## ChatABL: Abductive Learning via Natural Language Interaction with ChatGPT

Tianyang Zhong<sup>1</sup>, Yaonai Wei<sup>1</sup>, Li Yang<sup>1</sup>, Zihao Wu<sup>2</sup>, Zhengliang Liu<sup>2</sup>, Xiaozheng Wei<sup>1</sup>, Wenjun Li<sup>1</sup>, Junjie Yao<sup>8</sup>, Chong Ma<sup>1</sup>, Xiang Li<sup>3</sup>, Dajiang Zhu<sup>4</sup>, Xi Jiang<sup>8</sup>, Junwei Han<sup>1</sup>, Dinggang Shen<sup>5,6,7</sup>, Tianming Liu<sup>2</sup>, and Tuo Zhang \*1

<sup>1</sup>School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

<sup>2</sup>School of Computing, The University of Georgia, Athens 30602, USA

<sup>3</sup>Department of Radiology, Massachusetts General Hospital and Harvard

Medical School, Boston 02115, USA

<sup>4</sup>Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington 76019, USA

<sup>5</sup>School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China

<sup>6</sup>Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China
 <sup>7</sup>Shanghai Clinical Research and Trial Center, Shanghai, 201210, China
 <sup>8</sup>School of life science and technology, University of Electronic Science and Technology of China, Chengdu 611731, China

#### Abstract

Large language models (LLMs) such as ChatGPT have recently demonstrated significant potential in mathematical abilities, providing valuable reasoning paradigm consistent with human natural language. However, LLMs currently have difficulty in bridging perception, language understanding and reasoning capabilities due to incompatibility of the underlying information flow among them, making it challenging to accomplish tasks autonomously. On the other hand, abductive learning (ABL) frameworks for integrating the two abilities of perception and reasoning has seen significant success in inverse decipherment of incomplete facts, but it is limited by the lack of semantic understanding of logical reasoning rules and the dependence on complicated domain knowledge representation. This paper presents a novel method (ChatABL) for integrating LLMs into the ABL framework, aiming at unifying the three abilities in a more user-friendly and understandable manner. The proposed method uses the strengths of LLMs' understanding and logical reasoning to correct the incomplete logical facts for optimizing the performance of perceptual module, by summarizing and reorganizing reasoning rules

<sup>\*</sup>Corresponding author: tuozhang@nwpu.edu.cn

represented in natural language format. Similarly, perceptual module provides necessary reasoning examples for LLMs in natural language format. The variable-length handwritten equation deciphering task, an abstract expression of the Mayan calendar decoding, is used as a testbed to demonstrate that ChatABL has reasoning ability beyond most existing state-of-the-art methods, which has been well supported by comparative studies. To our best knowledge, the proposed ChatABL is the first attempt to explore a new pattern for further approaching human-level cognitive ability via natural language interaction with ChatGPT.

#### 1 Introduction

Large Language Models (LLMs), which learn from unprecedented levels of data to generate human-like responses, have emerged as advanced artificial general intelligence (AGI) systems [4, 5, 35, 38, 58]. LLMs play a pivotal role in language translation [44], problem-solving [37], naming entity recognition [48], and text generation [11]. Inspired by their remarkable progress in natural language processing, it will be an interesting topic to bridge the capacities of perception, language understanding and reasoning (PLR) to explore advanced intelligent behaviors of humans, which has profound significance for the development of new AGI systems.

From the perspective of human perception and cognitive ability [9, 29, 42, 61, 70], image processing includes understanding the spatial relationship between objects, identifying patterns and textures, and extracting features that describe objects in images [3, 58]. More importantly, these results are further executed through rigorous logical reasoning [21, 52, 64], and language interaction to achieve information exchange with the external world. These tasks require a deep collaborative understanding of perception, language understanding and reasoning components, which is challenging for LLMs that have been primarily trained on text data [6]. This limitation is mainly manifested in how to establish an effective communication framework to satisfy the compatibility of underlying information flow among them, which provides inspiration for further research, that is, more attention should be paid to the structural design of the model on the basis of LLMs [62].

On the other hand, the latest abductive learning (ABL) framework [69] unifies perception and reasoning in a mutually beneficial manner, which overcomes heavy-reasoning light-perception deficiencies of [15], and solves heavy-perception light-reasoning flaws of [23]. Substantial advances have been made in areas such as theft judicial sentencing [43], stroke evaluation in table tennis [57], neuro-symbolic learning tasks [7, 12, 28]. Nevertheless, it is insufficient for ABL to extract semantic information of individual reasoning rules expressed by logical clauses and their collaborative relationships, which are reflected in the following aspects: 1) construction of the knowledge base requires expert knowledge and complex transformations related to logical clauses; 2) representing mutual coupling information among logical rules is burdensome, particularly when attempting to match newly added reasoning rules with existing ones; 3) the joint optimization of perception and reasoning modules requires a huge amount of

computation.

The aim of this paper is to provide a scheme called ChatABL, which combines the strength of LLMs and ABL framework to address the PLR problem in a more user-friendly and understandable manner. In this scheme, images are fed into perception module, i.e., various neural networks. Since the generated high-dimensional tensor cannot be understood by LLMs, we convert them into incomplete logical facts rendered in natural language format, conveying the partial depiction of the external environment by perception model and the accuracy is subject to verification. Then, the textual representations serve as corrective information for LLMs, which synthesize contextual cues, reasoning rules, and sample databases to realize constraint and rectification functionality. And the original perception data and the corrected results are used to supervise materials to update perception model. In turn, the optimized perception model provides the necessary logical facts for LLMs in natural language format, which can be used to strengthen LLMs. Finally, as a preliminary effort, the variable-length handwritten equation deciphering task, the well-known "holy grail" problem of artificial intelligence field [14], is regarded as a testbed in this work to testify the validity of the proposed ChatABL. Experimental results and comprehensive analysis with other mainstream methods validate the superiority of this method showing remarkable reasoning ability in solving complex reasoning problems.

The proposed method uses LLM's robust logical reasoning capabilities to boost perception models, which allows us to extract the information of large unlabeled datasets. Another advantage of ChatABL is that the comprehensive information among extended rules and original domain knowledge can be leveraged to provide interactive reasoning explanations in natural language format. Overall, the main contributions of our work are summarized as follows:

- The ChatABL firstly explores a solution for bridging perception, language understanding, and reasoning ability via natural language interaction with LLM, providing insights for future research. The proposed method can complete complex reasoning tasks under the condition of small-sample data and incomplete knowledge.
- 2) A novel knowledge-constrained self-feedback optimization strategy is designed utilizing penalty-based dynamics prompt. The proposed method utilizes the penalty-based dynamics prompt to execute trialand-error and reasoning steps iteratively for refining logical facts by introducing self-feedback mechanism.
- 3) We have preliminarily rectified the previous erroneous conception by leveraging LLM to solve the handwritten equation deciphering task: LLM-based reasoning is a fallacy, and LLM chiefly serves as a user interface.

#### 2 Related Works

# 2.1 Perception and logical reasoning systems, and abductive learning

It is suggested that human perception system and logical reasoning system are independent and mutually reinforcing [2, 19, 52, 64]. Further, Marianna B. argues that these two systems work together to form our understanding of the world around us, with each system reinforcing and supporting the other [21].

As one of the holy grail problems in AI, combining machine learning and logical reasoning has drawn much attention [30, 36, 41, 50]. Most existing methods try to combine the two different systems by making one side to subsume the other [16, 23, 66]. Another typical approach is to use deep neural networks or other differentiable functional calculations to approximate symbolic calculi [22, 55]. However, most of them still require human-defined symbols as input [49]. Few of them can make full-featured logical reasoning, and they usually require large amounts of training data.

Recently, an abductive learning(ABL) method proposed by Zhou et.al [69] targeted at unifying these two AI paradigms in a mutually beneficial way. The ABL method has been applied to a handwritten equation decipherment (HED) task, an abstract expression of the Mayan calendar decoding [27], designed by Zhou and colleagues for demonstration, and the agent models can recognise numbers and resolve unknown mathematical operations simultaneously from images of simple hand-written equations, and can be generalised to longer equations and adapted to different tasks.

The handwritten equation decipherment task mentioned (**Figure 1**) is indeed a highly condensed and classic case that reflects the ideas of connecting perception and reasoning in intelligent behavior [7, 13]. The equations are constructed from images of symbols ("0", "1", "+" and "="), and they are generated with unknown operation rules, each example is associated with a label that indicates whether the equation is correct. A machine is tasked with learning from a training set of labelled equations, and the trained model is expected to predict unseen equations correctly.

As described above, the small sample size and variable length equations make the resolution of this task challenging. The ABL framework tries to address these challenges by connecting machine learning with an abductive logical reasoning module and bridging them with consistency optimisation [14]. The framework mainly consists of three parts: machine learning, logical abduction and optimisation. In the HED task, a CNN-based machine model [26] was used to generate pseudo-labels from image pixels. Logical abduction was achieved via an abductive logic program with Prolog computer language, and the RACOS optimization tool [65], using derivative-free optimisation algorithm, was deployed to solve the consistency optimization problem between pseudo-labels and background knowledge, which was designed as a logic program [7], involved only equation structure and recursive bit-wise operation definitions. Concretely, the background knowledge about equation structures is a set of Definite Clause Grammar (DCG) rules [1, 45], which recursively defined digits as sequences of "0" and "1", and each equation had the format of

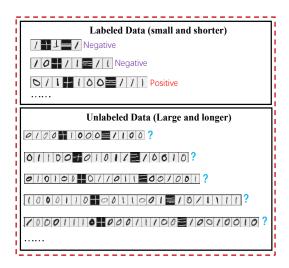


Figure 1: Illustration of the handwritten equation decipherment, whose task is to recognise symbols and discover the unknown mathematical operation behind the handwritten equations from labeled images of small and shorter equations, and can be generalised to unlabeled images of large and longer equations.

X+Y=Z, allowing for varying lengths of X, Y, and Z. The logic program operating in a digit-by-digit manner reversed X+Y using the last digit as the starting point [13].

The fact can be found that there are shortcomings in dealing with complex reasoning tasks by ABL method. The logic program could be too complex to interpret since it always involves complicated transformation with all kinds of logical predicate, hampering ABL's performance in generating accurate hypotheses and explanations. Therefore, it is necessary to find a new method to simplify the design and update of logical program.

#### 2.2 Large Language Models and ChatGPT

As one of the most influential LLMs today, ChatGPT provides a user-friendly human-machine interaction platform and API that brings the powerful capabilities of large language models to the public and has been rapidly integrated into various fields of application such as education, healthcare, and others to perform general natural language processing tasks [37, 11, 58, 40], including text classification, data augmentation, arithmetic reasoning, sentiment analysis, question and answering, summarization, and more.

Since the last century, researchers have been exploring the ability of machines to process natural language. Early methods relied on human induction of knowledge from data and teaching machines to perform tasks.

Later, with the advent of supervised learning, machines were able to learn automatically from annotated data. However, the amount of annotated data is limited compared to the vast amount of unannotated data, making it difficult for machines to learn universal knowledge. This changed with the emergence of pre-trained language models (PLMs), typically based on the Transformer architecture, and incorporating self-supervised learning. These models first learn universal knowledge from massive amounts of unannotated data, and then fine-tune on a small amount of annotated data, such as BERT [17], Google T5 [47], BART [34], and OpenAI GPT [46] series, with a parameter size of up to 1 billion.

After that, large language models (LLMs) with parameter sizes of over 100 billion have emerged, and they are still based on the Transformer architecture, but with vastly increased training data, parameter sizes, and model sizes. Researchers were surprised to find that as parameter size increases, the models exhibit new capabilities and continuous improvement in accuracy, which has led to the development of more LLMs, such as FLAN [39], GPT-3 [5], OPT [67], Bloom [51], PaLM [10], all of which have parameter sizes over 100 billion, with PaLM having 540 billion parameters.

Recent research [8, 40, 58] shows that ChatGPT have excellent high-level reasoning abilities, which can help researchers simplify experimental design and handle various non-symbolic problems. For example, Microsoft Research [6] found that GPT-4 can solve a series of mathematical reasoning problems from elementary school to university level. In this paper, our goal is to use ChatGPT to replace the reasoning module in the reverse translation learning process to determine if the output of the perception module aligns with the rules of the knowledge base.

#### 2.3 Reasoning via LLM

The general capabilities and wide applicability of LLMs incentivized researchers to explore high-level reasoning abilities. In fact, reasoning is considered to be one of the emergent abilities when language models scale up in size [6, 59].

The original GPT-3 model [5] has demonstrated much potential in common sense reasoning through in-context learning, which lays the foundation for future reasoning research with LLMs. In addition, Wei et al. [60] hypothesized and verified that when given carefully prepared prompts sequentially (chain-of-thoughts), LLMs can perform significantly better in arithmetic reasoning, deductive reasoning and common sense reasoning through decomposition of multi-step problems.

Recently, Wu et al. [63] compared the deductive reasoning abilities of large language models. This study investigates ChatGPT and GPT-4's performance in the specialized domain of radiology, using a natural language inference task and comparing them to fine-tuned models. Results reveal that GPT-4 outperforms ChatGPT, and smaller fine-tuned models (e.g., BERT) require significant data to achieve GPT-4 level performance. The findings imply that creating a generic reasoning model based on LLMs for diverse tasks across various domains is viable and practical for clinical applications.

Another work by Ma et al. [40] explored LLM's ability to comprehend radiology reports through an innovative dynamic prompting paradigm. The impression section of radiology reports is crucial for communication between radiologists and other physicians. However, it is costly to produce valid impressions from radiology findings and this process calls for automation. The authors utilize LLMs' in-context learning ability by creating dynamic contexts with domain-specific, individualized sample findings-impression pairs. This approach enables the model to acquire contextual knowledge from semantically similar examples in existing data. They also develop an iterative optimization algorithm for automatic evaluation and prompt composition to further refine the model. ImpressionGPT achieves state-of-the-art performance on MIMIC-CXR and OpenI datasets without additional training data or LLM fine-tuning, presenting a localization paradigm for LLMs applicable across various domains with specific language processing requirements.

However, there is a scarcity of research on applying LLMs to abductive reasoning. One study by Jung et al. [31] improved LLMs' ability to make logical explanations through *maieutic prompting*, which elicits abductive explanations to a problem through by recursively presenting the model with its own generated output as the inquiry. However, this study is limited to the scope of QA-style reasoning.

This work is among the first efforts to present a comprehensive abductive learning framework that not only avoid the difficult process of symbolic rule formulation but also enhance the model's interpretability. We believe that the combination of these two factors will demonstrate the potential of ChatGPT in reasoning problems.

## 3 Proposed Method

#### 3.1 Problem Setting

In the ChatABL framework for the HED task, the given input is defined as Input =  $\{X_l, X_u, KB_\theta\}$ , where tensor  $X_l$  denotes labeled data, tensor  $X_u$  denotes unlabeled data, the data size of  $X_u$  is much larger than that of  $X_l$ , and  $KB_\theta$  denotes knowledge base, which is used to constrain incomplete logical facts generated by perception module. Concretely,  $\boldsymbol{X}_{l} = \{(\boldsymbol{x}_{l1}, \boldsymbol{y}_{l1}), (\boldsymbol{x}_{l2}, \boldsymbol{y}_{l2}), \dots, (\boldsymbol{x}_{li}, \boldsymbol{y}_{li})\}$ , where variable  $x_{li}$  represents shorter variable-length handwritten equations, and  $y_{li}$  implies the corresponding labels. And the assignment of  $\mathbf{X}_l$  is to learn a mapping from x to y.  $X_u = \{x_{u1}, x_{u2}, \dots, x_{uj} \mid i \ll j\}$ , where  $x_{uj}$  denotes the unlabeled handwritten equations with longer length, are utilized to boost representative capability of the above mapping.  $KB_{\theta}$  consists of a series of domain rules in natural language format with learnable objective  $\theta$  in LLMs, which integrate fewer labeled data  $X_l$  and a large amount of unlabeled data  $X_u$  to optimize perceptual model f and mine unknown rules  $\theta$  for handwritten equation with the constraint of knowledge base using LLMs. The ChatABL algorithm yields the corresponding pseudo-labels to the unlabeled data by the classifier optimized by a small amount of labeled signals, and the produced labels may be incorrect due to the small number of training samples, which is difficult to guarantee good performance. Therefore, the ChatABL modifies the pseudo-labels and learns the reasoning rules of the knowledge base at the same time by LLMs, so that the consistency of them is maximized under the constraint of the knowledge base. Formally, the problem definition can be summarized as an optimization problem of searching **Output** under a given **Input**:

$$\min_{\boldsymbol{f},\theta} \quad \text{Loss}_{label} \left( \boldsymbol{y}_{li}, \boldsymbol{f}_{li} \right) + \text{Loss}_{\text{unlabel}} \left( \boldsymbol{\delta} \left( \boldsymbol{y}'_{uj} \right), \boldsymbol{f}_{uj} \right) \\
\text{s.t. argmax constraint } \left( \boldsymbol{\delta} \left( \boldsymbol{y}'_{uj} \right), \boldsymbol{f}_{li}, \boldsymbol{KB}_{\theta} \right) \tag{1}$$

where  $\mathbf{y}'_{uj}$  is the pseudo-label corresponding to the  $j^{th}$  unlabeled instance, which is generated by the perceptual module.  $\boldsymbol{\delta}\left(\cdot\right)$  indicates an implicit heuristic function learned by LLM, which aims to revise pseudolabels by logical reasoning process. In addition to correcting inconsistent pseudo-labels, this goal also helps the knowledge base to find unknown rules  $\theta$ . It can be seen from Eq.1 that the major challenge is how to mine the effective information of more massive and complex unlabeled image data under the  $KB_{\theta}$  constraints, and react to the iterative update of itself and reason the underlying mathematical laws behind logical events. Noted that LLMs can perform reasoning tasks based on their understanding of language without requiring optimization. It can be founded from Eq.1 that the introducing of LLM to the aforementioned problem will result in transformation of the optimization goal for ABL method, from the joint optimization problem of discrete variables and continuous variables to the optimization problem of continuous variables only for the perception model. The proposed issues in the Introduction part of this paper may be potentially addressed from this perspective. Further research details will be elaborated in the following sections.

#### 3.2 Overall Framework

ChatABL takes the images of variable length equations as input, extracts and embeds element features from the images, and outputs the judgment results of these equations and their underlying law. The key problems during constructing ChatABL are as follows.

- 1) KP1: How to make use of image information in limited labeled variable-length handwritten equations? Labeled large and longer handwritten equations are limited due to their high demand for proficient domain knowledge, which can be seen from the prototype problem (the deciphering process of the Mayan calendar [27]) that it is difficult for the most outstanding expert to directly give a definite answer or even solve it in most cases when facing the strange number. Therefore, the rational use of a small amount of labeled data poses a huge challenge to the generalization ability of agent systems.
- 2) KP2: How can we learn to leverage existing clues to tackle higher-level reasoning in complex reasoning tasks? In the case of unknown arithmetic operations, it is challenging for many individuals, particularly those unfamiliar with the domain, to find the law of unlabeled handwritten equations with limited rules and small and short la-

beled handwritten images. Therefore, how to effectively mine the existing conditions and learn the rules embedded within them to accomplish more challenging reasoning is an intriguing topic.

3) KP3: How to combine the perceptual information from images with domain knowledge rules in understandable manner? It is difficult to inject symbolic knowledge into the optimization of numerical values in machine learning models [57]. The major obstacle is that the current methods face difficulties in representing and synthesizing rules. Moreover, their corresponding reasoning processes fall short in terms of interpretability, rendering them challenging to express in natural language.

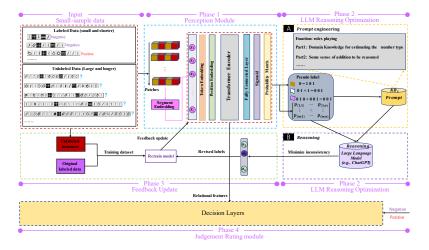


Figure 2: Schematic representation of the ChatABL framework for HED task, which leverages LLM's logical reasoning ability to correct the understanding of the images of the equations from the perception module, thereby optimizing the recognition of symbols and the mining ability of the mathematical operation behind them.

Considering the aforementioned issues, we constructed ChatABL into four phases including Perception Module, LLM Reasoning Optimization, Feedback Update and Judgement Rating module (Figure 2), which leverage ABL combined with LLM to extract and embed the variable length equations attributes using limited labeled data. Phase 1 involves using a recognition component as perception module to embed the equations. Then we introduce a logical reasoning component (KP1) based on domain knowledge rules described in natural language in Phase 2, which leverages reasoning ability of large language models to strengthen pseudolabels of unlabeled data (KP2). In Phase 3, the raw perception data and the revised labels are then used as supervision materials to iteratively improve the performance of the attribute recognition component in a feedback update manner (KP3). Finally, a classifier is trained in Phase 4, which learns evaluation in image-based methods to obtain quantified evaluation results. We judge the veracity of handwritten equations into

two levels as training labels and train the model to classify the level of unlabeled equations based on the embedding vectors in **Phase 1**.

The model trained in **Phase 3** is the same as the perception model introduced in **Phase 1**, and cross-entropy loss is utilized for its training [68]. Additionally, a conventional fully-connected layer is used for the decision layer, which has been elaborated elsewhere [13]. This paper first provides a brief introduction on the perception model and then focuses on the LLM reasoning optimization.

#### 3.3 Perception Model

The perception model serves as the backbone for recognizing image inputs. The primary goal of this perception model is to extract visual features from the data. Specifically, the extracted features will be used to construct logical facts and mathematical expressions, which are used for subsequent processing. In this study, we employ the Vision Transformer (ViT) [18] as the perception model to process the handwritten inputs. It is noteworthy that our framework is compatible with any vision model.

The ViT, originally adapted from the Transformer architecture in the domain of natural language processing, has demonstrated exceptional performance in various computer vision tasks, often surpassing traditional convolutional neural networks (CNNs). By treating an image as a sequence of tokens, the ViT effectively captures both local and global contextual information through its self-attention mechanism. Leveraging this powerful architecture allows us to effectively discern complex patterns and intricate spatial relationships present in handwritten mathematical expressions.

To maintain the independence of mathematical expressions in the handwritten input (the **Input** part in **Figure 2**), we split them into individuals  $X_u = \{x_{u1}, x_{u2}, \dots, x_{un}, \dots, x_{uj}\}$  using segment embedding. Meanwhile, we reshape  $x_{un} \in R^{H \times W \times C}$  into a sequence of flattened patches  $X_{upn} \in R^{N \times P^2 \times C}$  to serve as input for the ViT.  $N = \frac{HW}{P^2}$  and P denote the number of patches and the resolution of image patches, respectively. Every mathematical expression  $X_{upn} \in R^{N \times P^2 \times C}$  is made up of several patches representing binary digits ("0" and "1") or mathematical operators. The ViT generate pseudo-labels for subsequent processing as follows:

$$\boldsymbol{z}_0 = [\boldsymbol{x}_{cls}; \boldsymbol{x}_{up1} \boldsymbol{L}; \boldsymbol{x}_{up2} \boldsymbol{L}; \dots; \boldsymbol{x}_{upn} \boldsymbol{L}; \dots; \boldsymbol{x}_{upN} \boldsymbol{L}] + \boldsymbol{L}_{pos}$$
 (2)

$$z_m = MLP(LN(MSA(LN(z_{m-1})) + z_{m-1})) + LN(MSA(LN(z_{m-1})) + z_{m-1})$$
(3)

$$y' = SIG(FCL(z_M))$$
 (4)

where  $z_0$  is the latent vector mapped by a trainable linear projection from a sequence of flattened patches, consisting of token embedding and position embedding  $L_{pos}$  (Eq.2).  $[z_m, m = 1, 2, ..., M]$  represents

the results of the transformer encoder at different layers, which includes multiheaded self-attention (MSA), multilayer perceptron (MLP), and layernorm (LN) (**Eq.3**). y' generates the pseudo-label with a probability matrix based on the output from  $z_M$  (**Eq.4**). SIG and FCL denote the fully connected layer and sigmoid, respectively.

#### 3.4 LLM Reasoning Optimization

In this work, we employ penalty-based dynamic prompt and rulesconstrained self-feedback optimization strategy to enhance the adaptation of LLM for complex reasoning tasks. The main concept behind our approach is inspired by the human process of solving complex reasoning tasks, with "solving math problems" (SMP) being an illustrative example. The process involves interpreting known conditions and transforming them into several advantageous conditions for efficient problem-solving. Usually, mathematical conditions involved in complex problems are not typically encountered previously, or the specific forms utilized may vary from those previously encountered. In such cases, we rely on reasoning abilities to perform multiple iterations of trial and reasoning, under given rules and examples. If our reasoning results conform with the rules, we pass the test; if not, we would receive reminders indicating contradictions or deviations from the rules, which guides further reasoning and decision-making. In essence, the capacity of advanced human reasoning is characterized by iterative reasoning under rule constraints and reminders. This abstract representation of how humans solve complex reasoning tasks inspires us to design the relevant prompts and optimization strategies to enhance LLMs' powerful logical reasoning abilities. Overall, our method requires a small number of examples, facilitating LLM's robust logical reasoning ability with limited rules. Further details will be presented in the subsequent sections.

#### 3.4.1 Penalty-based Dynamic Prompt

Before introducing the Penalty-based Dynamic Prompt, we first explicate the input structure of LLM, which is accessible via a string to the API and divided into three components: System, User, and Assistant. The System message is typically invoked at the outset to demarcate the task and constrain the Assistant's behavior, while the User message serves to provide direction and constitutes the user's input. The model output is generated by the Assistant component, which serves as the basis for our dynamic prompt and iterative optimization framework.

Prior studies have used fixed-form prompts for straightforward tasks that could be easily generalized. However, these prompts lack the necessary prior knowledge for complex tasks and domain-specific datasets, resulting in low performance [40]. Thus, we propose a hypothesis that constructing dynamic prompts in a reward-punishment manner from relevant domain-specific corpora can enhance the model's comprehension and perception.

Inspired by SMP process, when solution result is inconsistent with existing rules, reminders are given to indicate the contradiction. These reminders correspond to the prompt of LLMs for adaptive adjustment. Therefore, the prompt can be updated by using this kind of feedback information as a guide for LLMs. Based on the premise, two prompts are designed in this paper: consistency discrimination prompt (CDP) and re-reasoning dynamic prompt (RDP), as shown in **Figure 3(A)(B)**. The primary function of the CDP is to determine whether the pseudo-labels

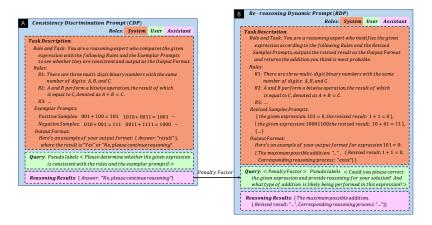


Figure 3: Details of penalty-based dynamic prompt, which consists of consistency discrimination prompt (CDP) and re-reasoning dynamic prompt (RDP). The CDP reminds whether to perform the reasoning process RDP by the penalty factor. They contains task description, query and reasoning results components.

obtained by the perceptual model are consistent with the given rules and examples. If inconsistency is detected, it serves as a reminder for the RDP (corresponding to the SMP process) to adjust dynamically in a punishment way. If consistency is detected, there is no need for the RDP. Both prompts are composed of task description, query and reasoning results. In the CDP, we first assigned the role of a reasoning expert to LLM for task description, after which we presented the reasoning task accompanied by a set of rules, exemplar prompts, and output format. The significance of this configuration lies in its ability to define the precise nature of the task undertaken by the model, serving as the fundamental framework for the entire prompt. After that, we generate the prefix query sentence: "Please determine whether the given expression is consistent with the rules base and the exemplar prompts ?", which is used to further specify consistency discrimination task. At the end of CDP, the reasoning result is integrated into the RDP query as a penalty factor. In the RDP, the task description format is similar to that of CDP, but with minor variations as the RDP accomplishes reasoning and correction tasks. Specifically, the rules applied in this task are consistent with those of CDP. The reasoning task is designed to correct a given expression. Figure 3(B) illustrates the exemplar prompts, and the output format comprises the corrected expression and its corresponding reasoning process. After that, we generate the query sentence: "Could you please correct the given expression and provide reasoning for your solution? And what type of addition operation is likely being performed in this expression?", which is used to further specify re-reasoning task. Note that the penalty factor from CDP needs to be added to the query before proceeding. At the end of the dynamic prompt, the results produced in this manner utilize the inconsistency between the given rules and pseudo-labels to strengthen the LLM's reasoning direction and enhance its performance.

#### 3.4.2 Knowledge-Constrained Self-Feedback Optimization

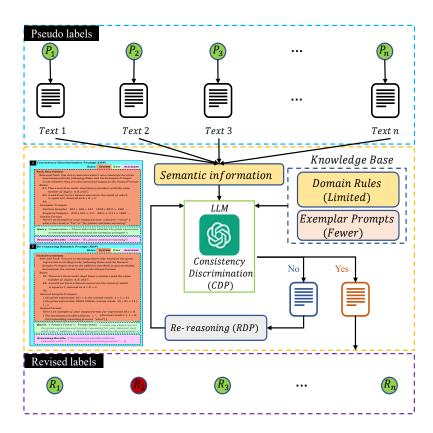


Figure 4: Details of knowledge-constrained self-feedback optimization strategy, which utilizes LLM with the penalty-based dynamics prompt to execute trial-and-error and reasoning steps iteratively for refining pseudo-labels by the consistency between abductive information and knowledge base.

In **section 3.4.1**, it can be seen that the design of the penalty-based dynamic prompt highlights a one-time effectiveness for LLM, which differs from humans' multiple trial-and-error reasoning process when solving

complex mathematical problems. This limitation may lead to uncertainty regarding whether the LLM's response aligns with our anticipated outcomes. This paper presents a novel knowledge-constrained self-feedback optimization strategy that leverages incomplete but certain knowledge, including existing rules and fewer examples, as discriminative criterion for reasoning incomplete facts, specifically pseudo-labels that require further validation. The proposed method utilizes the penalty-based dynamics prompt to execute trial-and-error and reasoning steps iteratively for refining pseudo-labels by introducing self-feedback mechanism. The epoch of iterations is dynamically determined based on consistency results between incomplete facts and knowledge, eliminating the requirement for laborious and time-consuming manual intervention.

In detail, the proposed method has been elaborately illustrated in Figure 4. Firstly, the pseudo-labels generated by the perception model are transformed into corresponding text labels (e.g., combinations of symbols "0", "1", "+", and "="), followed by semantic interpretation of the text information. To determine whether the correction process is required, we utilize the consistency between the semantic interpretation and the knowledge base (the same as the rules and exemplar prompts fields described in section 3.4.1), which relies on the CDP, described in Figure 3(A). However, in the limited labeled data, the perception model's ability to generate consistent pseudo-labels with respect to the rules and exemplar prompts for new instance is often inadequate. In such case, we use the corresponding reasoning results: "No, please continue reasoning", and it is produced by the CDP to identify inconsistencies, which serves as the query part of the RDP for re-reasoning task, as illustrated in Figure **3(B)**. Subsequently, the reasoning results generated by the LLM with RDP are repeatedly fed back into the CDP to validate consistency until the revised labels are in alignment with the knowledge base, which are eventually outputted as the revised labels.

The advantage of this knowledge-constrained self-feedback approach lies in the fact that the LLM conducts abducive reasoning on incomplete facts under the constraint of incomplete knowledge. This enables LLM to provide a reliable guarantee for the perception model in terms of knowledge constraints within the existing domain knowledge system, while also iteratively updating its own abstraction description of logical facts. As a result, the reasoning capability of LLM is significantly strengthened, which enhances its effectiveness in complex logical reasoning situations with incomplete knowledge.

## 4 Experimental Study

#### 4.1 Experiment Setup

We construct a sequence of raw images of digits and operators to the hand writing equations based benchmark handwritten character datasets [33, 54], as shown in **Figure 2**. We furnish ChatABL according to **Figure 3** with background domain knowledge in natural language format regarding arithmetic structural rules [13]. It is noteworthy that these rules do

not contain information about the type of computation performed within the equations; instead, ChatABL has to derive that information from the available data. An illustrative example is depicted in Figure 3. We evaluate the learning performance of ChatABL by benchmarking it against Meet [12], CNN-BiLSTM [13, 25], Transformer(TF) [56], and ABL [14], which involves pure perception methods (the former three), perception and reasoning methods (the last one). These state-of-the-art models are capable of addressing sequential input tasks. The training set utilized for all methods comprises equations ranging in length from 5 to 10, with each length containing 500 randomly-sampled equations. During the testing phase, all methods are tasked with predicting 5000 equations, ranging in length from 5 to 26, with 500 examples for each length. The same configuration is employed for prompt, reasoning optimization, and feedback updating. In our experiments, we adopt Accuracy, Precision, Recall, F1, and AUC [20] as evaluation metrics in our experiments, which essentially cover most of the metrics in the field.

#### 4.2 Comparison on Small Samples Performance

Table 1: Performance comparison of the ChatABL method to state-of-the-art methods as the proportion change of training dataset in terms of metrics (%) for handwritten equation recognition task.

${\bf Method/Metrics}$	Accuracy	Precision	Recall	F1	AUC
MEET-20	51.53	60.86	50.68	53.71	50.41
${\rm CNN\text{-}BiLSTM\text{-}20}$	54.71	66.94	52.84	59.06	53.19
ChatABL-20	68.92	78.01	66.90	69.43	68.44
ABL-20	91.36	98.28	87.34	91.93	87.85
MEET-50	54.70	69.91	53.13	60.38	53.63
${\rm CNN\text{-}BiLSTM\text{-}50}$	56.44	64.75	54.55	59.22	56.56
ChatABL-50	70.31	79.23	69.56	71.32	70.44
ABL-50	93.18	98.95	88.81	93.61	93.05
TF-80	63.18	86.76	58.49	69.87	64.63
MEET-80	64.47	90.45	59.27	71.62	65.18

Table 1 presents the experimental results of performance metrics for different comparative methods as the proportion changes in the training set, which consists of three subsets in the case of 20%, 50% and 80% labeled data as the training set. Compared to the ABL method based on first-order logical formula reasoning, the performance of ChatABL is lower than that of ABL, because the ABL method revises the error iteratively between images about handwriting equations and knowledge base by optimizing joint consistency on continuous and discrete variables. In contrast, the method proposed in this paper does not need to involve the optimization of complex discrete variables. The corresponding rules are described by natural language description, and then the perception mod-

ule is modified by using the reasoning ability of LLM, which also achieves good performance. This indicates that LLM has an advantage in unifying perception and understanding reasoning tasks through natural language interaction.

On the other hand, by comparing the results of the pure perception method and ChatABL when the labeling rate is low (the first two sections in **Table 1**), the gains of using both unlabeled data and knowledge rules for ChatABL are much higher. Further comparison on the third section shows that the ChatABL performance with 20% and 50% training set can even be superior the methods (TF,MEET) with 80% training set. The results verified that logic consistency can be very useful for providing a surrogate supervised signal through LLM reasoning process for revising the pseudo-labels from perception module. In short, LLMs based on pre-trained language models work better than previous studies, which allows the model to adequately learn the prior knowledge of the complex reasoning domain.

#### 4.3 Evaluation on Perceptual Model

In order to validate the performance effect of the perception module in the overall ChatABL framework on the HED tasks, we test the AlexNet [32], GoogleNet [53], ResNet [24], and ViT [18] models as the perception component of the ChatABL framework, respectively, because they are the most advanced benchmark models in the development of convolutional neural networks on solving tasks from image input. In particular, the ViT surpasses traditional CNNs in various computer vision tasks, because it treats images as token sequences that capture both local and global contextual information through self-attention mechanism. Additionally, we use the same metrics as the **4.2 section** for comparison.

Table 2: The influence of different perception modules and GPTs on ChatABL as the 20% proportion of training dataset in term of metrics (%).

Method/Metrics	Accuracy	Precision	Recall	F1	AUC
AlexNet	64.17	72.23	58.98	61.95	63.86
GoogleNet	66.54	76.58	62.19	67.09	65.96
ResNet	66.44	78.98	59.35	66.26	64.51
ViT	68.92	78.01	66.90	69.43	68.44
GPT-3.5-Turbo	68.92	78.01	66.90	69.43	68.44
GPT-4	70.01	79.51	67.82	70.63	69.85

As illustrated in the first part of **Table 2**, the overall performance impact of ChatABL is affected by different perception modules. In general, the performance improves progressively as better perception methods, only except that ChatABL with ResNet as the perception module presents slightly worse performance compared to that with GoogleNet.

And the approach of utilizing ViT as the perception model for Chat-ABL ultimately achieved the best performance among the comparative methods, with an accuracy rate approaching 70%. This represents an advancement in the context of HED tasks. Based on the results, it can be founded that an improved perception accuracy indeed impacted the performance of equation classification by ChatABL, which is similar to the effect of human sensory system on the whole central nervous system.

#### 4.4 The Impact of Different GPTs

In this section, we primarily compared the impact of the latest large language models on ChatABL. Since small-scale language models with limited parameters cannot provide meaningful reasoning processes [58], they were excluded from this experiment. We selected GPT-3.5-Turbo and GPT-4 as the reasoning component in the second phase of the ChatABL method. The corresponding parameter quantities of the two models are approximately 175 billion and 1 trillion, respectively [5]. Their parameter magnitudes differ by almost three orders of magnitude, thus comparing the impact of the two models on the overall performance is sufficient.

The lower part of **Table 2** reports the performance metrics. It can be seen that the ChatABL reasoning module's transition from GPT-3.5-Turbo to GPT-4 leads to an overall increase in all metrics, particularly with in greatest gain in accuracy from 68.92% to 70.01%. Taken together with the findings in **section 4.2**, these results demonstrate significant progress in studying the generalization ability of modern intelligent systems with small-sample data, and in synthesizing perception and reasoning via natural language as a medium. Overall, the reasoning capability of language models is proportional to their size, highlighting the critical factor of the logistic reasoning capability of LLMs.

#### 5 Discussion and Conclusion

This paper first explore a novel framework, ChatABL, which strives to achieve integration of perception, language understanding, and reasoning capabilities in user-friendly manner. To achieve this goal, we propose the incorporation of large language models into abductive learning, and design the prompt akin to human problem-solving, to enhance the effectiveness of ChatABL in complex mathematical reasoning task. And we have fundamentally rectified the previous erroneous conception: LLM-based reasoning is a fallacy and LLM chiefly serves as user interface.

However, the proposed method still has limitations to be solved. Compared to humans, the performance of large language models is heavily dependent on the context and design of the prompt. And there is significant room for improvement in prompt engineering in the future. Additionally, in our study, we acknowledge that there are areas where this paper lacks rigor. We have conducted a qualitative analysis of prompt design, but we have not done a quantitative analysis. And the number of tokens in LLM is limited, which may affect the performance of the model. Once the latest model, GPT5.0, is released in the future, we will conduct further in-depth

investigation. Furthermore, we envision expanding ChatABL to encompass multimodal data perception and integrated reasoning with diverse domain knowledge, providing novel directions for further research.

#### References

- [1] Abramson, H.: Definite clause translation grammars. University of British Columbia (1984)
- [2] Aksyuk, V.: Consciousness is learning: predictive processing systems that learn by binding may perceive themselves as conscious. arXiv preprint arXiv:2301.07016 (2023)
- [3] Biederman, I.: Human image understanding: Recent research and a theory. Computer vision, graphics, and image processing **32**(1), 29–73 (1985)
- [4] Blum, L., Blum, M.: A theoretical computer science perspective on consciousness and artificial general intelligence. arXiv preprint arXiv:2303.17075 (2023)
- [5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020)
- [6] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023)
- [7] Cai, L.W., Dai, W.Z., Huang, Y.X., Li, Y.F., Muggleton, S.H., Jiang, Y.: Abductive learning with ground knowledge base. In: IJCAI. pp. 1815–1821 (2021)
- [8] Cai, X., Liu, S., Han, J., Yang, L., Liu, Z., Liu, T.: Chestxraybert: A pretrained language model for chest radiology report summarization. IEEE Transactions on Multimedia (2021)
- [9] Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S., Delbridge, K.: Reactions to cognitive ability tests: the relationships between race, test performance, face validity perceptions, and test-taking motivation. Journal of Applied Psychology 82(2), 300 (1997)
- [10] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
- [11] Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S., Zhu, D., et al.: Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007 (2023)
- [12] Dai, W.Z., Muggleton, S.H.: Abductive knowledge induction from raw data. arXiv preprint arXiv:2010.03514 (2020)

- [13] Dai, W.Z., Xu, Q.L., Yu, Y., Zhou, Z.H.: Tunneling neural perception and logic reasoning through abductive learning. arXiv preprint arXiv:1802.01173 (2018)
- [14] Dai, W.Z., Xu, Q., Yu, Y., Zhou, Z.H.: Bridging machine learning and logical reasoning by abductive learning. Advances in Neural Information Processing Systems 32 (2019)
- [15] De Raedt, L., Kersting, K.: Probabilistic inductive logic programming. Springer (2008)
- [16] De Raedt, L., Kimmig, A.: Probabilistic (logic) programming concepts. Machine Learning 100, 5–47 (2015)
- [17] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
- [18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
- [19] Ferilli, S.: Gear: A general inference engine for automated multistrategy reasoning. Electronics 12(2), 256 (2023)
- [20] Fu, M., Liu, J., Zhang, H., Lu, S.: Multisensor fusion for magnetic flux leakage defect characterization under information incompletion. IEEE Transactions on Industrial Electronics 68(5), 4382–4392 (2020)
- [21] Ganapini, M.B., Campbell, M., Fabiano, F., Horesh, L., Lenchner, J., Loreggia, A., Mattei, N., Rahgooy, T., Rossi, F., Srivastava, B., et al.: Combining fast and slow thinking for human-like and efficient navigation in constrained environments. arXiv preprint arXiv:2201.07050 (2022)
- [22] Garcez, A.S.d., Gabbay, D.M., Ray, O., Woods, J.: Abductive reasoning in neural-symbolic systems. Topoi 26, 37–49 (2007)
- [23] Getoor, L., Taskar, B.: Introduction to statistical relational learning. MIT press (2007)
- [24] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015), http://arxiv. org/abs/1512.03385
- [25] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
- [26] Hossain, M.B., Naznin, F., Joarder, Y., Islam, M.Z., Uddin, M.J.: Recognition and solution for handwritten equation using convolutional neural network. In: 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR). pp. 250–255. IEEE (2018)
- [27] Houston, S.D., Mazariegos, O.F.C., Stuart, D.: The decipherment of ancient Maya writing. University of Oklahoma Press (2001)

- [28] Huang, Y.X., Dai, W.Z., Cai, L.W., Muggleton, S.H., Jiang, Y.: Fast abductive learning by similarity-based consistency optimization. Advances in Neural Information Processing Systems 34, 26574–26584 (2021)
- [29] Humphrey, R.: How work roles influence perception: Structuralcognitive processes and organizational behavior. American Sociological Review pp. 242–252 (1985)
- [30] Janiszewski, C., van Osselaer, S.M.: Abductive theory construction. Journal of Consumer Psychology 32(1), 175–193 (2022)
- [31] Jung, J., Qin, L., Welleck, S., Brahman, F., Bhagavatula, C., Bras, R.L., Choi, Y.: Maieutic prompting: Logically consistent reasoning with recursive explanations. arXiv preprint arXiv:2205.11822 (2022)
- [32] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM 60(6), 84–90 (2017)
- [33] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
- [34] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequenceto-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019)
- [35] Li, Y., Duan, Y.: The wisdom of artificial general intelligence: Experiments with gpt-4 for dikwp. arXiv preprint (2023)
- [36] Lin, Y., Ou, S.: Exploit domain knowledge: Smarter abductive learning and its application to math word problems. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2022)
- [37] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al.: Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852 (2023)
- [38] Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Liu, W., Shen, D., Li, Q., et al.: Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032 (2023)
- [39] Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H.W., Tay, Y., Zhou, D., Le, Q.V., Zoph, B., Wei, J., et al.: The flan collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:2301.13688 (2023)
- [40] Ma, C., Wu, Z., Wang, J., Xu, S., Wei, Y., Liu, Z., Guo, L., Cai, X., Zhang, S., Zhang, T., et al.: Impressiongpt: An iterative optimizing framework for radiology report summarization with chatgpt. arXiv preprint arXiv:2304.08448 (2023)
- [41] Magnani, L.: Why abductive cognition goes beyond just learning from data. In: Living Beyond Data: Toward Sustainable Value Creation, pp. 39–69. Springer (2022)

- [42] Nes, A., Sundberg, K., Watzl, S.: The perception/cognition distinction. Inquiry 66(2), 165–195 (2023)
- [43] Ouyang, L., Huang, R., Chen, Y., Qin, Y.: A sentence prediction approach incorporating trial logic based on abductive learning. Applied Sciences 12(16), 7982 (2022)
- [44] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., Tao, D.: Towards making the most of chatgpt for machine translation. arXiv preprint arXiv:2303.13780 (2023)
- [45] Pereira, F.C., Warren, D.H.: Definite clause grammars for language analysis—a survey of the formalism and a comparison with augmented transition networks. Artificial intelligence 13(3), 231–278 (1980)
- [46] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
- [47] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21(1), 5485–5551 (2020)
- [48] Rezayi, S., Dai, H., Liu, Z., Wu, Z., Hebbar, A., Burns, A.H., Zhao, L., Zhu, D., Li, Q., Liu, W., et al.: Clinicalradiobert: Knowledge-infused few shot learning for clinical notes named entity recognition. In: Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings. pp. 269–278. Springer (2022)
- [49] Russell, S.: Unifying logic and probability. Communications of the ACM 58(7), 88–97 (2015)
- [50] Sapir, M.: Machine learning is abduction inference. arXiv preprint arXiv:2206.07586 (2022)
- [51] Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
- [52] Sloman, S.A.: The empirical case for two systems of reasoning. Psychological bulletin **119**(1), 3 (1996)
- [53] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
- [54] Thoma, M.: The hasyv2 dataset. arXiv preprint arXiv:1701.08380 (2017)
- [55] Towell, G.G., Shavlik, J.W.: Knowledge-based artificial neural networks. Artificial intelligence 70(1-2), 119–165 (1994)
- [56] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)

- [57] Wang, J., Deng, D., Xie, X., Shu, X., Huang, Y.X., Cai, L.W., Zhang, H., Zhang, M.L., Zhou, Z.H., Wu, Y.: Tac-valuer: Knowledge-based stroke evaluation in table tennis. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 3688–3696 (2021)
- [58] Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023)
- [59] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
- [60] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022)
- [61] Williams, D.: Hierarchical minds and the perception/cognition distinction. Inquiry 66(2), 275–297 (2023)
- [62] Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
- [63] Wu, Z., Zhang, L., Cao, C., Yu, X., Dai, H., Ma, C., Liu, Z., Zhao, L., Li, G., Liu, W., Li, Q., Shen, D., Li, X., Zhu, D., Liu, T.: Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. arXiv preprint arXiv:2304.09138 (2023)
- [64] Xu, F., Liu, J., Lin, Q., Zhao, T., Zhang, J., Zhang, L.: Mind reasoning manners: Enhancing type perception for generalized zero-shot logical reasoning over text. arXiv preprint arXiv:2301.02983 (2023)
- [65] Yu, Y., Qian, H., Hu, Y.Q.: Derivative-free optimization via classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
- [66] Zadeh, L.A.: Fuzzy sets. Information and control 8(3), 338–353 (1965)
- [67] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
- [68] Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems 31 (2018)
- [69] Zhou, Z.H.: Abductive learning: towards bridging machine learning and logical reasoning. Science China Information Sciences 62, 1–3 (2019)
- [70] Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., Dong, Q., Kanwisher, N., Liu, J.: Heritability of the specific cognitive ability of face perception. Current Biology 20(2), 137–142 (2010)