Large Language Models Are Clinical Reasoners: Reasoning-Aware Diagnosis Framework with Prompt-Generated Rationales

Taeyoon Kwon[1*], Kai Tzu-iunn Ong[1*], Dongjin Kang[2], Seungjun Moon[2], Jeong Ryong Lee[3], Dosik Hwang[3], Beomseok Sohn[4], Yongsik Sim[4], Dongha Lee[1], Jinyoung Yeo[1]

[1]Department of Artificial Intelligence, Yonsei University
[2]Department of Computer Science and Engineering, Yonsei University
[3]Department of Electrical and Electronic Engineering, Yonsei University
[4]Department of Radiology, College of Medicine, Yonsei University

Abstract

Machine reasoning has made great progress in recent years owing to large language models (LLMs). In the clinical domain, however, most NLP-driven projects mainly focus on clinical classification or reading comprehension, and underexplore clinical reasoning for disease diagnosis due to the expensive rationale annotation with clinicians. In this work, we present a "reasoning-aware" diagnosis framework that rationalizes the diagnostic process via prompt-based learning in a time- and labor-efficient manner, and learns to reason over the prompt-generated rationales. Specifically, we address the clinical reasoning for disease diagnosis, where the LLM generates diagnostic rationales providing its insight on presented patient data and the reasoning path towards the diagnosis, namely **Clinical Chain-of-Thought (Clinical CoT)**. We empirically demonstrate LLMs/LMs' ability of clinical reasoning via extensive experiments and analyses on both rationale generation and disease diagnosis in various settings. We further propose a novel set of criteria for evaluating machine-generated rationales' potential for real-world clinical settings, facilitating and benefiting future research in this area.

Introduction

Reasoning is the ability to assess things logically based on available information of various types. Reasoning in clinical diagnosis, also known as clinical reasoning or diagnostic reasoning, is a dynamic thinking process between the observed clinical evidence and the identification of disease. It involves an integration of patient data, relevant medical knowledge, clinicians' experience, and other contextual or situational factors (Norman 2005; Cook, Sherbino, and Durning 2018). Poor clinical reasoning has been directly linked to misdiagnoses and eventually causing hospital adverse events including patient death (Balogh, Miller, and Ball 2015). Therefore, effective clinical reasoning is crucial for diagnosis in real clinical settings (Kassirer 1989).

Recently, deep learning (DL) models are widely utilized for disease diagnosis. However, a predominant portion of existing approaches formulates the process simply as image or text classification (Bakator and Radosav 2018; Kumar et al. 2022). These approaches entirely exclude the aforementioned clinical reasoning in their modeling and focus on fine-tuning high-capacity models for better feature extraction (Jang and Hwang 2022). However, such a data-driven approach can be limited by the data-scarcity problem in biomedical domains. Moreover, high-capacity models are shown to memorize the dataset, rather than solving the diagnosis task through logical reasoning (Mitra et al. 2020), and they cannot provide explanations justifying their diagnoses. Since whether a diagnosis can be explained and whether it matches the reasoning of humans are important for gaining clinicians' trust in DL techniques (Holzinger et al. 2017), this naive approach largely limits models' potential to be implemented for real-world applications.

Meanwhile, large language models have demonstrated their ability to perform multi-step reasoning as well as present the thinking process behind it, which is known as chain-of-thought (CoT) reasoning (Wei et al. 2022). Previous works have applied such reasoning ability to various domains (Wei et al. 2022; Kojima et al. 2022; Wu, Zhang, and Huang 2023; Liévin, Hother, and Winther 2023). In these works, LLMs serve as reasoners that generate natural language rationales guiding and explaining the solution. Despite such success, the use of LLMs to address clinical reasoning in disease diagnosis for real-world applications is still an under-explored area at the moment.

Motivated by these, in this work, we make a step toward clinical reasoning in disease diagnosis, where the models are aware of the clinical reasoning behind the diagnosis, as illustrated in Figure 1. To this end, we formulate the clinical reasoning in disease diagnosis as chain-of-thought reasoning, namely Clinical Chain-of-Thought (Clinical CoT). Our goal is to facilitate clinical reasoning by leveraging LLMs to reason over patient data, refer to relevant knowledge, and generate rationales that guide and explain the diagnosis.

Our contributions are two-fold: (i) We propose a practical framework for reasoning-aware diagnosis. Our framework involves clinical rationalization that augments the existing clinical data with clinical rationales, few-shot reasoning and diagnosis with LLMs, and distillation towards smaller models. (ii) We conduct a thorough analysis of the framework in our testbed diagnosis dataset to gain a deep understanding of the clinical reasoning task. We show that by reason-

*These authors contributed equally.

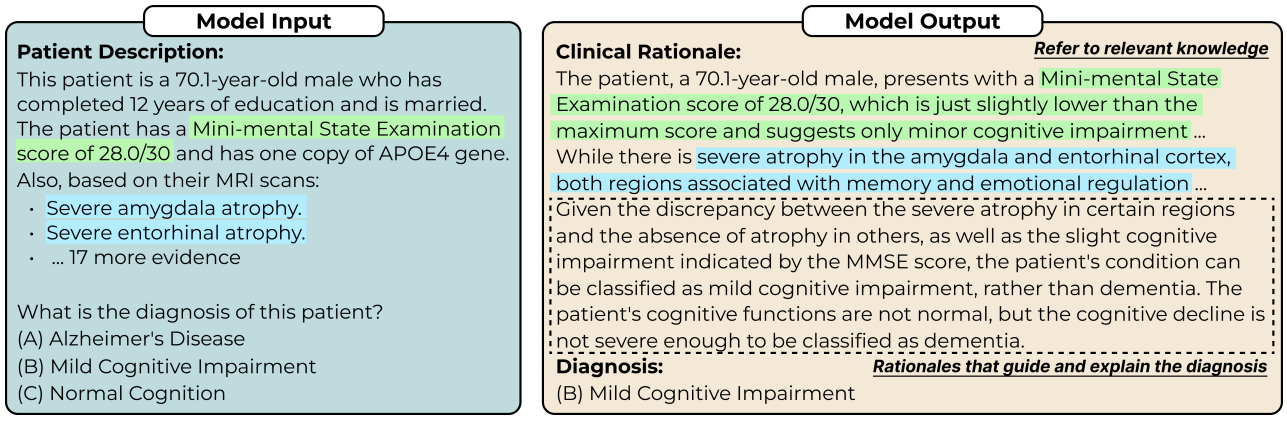


Figure 1: Clinical reasoning in disease diagnosis.

ing over presented clinical data, models can achieve better performance in disease diagnosis. Also, our extensive evaluation and analysis of generated rationales demonstrate that both the LLMs and distilled models can replicate the reasoning of clinical professionals in a human-like manner.

Problem Formulation

Clinical Reasoning for Disease Diagnosis

Most existing approaches for diagnosing diseases with DL models formulate the process simply as image or text classification (Bakator and Radosav 2018; Kumar et al. 2022). That is, given patient description $\mathcal{P}$, such as medical images or electronic health records, a diagnosis model θ is trained to predict the correct diagnosis $\mathcal{D}$:

$$\mathcal{D} \sim P_\theta(\cdot|\mathcal{P}) \tag{1}$$

However, this approach neglects the clinical reasoning connecting the presented patient description and the final diagnosis (Kassirer 1989). The absence of effective clinical reasoning can lead to diagnostic errors (*e.g.*, misdiagnoses), which are reported to contribute to around 10% of patient deaths and hospital adverse events (Norman 2005).

To address that, we exploit LLMs' reasoning capacity in clinical diagnosis, where the LLMs ought to perform clinical reasoning over presented clinical data. Formally, given patient description $\mathcal{P}$, the model first generates a rationale $\mathcal{R}$ decomposing the reasoning process over $\mathcal{P}$, and then makes its diagnosis $\mathcal{D}$ based on $\mathcal{P}$ and $\mathcal{R}$:

$$\mathcal{R} \sim P_\theta(\cdot|\mathcal{P}) \tag{2}$$

$$\mathcal{D} \sim P_\theta(\cdot|\mathcal{P}, \mathcal{R}) \tag{3}$$

Testbed: Alzheimer's Disease Diagnosis

Alzheimer's disease (AD) is an irreversible neurodegenerative disease associated with cognitive decline (DeTure and Dickson 2019). In this study, we choose AD diagnosis task as the testbed for clinical reasoning. This choice is based on the fact that AD diagnosis requires a thorough understanding of various aspects of the disease (Budson and Solomon 2012). In our study, patient description $\mathcal{P}$ consists of (1) textual descriptions derived from the MRI scan, such as *"This patient has **SEVERE** hippocampal atrophy"*;[1] (2) demographic information; (3) educational level; (4) results from the mini-mental state examination (MMSE); (5) the presence of APOE4 allele. The diagnosis $\mathcal{D}$ can be either *Alzheimer's Disease*, *Mild Cognitive Impairment* (MCI), or *Normal Cognition* (NC). Details on transforming MRI scans into textual descriptions are provided in the appendix.

Reasoning-Aware Diagnosis Framework

Framework Overview

Recent works successfully leverage LLMs' ability of CoT reasoning to generate free-text rationales that present the reasoning path and the necessary knowledge towards the answers in various reasoning tasks (Wei et al. 2022; Wu, Zhang, and Huang 2023). Our goal is to exploit such ability in clinical diagnosis, where the LLMs ought to generate rationales demonstrating its reasoning over presented clinical data. For that, we formulate the rationale generation in clinical diagnosis as Clinical CoT reasoning. Upon that, we propose a reasoning-aware diagnosis framework (Figure 2), which includes modules addressing different approaches to facilitate clinical reasoning.

Module I: Clinical Rationalization

To generate clinical CoT rationales, which deliver the diagnostic reasoning towards the correct diagnosis, by prompting a LLM to rationalize the presented clinical data.

Formally, given clinical data consisting of patient description $\mathcal{P}$ and a ground-truth label of the diagnosis $\mathcal{D}$, which can be either *Alzhemer's Disease*, *Mild Cognitive Impairment* (MCI), or *Normal Cognition* (NC), the LLM is prompted to generate clinical rationales $\mathcal{R}^*$ that demonstrate the reasoning process over $\mathcal{P}$ such that the final diagnosis $\mathcal{D}$ can be induced from $\mathcal{R}^*$:

$$\mathcal{R}^* = \underset{\mathcal{R}}{\text{argmax}}\, P_{\text{LLM}}(\mathcal{R}|\mathcal{P}, \mathcal{D}) \tag{4}$$

[1]Hippocampal atrophy refers to the shrinkage or loss of nerve cells in the hippocampus, a region related to memory formation and memory retrieval (Voss et al. 2017).

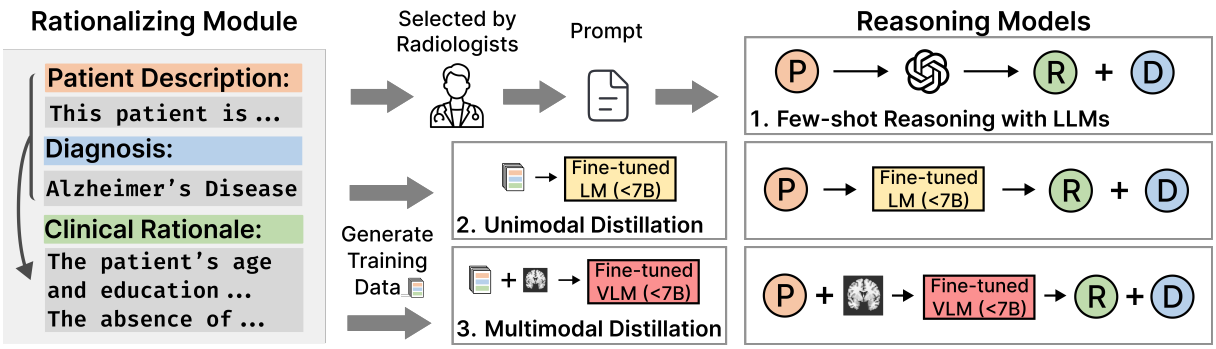


Figure 2: An overview of our framework ($\mathcal{P}$: Patient description; $\mathcal{D}$: Diagnosis; $\mathcal{R}$: Clinical rationale).

As output, we collect a set $\mathbb{D}$ of the processed samples, each of which is a triplet of patient description $\mathcal{P}$, a ground-truth label of the diagnosis $\mathcal{D}$, and the clinical rationales $\mathcal{R}$.

Module II-1: Few-shot CoT Reasoning

LLMs have demonstrated promising performance in tasks requiring logical reasoning with CoT prompting (Kojima et al. 2022; Wei et al. 2022). As a pioneer study towards clinical reasoning with LLMs, we investigate if such success can be replicated in the domain of clinical diagnosis. Thereby, our second module addresses few-shot disease diagnosis, where we prompt LLMs to perform clinical reasoning before the diagnosis.

Formally, given the patient description $\mathcal{P}$, an LLM is prompted to generate both a plausible clinical rationale $\hat{\mathcal{R}}$ and the name of the predicted diagnosis $\hat{\mathcal{D}}$:

$$\hat{\mathcal{R}} = \underset{\mathcal{R}}{\text{argmax}}\, P_{\text{LLM}}(\mathcal{R}|\mathcal{P}) \tag{5}$$

$$\Rightarrow \hat{\mathcal{D}} = \underset{\mathcal{D}}{\text{argmax}}\, P_{\text{LLM}}(\mathcal{D}|\mathcal{P}, \hat{\mathcal{R}}) \tag{6}$$

where $\Rightarrow$ indicates a sequential generation of tokens.

Module II-2: Unimodal-Student Distillation

Despite the impressive performance offered by LLMs in few-shot settings, it is non-trivial to deploy them for real-world applications due to the large size of parameters.[2] Recent works resolve this by using LLMs to augment the target dataset with rationales and use the augmented dataset to fine-tune smaller models, aiming to distill LLMs' reasoning capacity into models that are more affordable for practical uses (Hsieh et al. 2023). This approach is known as knowledge distillation (Hinton, Vinyals, and Dean 2015).

This module distills the knowledge of diagnostic reasoning from the LLM (teacher) into orders-of-magnitude smaller language models, with the goal of developing smaller CoT reasoners for real clinical settings. Applying our rationalizing module (Module I), we obtain clinical data for AD diagnosis that are augmented with clinical rationales. We purpose this augmented dataset as training data to train the student language models.

Formally, given patient description $\mathcal{P}$, the LM is trained to sequentially predict the clinical rationale $\mathcal{R}$ and the ground-truth label of the diagnosis $\mathcal{D}$. The language model is optimized by minimizing the generation loss $\mathcal{L}_{\text{LM-Distill}}$:

$$\mathcal{L}_{\text{LM-Distill}} = \mathop{\mathbb{E}}_{(\mathcal{P},\mathcal{D},\mathcal{R})\in\mathbb{D}}[-\log P_{\text{LM}}(\mathcal{R}, \mathcal{D}|\mathcal{P})] \tag{7}$$

Module II-3: Multimodal-Student Distillation

Besides training smaller CoT reasoners with language models, multimodal CoT, where both visual and textual inputs can be considered via vision-language models (VLMs), has also garnered attention. For instance, Zhang et al. (2023) showed that by including images alongside textual inputs, models with under 1B parameters can generate more effective CoT rationales and vastly outperform the previous state-of-the-art LLM with 175B parameters on a question-answering benchmark. Meanwhile, the diagnosis of many diseases including AD involves medical images such as MRI scans, fundus photographs, and X-ray images (Kumar et al. 2022). Therefore, we further extend knowledge distillation in clinical diagnosis to VLMs.

Formally, given patient description $\mathcal{P}$ and its corresponding MRI scan $\mathcal{V}$, the VLM is trained to sequentially predict the clinical CoT rationale $\mathcal{R}$ and the ground-truth label of the diagnosis $\mathcal{D}$ based on $\mathcal{P}$ and $\mathcal{V}$. The VLM is learnt by minimizing the generation loss $\mathcal{L}_{\text{VLM-Distill}}$:

$$\mathcal{L}_{\text{VLM-Distill}} = \mathop{\mathbb{E}}_{(\mathcal{P},\mathcal{V},\mathcal{D},\mathcal{R})\in\mathbb{D}}[-\log P_{\text{VLM}}(\mathcal{R}, \mathcal{D}|\mathcal{P}, \mathcal{V})] \tag{8}$$

Experiments

Experimental Settings

Dataset. We acquire 7,124 clinical data for Alzheimer's disease (AD) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al. 2008) and 428 from Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) (Ellis et al. 2009). Datasets from these organizations have profoundly facilitated the study of AD. Both datasets comprise three components: (1) MRI scans, (2) ground-truth labels of diagnosis, and (3) patient descriptions, including demographic information, educational level, results from the

[2]One LLM with 175B parameters requires at least 350GB GPU memory with tailored infrastructures (Zheng et al. 2022).

		ADNI							AIBL						
		Accuracy	Precision			Recall			Accuracy	Precision			Recall		
Model	Prompt	Total	AD	MCI	NC	AD	MCI	NC	Total	AD	MCI	NC	AD	MCI	NC
ChatGPT	0-shot	55.3	56.8	45.3	69.9	96.4	22.4	48.8	54.0	59.8	47.2	61.1	80.0	31.7	55.0
	1-shot	50.7	61.4	40.6	71.4	81.5	**62.9**	7.9	50.7	67.4	41.8	**70.6**	66.9	**74.7**	8.6
	3-shot	58.8	61.6	47.1	**71.5**	85.5	46.7	44.8	57.2	**68.3**	46.6	63.5	63.1	57.0	52.1
	5-shot	57.3	55.7	44.4	70.2	**97.2**	23.2	53.2	56.3	57.7	**47.3**	62.5	**86.9**	33.5	53.6
	Clinical CoT	**67.3**	**62.2**	**54.2**	62.7	90.3	26.6	**86.5**	**62.4**	63.8	43.7	53.2	79.2	25.3	**88.6**
GPT-4	0-shot	59.6	51.1	51.6	76.8	**99.6**	24.3	42.1	55.4	52.1	49.5	**71.1**	93.9	29.1	49.3
	1-shot	53.0	54.1	44.4	**81.0**	98.4	42.9	18.7	54.9	55.6	42.2	65.5	**96.2**	38.0	35.7
	3-shot	61.8	64.3	50.9	76.5	86.3	55.2	44.4	58.2	66.2	46.7	67.0	72.3	53.2	50.7
	5-shot	62.6	67.8	50.5	76.5	90.8	**57.5**	40.1	59.8	68.9	48.4	69.3	78.5	**58.9**	43.6
	Clinical CoT	**68.4**	**77.5**	**59.3**	67.4	76.2	40.5	**89.3**	**62.6**	**82.2**	**51.2**	60.9	63.8	39.2	**87.9**

Table 1: Evaluation on LLMs in zero-and-few-shot diagnosis. The Clinical CoT includes two exemplar shots.

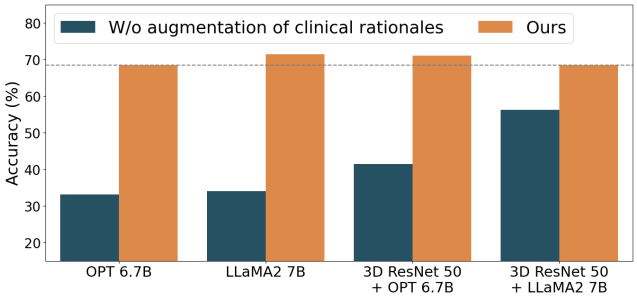


Figure 3: Performance of student models trained with and without clinical rationales, reported on ADNI. The dotted line is the performance of the teacher LLM (GPT-4).

mini-mental state examination (MMSE) and the presence of APRO4 allele. Data from ADNI are split into training, validation, and test sets. All the AIBL data are exclusively used as an out-of-domain test set rather than training. The ratio of AD:MCI:NC in both test sets is roughly 1:1:1.

Large language models. We choose ChatGPT (OpenAI 2023a) and GPT-4 (OpenAI 2023b) as our selection for LLMs. They have shown an impressive ability to perform chain-of-thought reasoning. For our clinical rationalization module (Module I), we adopt GPT-4. And we adopt both ChatGPT and GPT-4 for few-shot diagnosis (Module II). In all of our experiments, we set the temperature to 0.7, max tokens to 2000, and apply greedy decoding.

Unimodal-student models. We consider OPT (Zhang et al. 2022) and LLaMA2 (Touvron et al. 2023), two commonly used language models as our foundation model for the unimodal student. For our experiments, we use the 1.3B and 6.7B versions of OPT and the 7B version of LLaMA2.

Multimodal-student models. Following Tsimpoukelli et al. (2021), our multimodal student is based on the vision-language model, which consists of convolutional neural networks as vision encoder and a language model as text encoder. The vision encoder extracts image features from the MRI scan. During training, the vision encoder is trained to align its extracted feature with text features, such that the language model can effectively attend to features of both modalities. Since MRI scans are 3-dimensioanl images, we consider 3D ResNet (Hara, Kataoka, and Satoh 2017) as our foundation model of vision encoder. For the language model, we use the 1.3B and 6.7B versions of OPT and the 7B version of LLaMA2. Implementation details of our student models are provided in the appendix.

Results and Discussion

We now present the empirical findings of the following research questions that guide our experiments:
RQ1: *Does clinical rationales improve AD diagnosis?*
RQ2: *Does knowledge distillation benefit small models?*
RQ3: *Is our framework helpful in data-scarce scenarios?*
RQ4: *What causes misdiagnosis and how to get over it?*

LLMs' diagnostic performance (RQ1). Table 1 presents the experimental results. Firstly, we observe that LLMs without clinical CoT generally yield high recall in one class (mostly AD) with fairly low recall in the other two classes. This suggests that clinical CoT can prevent LLMs from being biased toward a specific diagnosis.

Secondly, LLMs with clinical CoT show huge improvements in accuracy compared to baselines with more shots. This follows the observation in Kojima et al. (2022), where LLMs with 2-shot CoT prompting largely outperform LLMs with 8-shot standard prompting in arithmetic domains. We demonstrate the same patterns in clinical diagnosis.

Performance of student models (RQ2). Table 2 presents the experimental results of student models. Firstly, among unimodal students: For in-domain data, OPT 6.7B shows comparable performance to the teacher model in total accuracy. LLaMA2 7B yields a significant performance gain in accuracy and precision in all three classes; For out-of-domain data, all unimodal students outperform the teacher LLM in accuracy. In addition, OPT 1.3B and LLaMA2 7B demonstrate better precision for all classes and higher recall for AD and MCI. Moreover, all the text-only unimodal

	ADNI (In-domain)							**AIBL** (Out-of-domain)						
	Accuracy	Precision			Recall			Accuracy	Precision			Recall		
Baselines	Total	AD	MCI	NC	AD	MCI	NC	Total	AD	MCI	NC	AD	MCI	NC
GPT-4 (Teacher Model)	68.4	77.5	59.3	67.4	76.2	40.5	89.3	62.6	82.2	51.2	60.9	63.8	39.2	87.9
3D ResNet-50	49.8	78.0	37.9	59.1	44.3	67.1	37.3	48.1	80.5	40.1	50.7	63.8	39.2	87.9
3D ResNet-152	51.9	77.0	37.1	52.4	55.6	45.1	55.1	47.6	61.1	39.8	61.1	33.8	68.3	37.1
Unimodal Students														
OPT 1.3B	66.5	78.9	52.3	**73.3**	65.5	**61.8**	72.6	**70.0**	**88.4**	56.9	**80.2**	76.2	**81.0**	52.1
OPT 6.7B	68.4	77.4	61.3	64.9	**85.5**	37.8	83.7	66.1	79.0	59.6	60.2	**83.8**	37.3	82.1
LLaMA2 7B	**71.5**	**83.2**	**63.0**	67.9	81.9	48.6	**84.9**	69.4	84.9	**60.0**	65.8	82.3	57.0	**71.4**
Multimodal Students														
3D ResNet-50 (0.05B)														
+ OPT 1.3B	68.6	86.1	53.3	76.4	75.4	71.4	59.1	65.6	87.3	52.9	70.7	69.2	73.4	53.5
+ OPT 6.7B	70.8	**89.5**	**56.0**	74.8	75.8	70.7	65.9	65.7	87.4	53.5	67.2	69.2	67.7	60.0
+ LLaMA2 7B	69.0	82.8	**56.0**	69.3	83.5	57.9	66.3	68.0	84.8	55.7	**73.9**	**81.5**	**74.1**	48.6
3D ResNet-152 (0.1B)														
+ OPT 1.3B	68.9	79.6	54.8	75.3	79.0	61.3	**66.6**	67.2	73.8	**66.4**	62.1	79.3	55.8	**73.1**
+ OPT 6.7B	**71.0**	88.9	55.7	**79.5**	74.2	**75.7**	63.0	65.6	**88.0**	53.0	66.1	74.4	67.1	55.7
+ LLaMA2 7B	68.5	83.3	54.9	69.2	**84.7**	60.6	60.7	**69.4**	85.5	57.9	72.0	**81.5**	72.2	55.0

Table 2: Evaluation on unimodal and multimodal students compared with GPT-4 and vision-only baselines in AD diagnosis. The GPT-4 (Teacher Model) refers to the performance of GPT-4 augmented with clinical CoT rationales in Table 1.

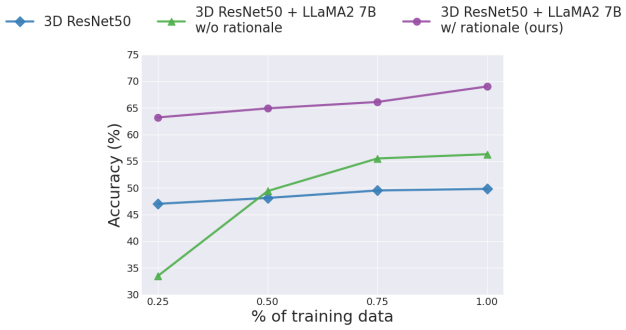


Figure 4: Data efficiency brought by clinical reasoning.

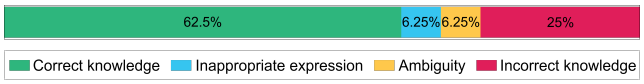


Figure 5: Analysis of rationales from GPT-4's misdiagnoses.

students outperform the baseline vision models in total accuracy, despite the image-intensive nature of AD diagnosis.

Secondly, for multimodal students: Similar to the finding from Zhang et al. (2023), with multimodal CoT, all VLMs with orders-of-magnitude smaller sizes of parameters outperform the LLM teacher in total accuracy as well as recall for AD and/or MCI. Also, all multimodal students achieve remarkably higher total accuracy than vision-only baselines.

Overall, although metrics in which student models win slightly differ when the adopted encoders change, we can conclude a general observation: most of the students exhibit higher accuracy and recall for AD and MCI than the LLM teacher. A higher recall indicates that the approach is more effective at minimizing false negatives, which suggests that the student models are better at avoiding misdiagnosing patients with AD or MCI as normal.

Figure 3 visualizes the difference in diagnostic accuracy of our student models with and without knowledge distillation from the LLM. We can clearly observe the significant improvement in performance when training the model with our data augmented with clinical rationales.

Data efficiency (RQ3). The lack of sufficient data is a long-standing problem in the biomedical domain. Thus, we experiment on our multimodal student with varying amounts of training data to examine if our framework is helpful in data-scarce scenarios. The results are in Figure 4.

Our multimodal student (purple), trained with clinical rationales, consistently outperforms the vision-only and vision-language baseline models no matter how much training data is used. Furthermore, when trained with only 25% of training data, the multimodal student exhibits higher accuracy than both baselines trained with 100% of data. Our approaches show comparable performance even when only a limited amount of training data is available. These experimental findings confirm the data efficiency brought by distilling LLMs' reasoning capacity to small diagnosis models, which is an important property in the biomedical domain.

Analysis of misdiagnosed cases (RQ4). Since the rationales are generated "during" the diagnosis, investigating the correlation between generated rationales and misdiagnoses is rather important. For that, we sample 32 rationales generated from misdiagnosed cases of GPT-4 and present the

analysis in Figure 5 (by two radiologists).

In failed cases, 75% of the rationales' contents are all medically correct. Among them, some contain expressions that generally will not be used in radiology reports, such as *"Interestingly, ..."* (Inappropriate expression; 6.25%) or show wrong usages of language due to the discrepancy between clinical and general domains (Ambiguity; 6.25%). For example, a rationale uses *"positive sign"* in a scenario that is "good for the patient". However, this term is often used in the context of "abnormal".

Only 25% of them contain at least one medically incorrect knowledge; This shows that *the misdiagnosis is not necessarily caused by ineffective rationales*. We presume optimizing the modeling on how to utilize rationales for the diagnosis (*e.g.*, dividing the generation of rationales and the diagnosis into two separate stages) may possibly lead to better performance. We leave this to future works.

Clinician Study: Quality of Rationales

Criteria for Clinical Reasoning Rationales

To investigate the LLM's role as a clinical reasoner, assessing if the quality of clinical rationales meets the standards of clinicians, is of significant importance. Prior works have incorporated human evaluations to assess the quality of machine-generated rationales (Zelikman et al. 2022; Wang et al. 2023). However, those works are limited to general domains, *i.e.*, commonsense reasoning, and mainly focused on whether the rationales justify the target.

In this work, we and a group of licensed radiologists propose a novel set of criteria specifically designed to evaluate machine-generated rationales for clinical diagnosis. We expect this to facilitate and benefit future research on eliciting rationales for clinical application.

- **CONSISTENCY**: How much a generated rationale is not contradictory to the presented data and model prediction (or the ground-truth diagnosis).
- **CORRECTNESS**: How medically correct the knowledge referred to in the rationale is.
- **SPECIFICITY**: How detailed and specific the insights provided in the generated rationale are.
- **HELPFULNESS**: How much a clinical rationale benefits the prediction towards the correct diagnosis.
- **HUMAN-LIKENESS**: How well a clinical rationale demonstrates the insight and understanding of the presented patient description or diagnosis in a way that matches the human behaviours.

Human Evaluation

We conduct human evaluation on 240 clinical reasoning rationales that are "generated during the diagnosis" with two radiologists. The results are illustrated in Figure 6. To assess them more critically, we sample rationales from "challenging cases" where "all 5 models" fail to predict the correct diagnosis (referred to as *"Misdiagnoses"*), and an equal amount of correctly diagnosed cases (referred to as *"Correct Diagnoses"*). As a reference point, we apply the rationalizing module (Module I) to generate their rationales with access to ground-truth diagnosis (denoted as *Ref*; gray-bar).

GPT-4 faithfully reflects available clinical evidence. We observe that *GPT-4* and *Ref* perform similarly within correct diagnoses. However, in *Misdiagnoses* group, rationales generated without access to ground-truth diagnosis (*GPT-4*) yield better scores even than those with access (with *Ref*), in every criterion. This phenomenon indicates that when a ground-truth diagnosis is challenging to predict based on the given patient description (*i.e.*, *Misdiagnosis* group), it is possible that prior knowledge of the diagnosis may not be beneficial for rationale annotation. We presume it stems from the discrepancy of available clinical evidence between radiologists (who annotated the dataset in the real world) and our study. As a result, when asked to condition on the ground-truth label, GPT-4 may operate as if it is being forced to construct a reasoning path that contradicts its understanding of the available clinical evidence. Therefore, the superior results of rationales generated without referencing to the ground truth in the *Misdiagnoses* group (*GPT-4*; yellow bars) manefest that GPT-4 can faithfully reflect the observed clinical evidence in the rationales.

Knowledge distillation enables better generalization. Intriguingly, the distilled student models also surpass *Ref* in *Misdiagnoses* cases. This implies that training with clinical rationales helps the student model to generate a more generalized rationale that is not biased toward a certain diagnosis. Their better diagnostic performance than the LLM teacher supports this finding (Table 2).

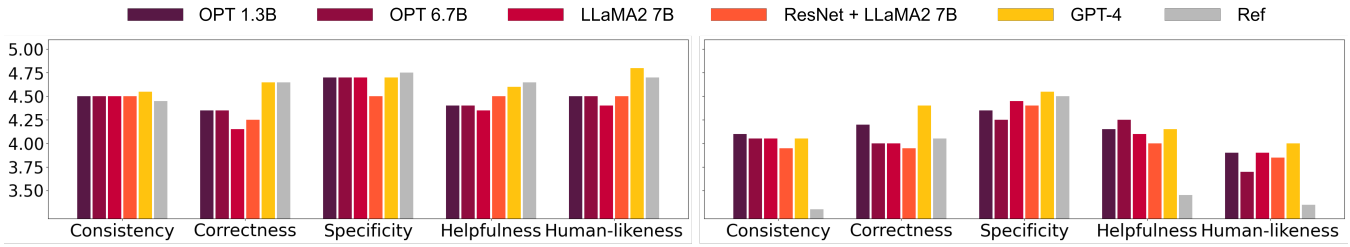
Our framework elicits effective rationales for real-world applications. Overall, rationales generated by the teacher LLM (GPT-4) and student models receive scores higher than 4 regarding almost every criterion (the lowest is 3.45 with the score range being 0-5). The promising helpfulness and correctness scores in the misdiagnoses group match our previous findings: *Misdiagnoses may not necessarily stem from ineffective rationales*, but rather from how we model the condition of rationales in diagnosis. Also, high specificity scores indicate that both the LLM and distilled students can present detailed rationales concluding their observations and insights (the average length of the clinical rationale in this study is 269.4 words).

Most importantly, the outstanding human-likeness scores, especially within the correct diagnoses, show that our rationales can effectively replicate the clinical reasoning of radiologists, and thus are more likely to seamlessly integrate into real-world radiology reports.

Case Study of Machine-generated Rationales

We highlight two important properties of the generated rationales. We describe how they are aligned with the reasoning of radiologists, present the quotes from them, and elaborate on how they can benefit DL-based diagnosis.

Interpret clinical evidence contextually. The presence of the APOE4 gene is a strong risk factor for susceptibility to Alzheimer's disease. However, in situations where other evidence suggests that the patient is normal, radiologists do not just blindly rely on APOE4, instead, they comprehensively



(a) ***Correct Diagnoses*** **(b)** ***Misdiagnoses***

Figure 6: Evaluations on clinical rationales. We report the average score (score range: 0-5).

consider all the evidence based on their expertise and experiences. Our rationales exhibit the same behavior:

> *"The patient carries one copy of APOE4 gene, which is known to increase the risk of AD. The absence of cognitive impairment symptoms and brain atrophy* ***suggest that this genetic risk has not led to any apparent neurodegeneration****."*

In conventional data-driven approaches for disease diagnosis, *i.e.*, text or image classification, it is non-trivial to annotate whether a certain feature is important in all possible scenarios. Our reasoning-aware diagnosis framework not only ameliorates this need, but also provides rationales that replicate such mechanism of human radiologists. Hence, we presume our framework has the potential to serve as a reliable tool for assisting clinicians in real-world data annotation.

Selectively summarize important evidence. In practice, after a thorough understanding and interpretation of all clinical evidence, radiologists select meaningful findings from their observations to make an accurate judgement. Such summarization of evidence can be found in our rationales:

> *"... in summary, the patient's cognitive decline, as evidenced by the MMSE score and the presence of mild hippocampal and severe atrophy in key areas related to memory, indicate a mild cognitive impairment."*

Machine-generated rationales are often too long because they contain several sentences presenting the necessary information. Liu et al. (2023) demonstrate that when LLMs are processing long text, information at the beginning or the end of the text can be more effectively utilized than those scattered around. Since today's LLMs continue growing their ability to process longer texts, being able to selectively summarise necessary information at the end of the rationales supports our framework's usefulness for future LLM-based studies in clinical domains.

Related Work

Alzheimer's Disease Diagnosis. Most DL-based methods for AD formulate the diagnosis simply as image classification and address the performance via transfer learning from general domains or tuning model architectures (Ebrahimi, Luo, and Chiong 2020; Jang and Hwang 2022). These approaches focus on better extracting the image features. However, AD diagnosis requires understanding and reasoning over a range of clinical data, such as APOE4 allele and MMSE alongside the MRIs (Budson and Solomon 2012; Weller and Budson 2018). To resolve that, several studies have exploited different aspects or features of AD: Zhu et al. (2016) and Zhang et al. (2018) approach AD diagnosis with multimodal data, *e.g.*, positron emission tomography and genetics data; Ong et al. (2023) leverage volume measurements of brain regions (*e.g.*, subcortical volume) extracted from MRIs as additional training objectives alongside the classification of AD via multi-task learning. Although these studies do facilitate diagnosis models to consider more aspects or features of the disease, none of them provides a clear picture of the reasoning behind the diagnosis.

Clinical NLP. The success of LMs has sparked a surge in applying NLP techniques to the biomedical field (Lee et al. 2020; Yue, Jimenez Gutierrez, and Sun 2020; Rajagopal et al. 2021; Kim et al. 2023; Feng et al. 2023). For example, Lee et al. (2020) fine-tune the commonly used BERT model (Kenton and Toutanova 2019) with medical corpus to endow it with biomedical knowledge, which is then implemented by Yue, Jimenez Gutierrez, and Sun (2020) to solve the clinical reading comprehension task; Rajagopal et al. (2021) and Feng et al. (2023) address the generation of explanations for various medical conditions via sequence-to-sequence language models with template-based approaches. More recently, upon the advancements of LLMs, Agrawal et al. (2022) have proposed to use LLMs to recognize named entities from clinical texts; Li et al. (2023) use LLMs to predict the synergy of drug pairs in rare human tissues that lack structured data and features.

Most prior work heavily relies on the knowledge from the pre-training corpus, ignoring whether the knowledge used is correct for the situation or follows human reasoning. In this work, we address the absence of clinical reasoning, especially in disease diagnosis, via large language models with prompt-based learning.

Conclusion

We present a reasoning-aware diagnosis framework to target the absence of clinical reasoning in most prior works. Upon that, we investigate LLMs' reasoning ability in clinical diagnosis via prompt-based learning and embark on various experiments with few-shot diagnosis and knowledge distillation. As a formal study of clinical reasoning towards real-world applications, we propose a series of novel criteria for assessing the quality of machine-generated clinical ratio-

nales. These criteria can facilitate and benefit future work in this area. Through human evaluation and extensive analysis of generated rationales, we establish a solid foundation for utilizing LLMs, both directly and indirectly, to model clinical reasoning in disease diagnosis.

Limitations. Our study has the following limitations: (1) Our prompt used to invoke LLMs' CoT reasoning only contains two rationale demonstrations due to their length (around 260 words each). This can potentially affect models' performance in rationale generation and diagnosis; (2) In our settings, the clinical rationale and the name of the predicted diagnosis are autoregressively generated. We do not explore other paradigms, such as jointly predicting them via multi-task learning or dividing the rationale generation and diagnosis into separate stages; (3) Although a group of licensed radiologists are involved in this study, we have not incorporated this framework into real-world clinical settings; (4) We do not include filtering mechanism to target ineffective rationales. Although our analysis shows that even in misdiagnosed cases, 75% of the rationales are medically correct, training student models with such a collection of rationales can potentially hinder their performance.

Ethical Statement

All of the patient data from both ADNI and AIBL are approved by the institutional review boards and de-identified for privacy. Additionally, patient information in all examples provided in this paper is partially masked manually. Assessment for potential societal impacts regarding data bias, accountability, legal challenges, and so on, is necessary before applying our method to real clinical settings.

Acknowledgements

This work is mainly supported by the Samsung Research Funding Center of Samsung Electronics (Project Number SRFC-TF2103-01), and partially supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)) and (No.2021-0-02068, Artificial Intelligence Innovation Hub) and (No.2022-0-00077, AI Technology Development for Commonsense Extraction, Reasoning, and Inference from Heterogeneous Data). Jinyoung Yeo is a corresponding author (jinyeo@yonsei.ac.kr).

References

Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; and Sontag, D. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1998–2022.

Bakator, M.; and Radosav, D. 2018. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, 2(3): 47.

Balogh, E. P.; Miller, B. T.; and Ball, J. R. 2015. Improving diagnosis in health care.

Budson, A. E.; and Solomon, P. R. 2012. New Criteria for Alzheimer's disease and Mild Cognitive Impairment: Implications for the Practicing Clinician. *The neurologist*, 18(6): 356.

Cook, D. A.; Sherbino, J.; and Durning, S. J. 2018. Management reasoning: beyond the diagnosis. *Jama*, 319(22): 2267–2268.

DeTure, M. A.; and Dickson, D. W. 2019. The neuropathological diagnosis of Alzheimer's disease. *Molecular neurodegeneration*, 14(1): 1–18.

Ebrahimi, A.; Luo, S.; and Chiong, R. 2020. Introducing transfer learning to 3D ResNet-18 for Alzheimer's disease detection on MRI images. In *2020 35th international conference on image and vision computing New Zealand (IVCNZ)*, 1–6. IEEE.

Ellis, K. A.; Bush, A. I.; Darby, D.; De Fazio, D.; Foster, J.; Hudson, P.; Lautenschlager, N. T.; Lenzo, N.; Martins, R. N.; Maruff, P.; et al. 2009. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International psychogeriatrics*, 21(4): 672–687.

Feng, S. Y.; Khetan, V.; Sacaleanu, B.; Gershman, A.; and Hovy, E. 2023. CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 313–327.

Hara, K.; Kataoka, H.; and Satoh, Y. 2017. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, 3154–3160.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Holzinger, A.; Biemann, C.; Pattichis, C. S.; and Kell, D. B. 2017. What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-k.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017. Toronto, Canada: Association for Computational Linguistics.

Jack Jr, C. R.; Bernstein, M. A.; Fox, N. C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P. J.; L. Whitwell, J.; Ward, C.; et al. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4): 685–691.

Jang, J.; and Hwang, D. 2022. M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20718–20729.

Kassirer, J. P. 1989. Diagnostic reasoning. *Annals of internal medicine*, 110(11): 893–900.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, 2.

Kim, M.; Ong, K. T.-i.; Choi, S.; Yeo, J.; Kim, S.; Han, K.; Park, J. E.; Kim, H. S.; Choi, Y. S.; Ahn, S. S.; et al. 2023. Natural language processing to predict isocitrate dehydrogenase genotype in diffuse glioma using MR radiology reports. *European Radiology*, 33(11): 8017–8025.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kumar, Y.; Koul, A.; Singla, R.; and Ijaz, M. F. 2022. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, 1–28.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, T.; Shetty, S.; Kamath, A.; Jaiswal, A.; Jiang, X.; Ding, Y.; and Kim, Y. 2023. CancerGPT: Few-shot Drug Pair Synergy Prediction using Large Pre-trained Language Models. arXiv:2304.10946.

Liévin, V.; Hother, C. E.; and Winther, O. 2023. Can large language models reason about medical questions? arXiv:2207.08143.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

Mitra, A.; Banerjee, P.; Pal, K. K.; Mishra, S.; and Baral, C. 2020. How Additional Knowledge can Improve Natural Language Commonsense Question Answering? arXiv:1909.08855.

Norman, G. 2005. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4): 418–427.

Ong, K. T.-i.; Kim, H.; Kim, M.; Jang, J.; Sohn, B.; Choi, Y. S.; Hwang, D.; Hwang, S. J.; and Yeo, J. 2023. Evidence-empowered Transfer Learning for Alzheimer's Disease. arXiv:2303.01105.

OpenAI. 2023a. ChatGPT. https://openai.com/blog/chatgpt.

OpenAI. 2023b. GPT-4 Technical Report.

Rajagopal, D.; Khetan, V.; Sacaleanu, B.; Gershman, A.; Fano, A.; and Hovy, E. 2021. Cross-domain reasoning via template filling. *arXiv preprint arXiv:2111.00539*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212.

Voss, J. L.; Bridge, D. J.; Cohen, N. J.; and Walker, J. A. 2017. A closer look at the hippocampus and memory. *Trends in cognitive sciences*, 21(8): 577–588.

Wang, P.; Wang, Z.; Li, Z.; Gao, Y.; Yin, B.; and Ren, X. 2023. SCOTT: Self-Consistent Chain-of-Thought Distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5546–5558. Toronto, Canada: Association for Computational Linguistics.

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

Weller, J.; and Budson, A. 2018. Current understanding of Alzheimer's disease diagnosis and treatment. *F1000Research*, 7.

Wu, D.; Zhang, J.; and Huang, X. 2023. Chain of Thought Prompting Elicits Knowledge Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6519–6534. Toronto, Canada: Association for Computational Linguistics.

Yue, X.; Jimenez Gutierrez, B.; and Sun, H. 2020. Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4474–4486. Online: Association for Computational Linguistics.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35: 15476–15488.

Zhang, C.; Adeli, E.; Zhou, T.; Chen, X.; and Shen, D. 2018. Multi-Layer Multi-View Classification for Alzheimer's Disease Diagnosis. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. arXiv:2302.00923.

Zheng, L.; Li, Z.; Zhang, H.; Zhuang, Y.; Chen, Z.; Huang, Y.; Wang, Y.; Xu, Y.; Zhuo, D.; Xing, E. P.; et al. 2022. Alpa: Automating inter- and intra-operator parallelism for distributed deep learning. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, 559–578.

Zhu, X.; Suk, H.-I.; Lee, S.-W.; and Shen, D. 2016. Canonical feature selection for joint regression and multi-class identification in Alzheimer's disease diagnosis. *Brain imaging and behavior*, 10: 818–828.