT [58] models object relationships for zero-shot inference, while Video-of-Thought (VoT) [59] structures temporal reasoning for video tasks. Although these CoT extensions have improved LLMs' multimodal reasoning, current prompting strategies still struggle with capturing nuanced differences in multimodal data, particularly when intricate spatial and temporal interactions between modalities demand more structured reasoning. In such cases, more context-aware reasoning mechanisms are required to accurately interpret the nuanced dependencies that span across multiple modalities.

Existing radiology education frameworks face challenges related to inconsistent training, limited mentorship, and the lack of personalized feedback. AI-driven systems, particularly LLMs and LMMs, offer promising solutions but currently fail to provide fine-grained perceptual reasoning feedback. While advances in CoT reasoning have improved model interpretability, no prior work has explored integrating eye gaze data with textual information

to generate personalized prompts for structured decision-making. This paper aims to address these gaps by proposing an AI-driven framework that leverages multimodal CoT reasoning to deliver personalized, adaptive feedback in radiology education, bridging the divide between self-directed learning and expert mentorship.

3. Method

Test-time scaling is a critical technique for adapting LLMs and LMMs to perform complex multimodal reasoning without requiring additional training or fine-tuning. To address the limitations of existing approaches, we introduce the Structural Chain-of-Thought (SCoT) prompting strategy, a structured framework that enables these models to systematically analyze and compare multimodal data, such as gaze patterns and diagnostic interpretations, between students and teachers. SCoT identifies perceptual errors and provides personalized feedback to student radiologists, enhancing their diagnostic reasoning skills. The design of SCoT is grounded in mathematical foundations and structured representations that establish structural priors, ensuring the model effectively captures relationships between fixation patterns and report transcriptions. These priors guide the reasoning process, enabling the model to detect subtle discrepancies in visual attention and diagnostic reasoning. We then present the overall methodology, detailing how SCoT organizes information into a structured reasoning framework.

3.1. Mathematical Foundation

The setup for analyzing the gaze data of teaching and student radiologists is defined as follows:

- $D_T, D_S \in \mathbb{R}^{t \times d}$ represent the gaze data matrices for the teaching and student radiologists, respectively. Here, t is the number of time steps, and d is the dimensionality, encompassing spatial fixation points and fixation durations. Each column in d represents different features of the gaze data, including the x and y coordinates for the fixation point and the fixation duration.

- R_T, R_S denotes the report transcriptions of the teaching and student radiologists encompassing the sentence level timestamps, which provide context to interpret the gaze patterns.

Our goal is to reduce the burden on the LLM/LMM and increase its sensitivity in comparing two multimodal signals so that it can capture the subtle differences between the student and teacher radiologist's eye gaze and report transcriptions. This allows the model to provide useful reasoning to the student radiologist about why they missed a diagnosis or key detail based on their gaze patterns. This involves analyzing both gaze and report data in a way that enhances the model's ability to detect small but significant differences between D_T and D_S and between R_T and R_S. This is achieved by using a graph structure to model the diagnostic thinking process of each radiologist.

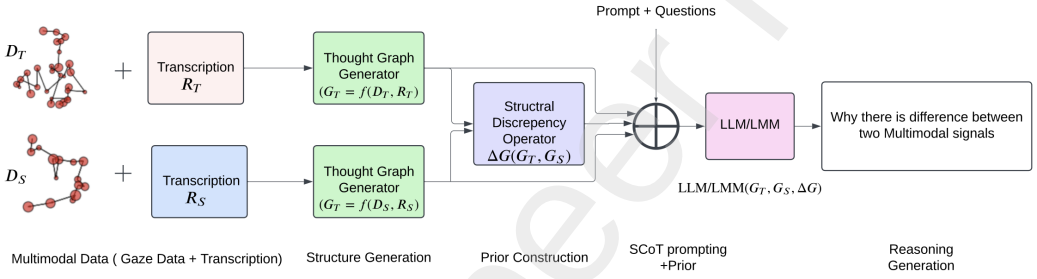

Figure 2: Overview of our proposed methodology, Structural Chain of Thought (SCoT) with Prior, which enhances LLM/LMM sensitivity to subtle differences in complex multimodal signals. The process consists of three phases: (1) Structure Generation: creating structured representations of multimodal data; (2) Prior Construction: generating a structural prior that highlights the most relevant features; and (3) SCoT Prompting with Prior: leveraging the structural prior to guide the model's attention to fine-grained details, thereby improving its ability to reason accurately in complex multimodal tasks.

3.2. SCoT Methodology

As shown in Figure 3, the SCoT methodology involves three main phases for structuring multimodal inputs, calculating the structural discrepancy used as a prior for the LLM/LMM, allowing for a reasoning-driven comparison by LLMs or LMMs between two multimodal signals.

3.2.1. Structure Creation and Attribute Generation

Multimodal Signal Structuring: We define a multimodal representation $G_T = f(D_T, R_T)$ for the teaching radiologist and $G_S = f(D_S, R_S)$ for the student radiologist, where f is a transformation that consolidates gaze and report data into a structured graph form, referred to as the "thought graph".

This graph encodes the diagnostic process, mapping gaze fixation patterns to report interpretations. Each thought in the thought graph corresponds to a sentence in the radiology report separated with the period ("."), representing a particular diagnosis. The representation of thought graphs is illustrated in Figure 3, which visually depicts how gaze fixation patterns and report interpretations are structured into subgraphs.

Graph Structure for Thought Representation: Define a directed graph $G = (V, E)$ to represent each radiologist's thought process, where:

- Each node $v_i \in V$ represents a fixation point, with attributes (p_i, w_i), where p_i denotes the spatial location of the fixation and w_i represents the duration of the fixation.
- Nodes are organized into subgraphs, each linked to a diagnostic phrase in the radiology report. Thus, each subgraph signifies a decision or observation in a specific diagnostic context, reflecting the interplay between gaze data and report interpretation.
- Directed edges $(v_i, v_j) \in E$ within a subgraph represent transitions between fixation points, modeling the temporal sequence in the radiologist's diagnostic process. Each edge (v_i, v_j) is assigned a weight d_{ij} representing the Euclidean distance between fixation points v_i and v_j, indicating how concentrated or dispersed the radiologist's focus was during diagnosis.

3.2.2. Structural Discrepancy Operator for Prior Construction

To enable LLM/LMM to focus on task-critical differences, we compute a structural discrepancy operator on the graphs G_T and G_S. This operator extracts a prior that encodes discrepancies, shifting the model's attention to essential variances without requiring sequential processing of the full multimodal data. Our structural discrepancy operator is defined as follows:

$$\Delta G(G_T, G_S) = (f_{\text{missing\_subgraphs}}(G_T, G_S), \\ f_{\text{missing\_nodes}}(G_T, G_S),\ f_{\text{reduced\_weights}}(G_T, G_S))$$

Where:

- $f_{\text{missing\_subgraphs}}(G_T, G_S)$ identifies the missing subgraphs in G_T that are not isomorphic to any subgraph in G_S,

- $f_{\text{missing\_nodes}}(G_T, G_S)$ identifies the nodes in G_T that are part of missing subgraphs and are not present in G_S,
- $f_{\text{reduced\_weights}}(G_T, G_S)$ identifies the nodes that are common to both graphs but have reduced weights in G_S.

Each function captures a distinct aspect of structural divergence:

Subgraph Matching: Identifying Missing Subgraphs.

$$f_{\text{missing\_subgraphs}}(G_T, G_S) = \{s \subseteq G_T \mid \nexists t \subseteq G_S,\ s \cong t\}$$

Where s is a subgraph of G_T; t is a subgraph of G_S; $\cong$ denotes graph isomorphism (i.e., s and t are structurally identical).

The function $f_{\text{missing\_subgraphs}}(G_T, G_S)$ returns subgraphs of G_T that do not have any isomorphic counterpart in G_S. Two graphs $G_T = (V_T, E_T)$ and $G_S = (V_S, E_S)$ are isomorphic if there exists a one-to-one mapping $f : V_T \to V_S$ such that: $(u, v) \in E_T \iff (f(u), f(v)) \in E_S$ The mapping f preserves the graph's structure.

Node Matching: Identifying Missing Nodes.

$$f_{\text{missing\_nodes}}(G_T, G_S) = \{v \in V_s \mid v \in s, \\ s \in f_{\text{missing\_subgraphs}}(G_T, G_S) \text{ and } v \notin V_t\}$$

Where s is a subgraph of G_T; V_s and V_t are the sets of nodes in G_T and G_S, respectively; $f_{\text{missing\_subgraphs}}(G_T, G_S)$ gives the set of subgraphs of G_T that do not have an isomorphic counterpart in G_S. For each subgraph s in G_T, the function checks if the nodes v in s are present in G_S (i.e., if $v \in V_t$). The nodes v that belong to these subgraphs and are not present in G_S are identified as missing nodes.

Node Weight Comparison: Detecting Reduced Weights.

$$f_{\text{reduced\_weights}}(G_T, G_S) = \{v \in V_T \cap V_S \mid \\ w(v, G_S) < w(v, G_T)\}$$

Where $w(v, G)$ represents the weight (e.g., fixation duration) of the node v in graph G; V_T and V_S are the sets of nodes in G_T and G_S, respectively.

This function identifies the nodes that are common to both graphs G_T and G_S but have reduced weights in G_S compared to G_T. Specifically, it returns the set of nodes with a lower weight in G_S than in G_T.

In this framework, the discrepancy operator ΔG serves as a targeted tool, isolating key structural discrepancies between G_T and G_S, and serving as a prior for the LLM/LMM to streamline its reasoning by focusing on clinically significant differences.

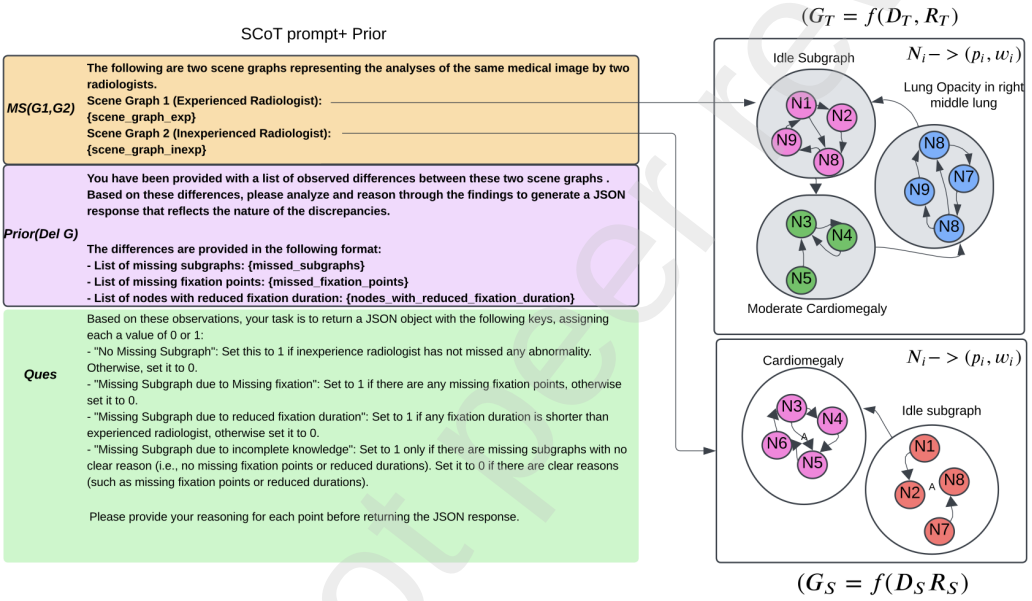

Figure 3: Illustration of the SCoT prompt. This prompt consists of two multimodal structures (thought graphs) paired with a structural prior, guiding the model to identify and reason about the differences between the multimodal data of the student and teacher radiologists. The thought graphs are composed of interconnected subgraphs, each representing a specific thought or diagnosis within the radiology report. This structure enables the model to analyze and compare subtle variations in the data, ultimately helping to uncover the underlying causes of perceptual errors in radiology education.

3.2.3. SCoT Prompting with Prior

As shown in Figure 3, we enable effective multimodal reasoning by providing the LLM/LMM with structured representations G_T, G_S and differential prior ΔG, utilizing chain-of-thought (CoT) prompting to guide the model in identifying subtle but diagnostic differences between student and teacher radiologist's multimodal data.

3.2.4. Prompted Comparison with CoT

To enhance the model's ability to identify and explain differences between the student and teaching radiologists, we employ Chain-of-Thought (CoT) prompting. This guides the LLM/LMM to reason through the structural discrepancy ΔG and provide an interpretable error attribution. We define this process as:

$$E_{\text{error}} = \text{CoT}(G_T, G_S, \Delta G)$$

where E_{error} represents a structured output detailing deviations in diagnostic reasoning. The model focuses on clinically significant discrepancies emphasized by ΔG, ensuring that the feedback is both interpretable and relevant to diagnostic decision-making. A sample prompt used in the SCoT framework is illustrated in Figure 3.

This structured error attribution, E_{error}, enables the student radiologist to pinpoint divergence points in their diagnostic process, linking these variations to differences in gaze patterns and report interpretations. The feedback helps refine their reasoning, highlighting areas for improvement based on direct comparisons with the expert's thought graph.

3.3. Reasoning evaluation

The reasoning process should be validated using structured questions (binary questions) to check whether the model's attribution is accurate. To evaluate the model's output, we simulate data with predefined discrepancies within classes of errors defined in the Dataset section. The model is then tasked with classifying and explaining these simulated errors. To ensure reasoning accuracy, we pose a set of binary questions (0 or 1 responses) in JSON format. Multi-label classification metrics, defined in the Experiments section, are used to quantitatively assess the model's output. These questions serve as evaluative labels, and the model's responses are evaluated using multi-label classification metrics (precision, recall, F1 score) to benchmark its reasoning performance and accuracy.

4. Datasets and Experiments

4.1. Dataset

The EGD-CXR dataset [60] consists of 1,083 chest X-ray (CXR) images paired with synchronized eye-tracking and transcription data, annotated by

an experienced radiologist. The eye-tracking data includes fixation coordinates (x, y), the fixation duration, and the elapsed time (in seconds) from the start of the recording captured using an eye tracker. Additionally, radiological reports were generated by transcribing the radiologist's verbal dictations while analyzing specific regions of the CXR images, focusing on abnormalities. These two multimodal signals—fixation data and textual reports—are used to evaluate the effectiveness of our method in discerning complex interactions among different types of signals.

4.2. Data Processing

The primary goal of this work is to develop a framework for comparing the student and teacher radiology report and eye gaze pattern to provide personalized recommendations to the student radiologist. However, due to the unavailability of student radiologist data with perceptual errors, we simulate our own error dataset. The data processing is carried out in two main steps: Fixation-Transcription Mapping and Error Data Synthesis.

4.2.1. Fixation-Transcription Mapping

In this step, we align sentence-level, timestamped text to the corresponding fixation data of the radiologist on the medical image at specific instances. For each sentence in the radiological report, we extract its start and end timestamps, and then map these to the fixation data using the elapsed time recorded by the eye tracker. The output is a synchronized dataset that pairs spoken analysis of abnormalities with eye gaze fixation data. From the original 1,083 samples in the EGD-CXR dataset, we successfully created 1,025 synchronized samples.

Error Type	**Number of Samples**
Missing Fixation	432
Reduced Fixation Duration	432
Incomplete Knowledge	161
No Error	216
Total Samples	1241

Table 1: Distribution of error types and corresponding sample sizes in the dataset.

4.2.2. Error Data Synthesis

The process of error data synthesis simulates perceptual errors in the EGD-CXR dataset by introducing three types of errors: (i) missed abnormalities due to missing fixation, (ii) missed abnormalities due to reduced fixation, and (iii) missed abnormalities due to incomplete knowledge. To generate these errors, sentences are randomly removed from the accurate transcription to represent missed abnormalities. Based on the error type, the corresponding fixation data is modified: for the "Missing Fixation" error, the fixation data within the relevant time range is removed; for the "Reduced Fixation Duration" error, the fixation duration is reduced by 50%; and for the "Incomplete Knowledge" error, no changes are made to the fixation data.

The dataset includes a mix of error types and samples with no errors. Table 1 provides the distribution of the error types and the corresponding sample sizes. As shown, there are 432 samples each for "Missing Fixation" and "Reduced Fixation Duration," while "Incomplete Knowledge" has 161 samples, and "No Error" includes 216 samples. In total, the dataset contains 1241 samples, providing a diverse set of cases to evaluate how well models can discern different types of perceptual errors.

4.3. Experiments

In our experiments, we evaluate the performance of our proposed SCoT prompting methodology and compare it to two baseline prompting approaches: zero-shot (ZS) and few-shot (FS) CoT. For the ZS CoT baseline, we use a generic CoT prompt to gauge the performance of SOTA CoT prompting on multimodal thought graph comparison. For the few-shot CoT baseline, we provide specific examples that demonstrate how to compare radiologists' eye-gaze patterns and transcriptions, showing step-by-step reasoning. These zero-shot and few-shot prompts are detailed in the supplementary section. We apply these prompting strategies across various LLMs and LMMs with different sizes, including Llama 3.2-Instruct (3B) [61], Llama 3.2 11B-Vision-Instruct [61], Mistral 7B-Instruct-v0.3 [62], and GPT-4 [63], enabling us to examine performance variations across model sizes, from smaller models to GPT-4. To ensure consistency, we set the temperature to 0.2 across all models. We accessed GPT-4 via OpenAI API and other models, such as LLAMA and Mistral, through Together.ai [64].

Method	Evaluation Metrics				
	Accuracy ↑	Precision ↑	Recall ↑	F1 Score ↑	Hamming Loss ↓
Mistral-7B-Instruct-v0.3-ZS CoT	30.33	43.10	44.45	33.59	0.33
Mistral-7B-Instruct-v0.3-FS CoT	33.49	52.36	44.20	44.27	0.27
Mistral-7B-Instruct-v0.3-ZS-SCoT (Our)	76.39	85.00	84.53	83.70	0.08
LLAMA-3.2-11B-Vision-Instruct-ZS CoT	40.78	55.17	50.38	47.19	0.25
LLAMA-3.2-11B-Vision-Instruct-FS CoT	43.31	59.55	55.96	55.08	0.23
LLAMA-3.2-11B-Vision-Instruct-ZS-SCoT (Our)	80.00	88.81	90.52	88.97	0.06
GPT-4o-Mini-ZS CoT	25.46	60.10	81.60	64.84	0.30
GPT-4o-Mini-FS CoT	48.32	56.00	71.07	61.91	0.21
GPT-4o-Mini-ZS-SCoT (Our)	96.48	97.62	96.90	97.24	0.01

Table 2: Performance Comparison of SCoT with Baseline Methods Across Multiple Models. This table evaluates the SCoT framework against standard CoT prompting in zero-shot and few-shot settings on the synthesized error dataset, highlighting its effectiveness in improving multimodal reasoning across different LLM/LMM models.

4.3.1. Evaluation Metrics

Since we frame the task as a multilabel classification problem, we calculate precision, recall, F1 score for each class and hamming loss to evaluate the multilabel classification performance. In Table 2, we report the macro precision, recall, and F1 score to provide a consolidated view of model performance. To compute precision, recall, F1 scores for each class, we utilized the `classification_report` function from scikit-learn. This function allows us to assess the classification performance for each individual class, presenting metrics such as precision, recall, and F1 score. To evaluate the overall performance of the model, we employed the `accuracy_score` function from scikit-learn. It provides the sample-based accuracy, where a sample is considered correct only if all labels for that sample are correctly predicted. This is strict, as even one incorrect label for a sample will make it count as incorrect. In addition to accuracy, we calculated the Hamming loss to measure the fraction of incorrect predictions across all classes and labels. The `hamming_loss` function from scikit-learn allows us to quantify the overall label prediction errors, considering individual label mismatches rather than sample-based accuracy alone.

5. Results & Discussion

5.1. Quantitative Results

Our results demonstrate that our proposed prompting strategy SCoT consistently outperforms conventional CoT prompting across multiple LLM/LMM

models, as shown in Table 2. By incorporating structural priors, SCoT effectively guides model attention toward relevant multimodal features, leading to significant improvements in key classification metrics, including accuracy, precision, recall, and F1 score. Importantly, these performance gains are observed across both zero-shot (ZS) and few-shot (FS) CoT, highlighting the robustness of SCoT in effectevely handling the student and teacher thought graphs with subtle differences. A particularly notable finding is that SCoT outperforms few-shot CoT prompting, even in a zero-shot setting. This suggests that structuring the reasoning pathway provides a more effective strategy than simply increasing the number of demonstrations.

Among the tested models, GPT-4o-Mini-ZS-SCoT achieves the highest overall performance, with an accuracy of 96.48 and an F1 score of 97.24, demonstrating the effectiveness of structured reasoning in guiding model decision-making. Similarly, LLAMA-3.2-11B-Vision-Instruct-ZS-SCoT outperforms both its zero-shot and few-shot CoT counterparts, achieving an accuracy of 80.00 and an F1 score of 88.97. Mistral-7B follows the same trend, with SCoT yielding substantial improvements over standard CoT prompting. These results highlight the ability of our structured approach to capture nuanced differences in thought graphs—differences that conventional reasoning strategies often fail to recognize. Furthermore, we observe that larger models, such as LLAMA-3.2-11B and GPT-4o-Mini, tend to benefit more from SCoT than smaller models like Mistral-7B. This suggests that greater model capacity may enhance the ability to leverage structured priors effectively. However, direct comparisons between models must be interpreted with caution, as differences in post-training mechanisms and architectural refinements can also influence performance. To disentangle the effect of model size from other factors, we conduct a controlled ablation study, which is detailed in the following section.

5.1.1. Class Specific Performance

In this section, we evaluate the performance of our baseline prompting strategies and SCoT across all error classes. Table 3 presents the class-wise performance metrics using GPT-4o-Mini. This comparison highlights how well different prompting methods detect perceptual errors made by student radiologists, providing insights into their effectiveness for multimodal reasoning.

The results indicate that ZS and FS CoT struggle to accurately distinguish different types of gaze-based perceptual errors. In particular, the "In-

Method	GPT-4o-Mini				
	Class Label	Accuracy ↑	Precision ↑	Recall ↑	F1 Score ↑
ZS CoT	Missing Fixation	64.97	54.83	95.83	69.76
	Reduced Fixation	70.14	68.42	54.17	60.47
	Incomplete Knowledge	43.51	18.52	76.40	29.82
	No Error	99.70	98.63	100	99.31
FS CoT	Missing Fixation	65.87	55.19	94.77	69.76
	Reduced Fixation	75.54	69.01	74.58	71.69
	Incomplete Knowledge	76.13	18.60	14.91	16.55
	No Error	95.06	81.20	100	89.63
ZS SCoT	Missing Fixation	98.92	99.53	97.92	98.72
	Reduced Fixation	98.53	100	96.53	98.23
	Incomplete Knowledge	98.14	94.94	93.17	94.04
	No Error	99.12	96.00	100	97.96

Table 3: Class-Specific Performance of SCoT and Baseline Methods Using GPT-4o-Mini. This table compares the SCoT framework with ZS CoT and FS CoT prompting strategies across perceptual error classes, emphasizing its superior ability to detect and classify errors, particularly in the 'Incomplete Knowledge' category.

complete Knowledge" class presents a unique challenge, as it requires reasoning beyond gaze alignment. A student may exhibit similar gaze patterns to an expert but still miss a finding due to a lack of knowledge. This demands long-context reasoning, which baseline prompting strategies fail to handle effectively, as seen in their poor F1 scores for this class. In contrast, our proposed SCoT methodology significantly improves performance across all classes. Notably, it provides a major boost for "Incomplete Knowledge," where prior information helps establish the connection between gaze behavior and underlying expertise gaps. SCoT achieves an F1 score of 94.04, substantially outperforming ZS CoT (29.82) and FS CoT (16.55). Additionally, it achieves near-perfect classification for "Missing Fixation" and "Reduced Fixation," demonstrating its capability to distinguish between different error types. These findings underscore the effectiveness of SCoT in handling complex, long-context reasoning tasks crucial for radiology training. By leveraging structured reasoning with prior knowledge, our method enables more accurate identification of perceptual errors, making it a promising tool for automated feedback in medical education.

5.2. Qualitative Results

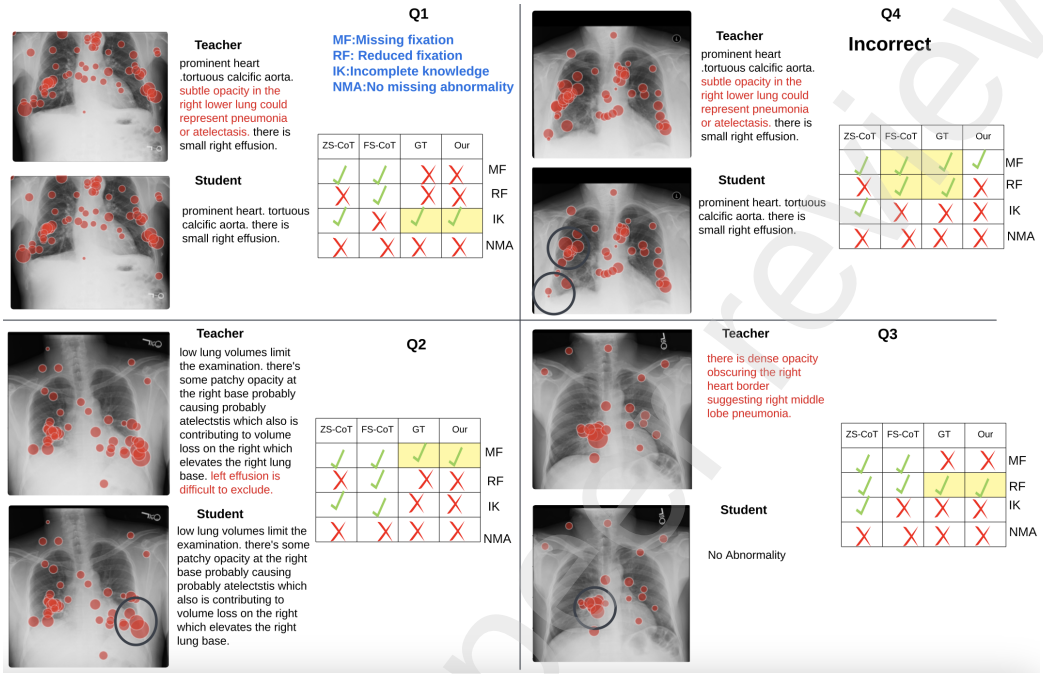

Figure 4: Qualitative comparison of our method (SCoT) with baseline methods (ZS CoT and FS CoT). The figure is divided into four quadrants (Q1, Q2, Q3, Q4), each presenting a table with class-wise predictions for each method. Red-colored text in the radiology report indicates abnormalities missed by the student, while the circles on the student's gaze pattern highlight subtle differences in gaze that contributed to the missed findings. Cells in the table are highlighted for methods that achieved perfect accuracy without any false positives.

To further evaluate the effectiveness of our SCoT prompting strategy, we visualize few examples along with GPT-4o-Mini's predictions compared to baseline methods in Figure 4. The figure is divided into four quadrants (Q1, Q2, Q3, Q4), each illustrating different cases that highlight the strengths and limitations of our approach. In Quadrant 1, SCoT successfully identifies that the student missed an opacity in the right lower lobe due to incomplete knowledge, rather than due to missing or brief fixations. This distinction is crucial, as the student's and teacher's gaze patterns are well-aligned, indicating that the student visually explored the region but lacked the necessary knowledge to recognize the abnormality. Unlike baseline methods, which

struggle with this complex reasoning step, SCoT enables the model to follow longer reasoning chains and correctly infer the underlying cause of the error. Quadrants 2 and 3 further illustrate SCoT's advantage in detecting subtle differences in fixation patterns. For example, in these cases, the student's fixations in key regions—such as the lower lung—are slightly reduced compared to the expert's. While baseline models fail to capture these fine-grained variations, SCoT effectively recognizes these small but meaningful deviations and adjusts its predictions accordingly, demonstrating a heightened sensitivity to nuanced gaze differences. In Quadrant 4, we present a failure case where SCoT misclassifies a brief fixation as a missed fixation. This error suggests that while SCoT significantly enhances multimodal reasoning, some edge cases remain challenging, particularly when distinguishing between brief and entirely absent fixations. These findings highlight areas for further refinement, such as incorporating additional priors or fine-tuning the model's sensitivity to fixation duration thresholds.

5.2.1. Reasoning and Feedback Visualization

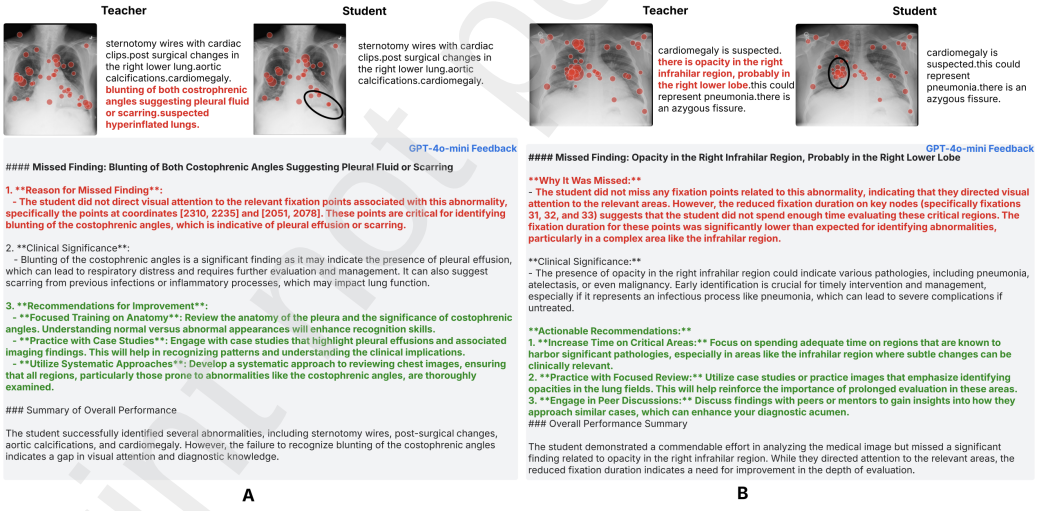

Figure 5: Illustration of the reasoning and feedback generated by GPT-4o-Mini using our SCoT with prior prompting strategy. In both Case A and Case B, the model accurately identifies the type of perceptual error and provides an explanation aligned with the student's gaze behavior. Red text highlights the missed finding and the reasoning behind it, while green text represents actionable recommendations for the student to improve their diagnostic approach.

One of the key applications of our proposed strategy is the development of a personalized teaching assistant for student radiologists. This assistant is designed to identify missed abnormalities, analyze the underlying reasons for these oversights, and provide tailored feedback based on the student's gaze patterns and search strategies. In this section, we present examples of the reasoning and feedback provided by the LLM/LMM using our proposed prompting strategy. The feedback not only helps detect errors, but also explains why these errors occurred, offering personalized insights and recommendations for improvement. Figure 5 illustrates the response produced by GPT-4o-Mini using our prompting approach. In both Case A and Case B, the model not only identifies the nature of the student's error but also provides a meaningful explanation that aligns with the gaze data. This feedback is crucial in guiding students toward a deeper understanding of their perceptual mistakes and refining their search strategies. In Case B, for example, the model correctly identifies that the abnormality was missed due to brief fixation rather than complete omission. Although some fixation points are present in the infra-hilar region, they are brief fixations to allow for accurate abnormality detection. This pattern suggests the occurrence of rapid saccades—quick, unstructured eye movements that prevent focused visual processing. In clinical practice, such scanning behaviors can lead to the oversight of subtle yet clinically significant abnormalities, particularly in complex cases where fine-grained attention is necessary. Beyond just identifying errors and corresponding reasoning, the model also provides actionable recommendations to the student. It suggests increasing fixation time on critical regions, engaging in more case studies, and discussing findings with peers and mentors to reinforce learning.

5.3. Response time comparison

The response times for the three prompting strategies (ZS CoT, FS CoT, and ZS SCoT) show notable differences, as illustrated in Figure 6. The ZS CoT baseline has a mean response time of 5.908 seconds with a variance of 3.271, indicating that while its response time is relatively quick, it fluctuates considerably depending on the complexity of the task. FS CoT, on the other hand, demonstrates a slightly lower mean response time of 5.247 seconds with a variance of 1.052, suggesting that the few-shot approach is more consistent in its performance, but it still lacks the multimodal reasoning capabilities that ZS SCoT provides. The ZS SCoT strategy, with a mean response time of 6.481 seconds and a variance of 3.364, shows a slightly higher process-

ing time but offers a more stable performance across varying tasks, despite the added complexity of multimodal data processing. The slight increase in response time can be attributed to the more complex graph processing involved, which is more demanding than handling simple text. However, the increase in response time is justified by the strategy's improved accuracy, as it is better able to detect and analyze complex perceptual errors compared to the baseline methods. In clinical practice, where precision is paramount, the stability and accuracy provided by ZS SCoT despite the slightly higher response time make it a preferable option for educational feedback in radiology, offering a balance between performance and processing demands.

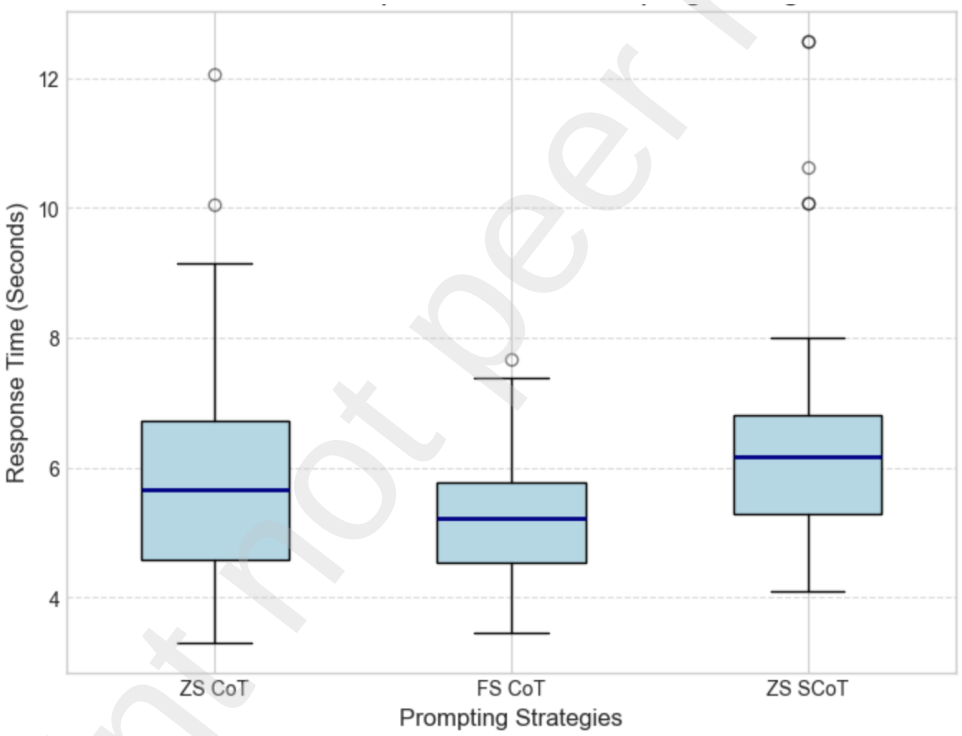

Figure 6: Comparison of response times for GPT-4o-Mini using ZS CoT, FS CoT, and our proposed zero-shot structured chain-of-thought ZS SCoT prompting strategies.

6. Ablations

We perform comprehensive ablation study on the simulated error dataset with the GPT-4o-Mini model. Detailed performance of each ablation experiment is included in the supplementary experiment section.

TG	I	Accuracy ↑	Precision ↑	Recall ↑	F1-Score ↑	Hamming Loss ↓
✓	✗	96.48	97.62	96.90	97.24	0.01
✓	✓	95.43	96.20	95.20	95.70	0.02

Table 4: Ablation Study: Impact of Different Data Components on Performance Metrics. TG represent the Thought graph and I represent the CXR Image.

Ablation with data: In this ablation experiment, we assess the impact of incorporating CXR image data (I) in our proposed SCoT prompting strategy. As shown in Table 5, the inclusion of CXR image (I) did not lead to a significant improvement in performance or enhancement of the model's reasoning process. This finding suggests that the textual and gaze pattern information utilized within the SCoT strategy already provides sufficient context for making accurate decisions, particularly when identifying perceptual errors in radiology. The lack of notable improvement from adding image data underscores a key strength of our model: its ability to function effectively with limited multimodal input, particularly in a zero-shot reasoning context. This characteristic allows our approach to remain lightweight and highly adaptable, making it applicable across a range of domains without relying on specialized image data. The model's ability to focus on textual inputs, gaze patterns, and reasoning strategies enables it to perform well, even in situations where image data may not be easily accessible or practical to use.

Prior: In this ablation study, we investigate how incorporating a structural prior, achieved through a Structural Discrepancy Operator, affects the performance of our SCoT framework. We compare two configurations: (1) using only two Thought Graphs(TG) without the prior(P) and (2) using both the Thought Graphs and the structural prior. As shown in Table 5, combining the Thought Graphs with the structural prior yields the best performance across all metrics, including accuracy, precision, recall, and F1 score. This result highlights the importance of the structural prior in enabling the model to capture fine-grained differences between multimodal signals effectively.

With model size: We evaluated our approach using models of varying sizes to understand the impact of model parameters on performance. Specifically, we tested smaller models, such as LLAMA 3.2 with 3 billion parameters, and larger models, like LLAMA 3.2 11B Vision-Instruct with 11 billion parameters. As the model size and parameter count increase, we observe a

P	TG	Accuracy ↑	Precision ↑	Recall ↑	F1-Score ↑	Hamming Loss ↓
✗	✓	23.57	62.57	79.32	63.94	0.29
✓	✓	96.48	97.62	96.90	97.24	0.01

Table 5: Ablation Study: Impact of structural prior on SCoT Prompting. TG represent the Thought graph and P represent the Priors.

notable improvement in performance. As shown in Figure 7B, the model with 11 billion parameters significantly outperforms the model with 3 billion parameters. This suggests that larger models benefit from enhanced capacity, contributing to better overall results.

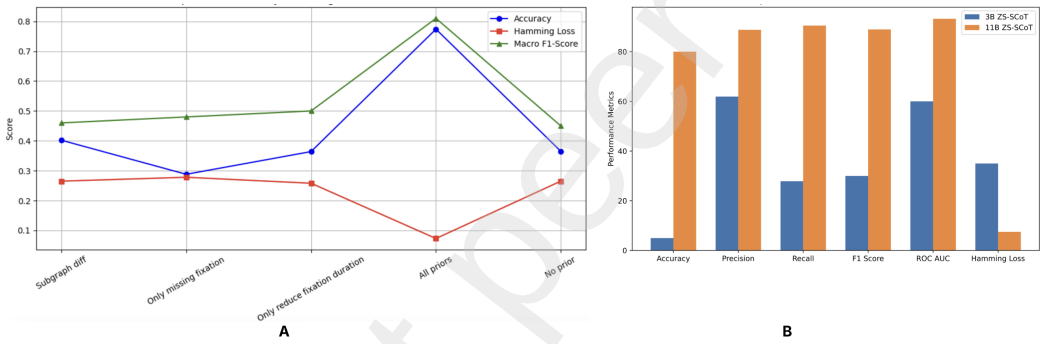

Figure 7: Effect of model size and priors on the performance of SCoT. (A) The impact of various priors on the model's multimodal reasoning ability in radiology, demonstrating that priors such as reduced fixation duration lead to significant performance improvements. (B) A comparison of model performance across different sizes, from the 3-billion parameter LLAMA 3.2 to the 11-billion parameter LLAMA 3.2 11B Vision-Instruct, highlighting the substantial performance gains achieved with larger models.

Effect of Prior: In this ablation experiment, we investigate the marginal effect of different priors in radiology education to understand which components most effectively enhance the multimodal reasoning capabilities of LLMs and LMMs. This study is conducted using GPT-4o-Mini, which demonstrated significant performance improvements when priors were incorporated into the structured prompting approach. As illustrated in Figure 7A, our results reveal that different priors contribute to performance to varying degrees. For instance, priors such as reduced fixation duration notably improve the model's ability to reason multimodally, while other priors show a

relatively smaller effect. Crucially, when all three priors are combined, the model achieves the highest performance, enhancing its capability to differentiate subtle differences in the eye gaze patterns of student and teacher radiologists. This suggests that, while each prior serves a role in improving the model's interpretative abilities, certain priors have a more pronounced influence on guiding the model's reasoning within the radiology education framework. These findings underscore the importance of identifying the most impactful priors for optimizing model performance, particularly in medical AI applications. By understanding the relative contributions of each prior, we can make more targeted improvements to AI-driven educational tools, enhancing their effectiveness in radiology training and potentially in other medical domains where accurate decision-making and nuanced reasoning are critical.

7. Limitations & Future work

While this study primarily focuses on the methodological and theoretical development of a framework designed to enhance the multimodal reasoning capabilities of LLMs and LMMs for radiology education, there are several limitations that need to be acknowledged. One key limitation is the lack of a publicly available dataset specifically capturing student radiologists' errors and gaze patterns. As a result, we simulated an error dataset, which, while useful, may not fully represent the complexities of real-world error scenarios and the nuanced gaze patterns of students during radiological tasks. This work serves as a proof of concept, and further research is required to adapt the framework for direct clinical application, where real-world data would be crucial for evaluating its practical effectiveness. Additionally, our model incorporates a limited number of priors, focusing on structural discrepancy operators. While this approach is effective in the context of our study, it may not account for all the possible types of prior knowledge needed in real-world scenarios. The structural discrepancy operator is, however, quite generalizable, and future research can build upon it by integrating more sophisticated prior models that account for additional complexities within radiology training and other medical fields. Thus, while this framework provides a solid foundation, further refinement and validation are necessary to ensure its broader applicability and effectiveness in clinical and educational settings.

Looking ahead, we plan to conduct a user study to evaluate the real-world utility of our framework by developing a full web-based application.

This application would enable interaction with students in a controlled setting, allowing us to collect valuable feedback on model performance, user experience, and its potential impact on radiology education.

8. Conclusion

In conclusion, this paper introduces the SCoT framework, designed to enhance the multimodal reasoning capabilities of LLMs and LMMs in radiology education. By integrating thought graphs and structural priors, SCoT enables models to perform fine-grained comparisons, identifying subtle discrepancies in diagnostic tasks. Our experiments demonstrate that SCoT significantly improves the model's ability to provide context-sensitive feedback, utilizing gaze patterns and transcriptions to offer personalized insights into the decision-making processes of both novice and expert radiologists. While the framework showed a slight increase in processing time compared to baseline methods, this was justified by its improved accuracy and capability in detecting complex perceptual errors. The results highlight the potential of SCoT to enhance radiology training by providing students with detailed feedback on missed abnormalities, ultimately helping them refine their diagnostic skills. To further advance research, we are releasing a simulated dataset and envision that SCoT will continue to evolve as a valuable tool in radiology education, improving the quality and efficiency of training for future radiologists.

9. Ethics statement

This study did not involve direct data collection from human participants. It utilized publicly available datasets (EGD-CXR, REFLACX) that were collected with appropriate ethical approvals and de-identified to protect participant privacy. Therefore, additional ethical approval was not required for this study.

10. Acknowledgment

This work was supported in part by the National Institutes of Health under Grant 1R01CA277739. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

References

[1] R. B. Gunderman, P. Patel, Perception's crucial role in radiology education, Academic radiology 26 (1) (2019) 141–143.

[2] S. Waite, Z. Farooq, A. Grigorian, C. Sistrom, S. Kolla, A. Mancuso, S. Martinez-Conde, R. G. Alexander, A. Kantor, S. L. Macknik, A review of perceptual expertise in radiology-how it develops, how we can test it, and why humans still matter in the era of artificial intelligence, Academic Radiology 27 (1) (2020) 26–38.

[3] S. Waite, A. Grigorian, R. G. Alexander, S. L. Macknik, M. Carrasco, D. J. Heeger, S. Martinez-Conde, Analysis of perceptual expertise in radiology–current knowledge and a new perspective, Frontiers in human neuroscience 13 (2019) 213.

[4] C. Chew, P. J. O'Dwyer, D. Young, Radiology and the medical student: do increased hours of teaching translate to more radiologists?, BJR—Open 3 (1) (2021) 20210074.

[5] C. M. Straus, E. M. Webb, K. L. Kondo, A. W. Phillips, D. M. Naeger, C. W. Carrico, W. Herring, J. A. Neutze, G. R. Haines, G. D. Dodd, Medical student radiology education: Summary and recommendations from a national survey of medical school and radiology department leadership, Journal of the American College of Radiology 11 (6) (2014) 606–610. doi:https://doi.org/10.1016/j.jacr.2014.01.012.
URL https://www.sciencedirect.com/science/article/pii/S1546144014000192

[6] C. Chew, P. J. O'Dwyer, E. Sandilands, Radiology for medical students: Do we teach enough? a national study, British Journal of Radiology 94 (1119) (2021) 20201308. arXiv:https://academic.oup.com/bjr/article-pdf/94/1119/20201308/57375716/bjr.20201308.pdf, doi:10.1259/bjr.20201308.
URL https://doi.org/10.1259/bjr.20201308

[7] R. B. Gunderman, A. R. Siddiqui, D. E. Heitkamp, H. D. Kipfer, The vital role of radiology in the medical school curriculum, American Journal of Roentgenology 180 (5) (2003) 1239–1242, pMID: 12704030. arXiv:https://doi.org/10.2214/ajr.180.5.1801239, doi:10.2214/ajr.

180.5.1801239.
URL https://doi.org/10.2214/ajr.180.5.1801239

[8] R. Alexander, S. Waite, M. A. Bruno, E. A. Krupinski, L. Berlin, S. Macknik, S. Martinez-Conde, Mandating limits on workload, duty, and speed in radiology, Radiology 304 (2) (2022) 274–282.

[9] C. Amedu, B. Ohene-Botwe, Harnessing the benefits of chatgpt for radiography education: A discussion paper, Radiography 30 (1) (2024) 209–216.

[10] P. Lakhani, B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks, Radiology 284 (2) (2017) 574–582, pMID: 28436741. arXiv:https://doi.org/10.1148/radiol.2017162326, doi:10.1148/radiol.2017162326.
URL https://doi.org/10.1148/radiol.2017162326

[11] R. Bertram, J. Kaakinen, F. Bensch, L. Helle, E. Lantto, P. Niemi, N. Lundbom, Eye movements of radiologists reflect expertise in ct study interpretation: A potential tool to measure resident development, Radiology 281 (3) (2016) 805–815. doi:10.1148/radiol.2016151255.
URL http://dx.doi.org/10.1148/radiol.2016151255

[12] T. Tanida, P. Müller, G. Kaissis, D. Rueckert, Interactive and explainable region-guided radiology report generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7433–7442.

[13] Z. Wang, L. Liu, L. Wang, L. Zhou, Metransformer: Radiology report generation by transformer with multiple learnable expert tokens, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11558–11567.

[14] A. Awasthi, S. Ahmad, B. Le, H. Nguyen, Decoding radiologists' intentions: A novel system for accurate region identification in chest x-ray image analysis, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE, 2024, pp. 1–5.

[15] A. Awasthi, N. Le, Z. Deng, C. C. Wu, H. Van Nguyen, Enhancing radiological diagnosis: A collaborative approach integrating ai and human

expertise for visual miss correction, arXiv preprint arXiv:2406.19686 (2024).

[16] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al., Toward expert-level medical question answering with large language models, Nature Medicine (2025) 1–8.

[17] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (7972) (2023) 172–180.

[18] F. Schuur, M. H. Rezazade Mehrizi, E. Ranschaert, Training opportunities of artificial intelligence (ai) in radiology: a systematic review, European radiology 31 (2021) 6021–6029.

[19] A. Abd-alrazaq, R. AlSaad, D. Alhuwail, A. Ahmed, P. M. Healy, S. Latifi, S. Aziz, R. Damseh, S. Alabed Alrazak, J. Sheikh, Large language models in medical education: Opportunities, challenges, and future directions, JMIR Med Educ 9 (2023) e48291. doi:10.2196/48291.
URL https://mededu.jmir.org/2023/1/e48291

[20] D. H. Ballard, A. Antigua-Made, E. Barre, E. Edney, E. B. Gordon, L. Kelahan, T. Lodhi, J. G. Martin, M. Ozkan, K. Serdynski, et al., Impact of chatgpt and large language models on radiology education: Association of academic radiology—radiology research alliance task force white paper, Academic radiology (2024).

[21] J. Grunhut, O. Marques, A. T. M. Wyatt, Needs, challenges, and applications of artificial intelligence in medical education curriculum, JMIR Med Educ 8 (2) (2022) e35587. doi:10.2196/35587.
URL https://mededu.jmir.org/2022/2/e35587

[22] W. B. Gefter, B. A. Post, H. Hatabu, Commonly missed findings on chest radiographs: causes and consequences, Chest 163 (3) (2023) 650–661.

[23] G. Tourassi, S. Voisin, V. Paquit, E. Krupinski, Investigating the link between radiologists' gaze, diagnostic decision, and image content, Journal of the American Medical Informatics Association 20 (6) (2013) 1067–1075.

[24] M. Małkiński, S. Pawlonka, J. Mańdziuk, Reasoning limitations of multimodal large language models. a case study of bongard problems, arXiv preprint arXiv:2411.01173 (2024).

[25] Y. Hao, J. Gu, H. W. Wang, L. Li, Z. Yang, L. Wang, Y. Cheng, Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, arXiv preprint arXiv:2501.05444 (2025).

[26] A. Van der Gijp, C. Ravesloot, H. Jarodzka, M. Van der Schaaf, I. Van der Schaaf, J. P. van Schaik, T. J. Ten Cate, How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology, Advances in Health Sciences Education 22 (2017) 765–787.

[27] C. Ramirez-Tamayo, S. H. A. Faruqui, S. Martinez, A. Brisco, N. Czarnek, A. Alaeddini, J. R. Mock, E. J. Golob, K. L. Clark, Incorporation of eye tracking and gaze feedback to characterize and improve radiologist search patterns of chest x-rays: A randomized controlled clinical trial, Journal of the American College of Radiology 21 (6) (2024) 942–946.

[28] G. R. Norman, K. W. Eva, Diagnostic error and clinical reasoning, Medical education 44 (1) (2010) 94–100.

[29] Y. Xu, Z. Jiang, D. S. W. Ting, A. W. C. Kow, F. Bello, J. Car, Y.-C. Tham, T. Y. Wong, Medical education and physician training in the era of artificial intelligence, Singapore Medical Journal 65 (3) (2024) 159–166.

[30] D. A. Asch, J. Grischkan, S. Nicholson, The cost, price, and debt of medical education, New England Journal of Medicine 383 (1) (2020) 6–9. arXiv:https://www.nejm.org/doi/pdf/10.1056/NEJMp1916528, doi:10.1056/NEJMp1916528.
URL https://www.nejm.org/doi/full/10.1056/NEJMp1916528

[31] K. A. Ericsson, R. T. Krampe, C. Tesch-Römer, The role of deliberate practice in the acquisition of expert performance., Psychological review 100 (3) (1993) 363.

[32] K. A. Ericsson, Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains, Academic medicine 79 (10) (2004) S70–S81.

[33] A. Gegenfurtner, E. Lehtinen, H. Jarodzka, R. Säljö, Effects of eye movement modeling examples on adaptive expertise in medical image diagnosis, Computers & Education 113 (2017) 212–225.

[34] T. F. Eder, K. Scheiter, J. Richter, C. Keutel, F. Hüttig, I see something you do not: Eye movement modelling examples do not improve anomaly detection in interpreting medical images, Journal of Computer Assisted Learning 38 (2) (2022) 379–391.

[35] D. Darici, M. Masthoff, R. Rischen, M. Schmitz, H. Ohlenburg, M. Missler, Medical imaging training with eye movement modeling examples: A randomized controlled study, Medical Teacher 45 (8) (2023) 918–924.

[36] K. Doi, H. MacMahon, S. Katsuragawa, R. M. Nishikawa, Y. Jiang, Computer-aided diagnosis in radiology: potential and pitfalls, European Journal of Radiology 31 (2) (1999) 97–109. doi:https://doi.org/10.1016/S0720-048X(99)00016-9.
URL https://www.sciencedirect.com/science/article/pii/S0720048X99000169

[37] M. L. Giger, Computer-aided diagnosis in radiology, Academic Radiology 9 (1) (2002) 1–3.

[38] L. Gorospe-Sarasúa, J. Muñoz-Olmedo, F. Sendra-Portero, R. de Luis-García, Challenges of radiology education in the era of artificial intelligence, Radiología (English Edition) 64 (1) (2022) 54–59. doi:https://doi.org/10.1016/j.rxeng.2020.10.012.
URL https://www.sciencedirect.com/science/article/pii/S2173510721000859

[39] J. Schreiter, F. Heinrich, B. Hatscher, D. Schott, C. Hansen, Multimodal human–computer interaction in interventional radiology and surgery: a systematic literature review, International Journal of Computer Assisted Radiology and Surgery (2024) 1–10.

[40] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, F. Xia, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, arXiv preprint arXiv:2201.11903 (2022).

[41] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: Advances in Neural Information Processing Systems, Vol. 35, 2022, pp. 22199–22213.

[42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, arXiv preprint arXiv:2201.11903 (2022).
URL https://arxiv.org/abs/2201.11903

[43] R. Zhang, X. Hu, B. Li, S. Huang, H. Deng, Y. Qiao, P. Gao, H. Li, Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15211–15222.

[44] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. hsin Chi, D. Zhou, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171 (2022).

[45] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, J. Zhao, Large language models are better reasoners with self-verification, arXiv preprint arXiv:2212.09561 (2022).

[46] Z. Ling, Y. Fang, X. Li, Z. Huang, M. Lee, R. Memisevic, H. Su, Deductive verification of chain-of-thought reasoning, Advances in Neural Information Processing Systems 36 (2023) 36407–36433.

[47] N. Miao, Y. W. Teh, T. Rainforth, Selfcheck: Using llms to zero-shot check their own step-by-step reasoning, arXiv preprint arXiv:2308.00436 (2023).

[48] O. Rubin, J. Herzig, J. Berant, Learning to retrieve prompts for in-context learning, arXiv preprint arXiv:2112.08633 (2021).

[49] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, O. Bousquet, Q. Le, E. Chi, Least-to-most prompting enables complex reasoning in large language models, arXiv preprint arXiv:2205.10625 (2022).

[50] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, arXiv preprint arXiv:2305.10601 (2023).

[51] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, et al., Graph of thoughts: Solving elaborate problems with large language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 17682–17690.

[52] X. Ning, Z. Lin, Z. Zhou, Z. Wang, H. Yang, Y. Wang, Skeleton-of-thought: Large language models can do parallel decoding, Proceedings ENLSP-III (2023).

[53] X. Wan, R. Sun, H. Dai, S. Ö. Arik, T. Pfister, Better zero-shot reasoning with self-adaptive prompting, in: Annual Meeting of the Association for Computational Linguistics, 2023.

[54] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work?, arXiv preprint arXiv:2202.12837 (2022).

[55] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem, S.-F. Chang, M. Bansal, H. Ji, Language models with image descriptors are strong few-shot video-language learners, arXiv preprint arXiv:2205.10747 (2022).

[56] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, S. Yang, Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models, arXiv preprint arXiv:2310.16436 (2023).

[57] C. Mitra, B. Huang, T. Darrell, R. Herzig, Compositional chain of thought prompting for large multimodal models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[58] Z. Jia, J. Liu, H. Li, Q. Liu, H. Gao, Dcot: Dual chain-of-thought prompting for large multimodal models, in: The 16th Asian Conference on Machine Learning (Conference Track), 2024.

[59] H. Fei, S. Wu, W. Ji, H. Zhang, M. Zhang, M.-L. Lee, W. Hsu, Video-of-thought: Step-by-step video reasoning from perception to cognition, arXiv preprint arXiv:2501.03230 (2024).

[60] A. Karargyris, S. Kashyap, I. Lourentzou, J. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, M. Moradi, Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development, arXiv preprint arXiv:2009.07386 (2020).

[61] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Y. et al., The llama 3 herd of models (2024). arXiv:2407.21783.
URL https://arxiv.org/abs/2407.21783

[62] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b (2023). arXiv:2310.06825.
URL https://arxiv.org/abs/2310.06825

[63] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. A. et al., Gpt-4 technical report (2024). arXiv:2303.08774.
URL https://arxiv.org/abs/2303.08774

[64] Together ai, accessed: 2024-11-01 (2022).
URL https://www.together.ai/