# A Survey for Multimodal Mathematical Reasoning

Anonymous CVPR submission

Paper ID *****

## Abstract

*Multimodal numerical reasoning, the ability to reason with and integrate information across multiple modalities, has become an increasingly important area of research in both natural language processing (NLP) and computer vision (CV) domains. Multimodal numerical reasoning is designed to extract information from multiple input modalities, such as text, image, etc., and merge them into a comprehensive conclusion. In this survey, we review and provide an overview of the recent advancements in multimodal numerical reasoning, including datasets,evaluation metrics and methods. In particular, we focus on the emerging capabilities of large language models (LLMs) in out-of-the-box tasks of arithmetic, common sense, and symbolic reasoning.While we conducted experiments on GPT-3.5 turbo's mathematical information extraction for a single modality limited by the openness of model functions.we also outline some of the remaining limitations and future research directions in this field, including the need for more comprehensive benchmarks and the development of models that can reason with more complex and diverse modalities.*

## 1. Introduction

Multimodal reasoning [2] is a fascinating area of research that combines different modalities, such as language, vision, and numerical data, to perform complex reasoning tasks. The ability to reason across multiple modalities is a hallmark of human intelligence, and as such, it has become a critical area of research in artificial intelligence.Recent advances in natural language processing have enabled the development of cutting-edge language models [13] with impressive capabilities in reasoning over textual data. To address this challenge, a surge of numerical and arithmetic reasoning datasets have been proposed in recent years to benchmark the capabilities of deep learning models.Multimodal numerical reasoning [8, 54] is particularly challenging, where the integration of quantitative data with other modalities can greatly enhance the reasoning process. The recent advancements in multimodal numerical reasoning using LLMs hold great promise for a wide range of applications.

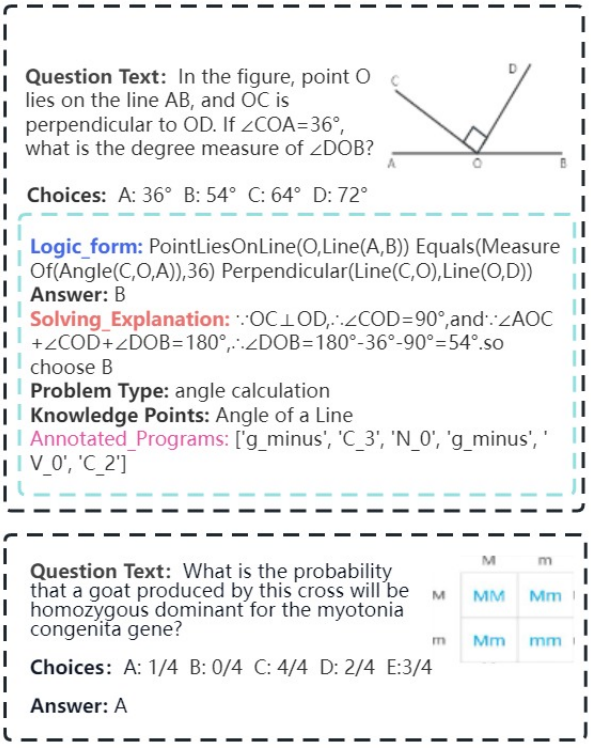



Figure 1. The main components of the content of the multimodal numerical reasoning dataset are displayed. The top is the form of the geometric data set, and the bottom is the form of the scientific visual questioning reasoning data set.

In recent years, there has been a surge of interest in multimodal numerical reasoning due to advances in deep learning and the availability of large multimodal datasets. This has led to the development of new models and techniques for multimodal numerical reasoning.In particular, LLMs emerge abilities in some arithmetic, general knowledge, and symbolic reasoning tasks [57].Some large integrated models with visual modules [59, 67] can understand

text and images together to solve problems, identify relationships between visual elements in a graph, and even perform complex arithmetic calculations. This makes them particularly well-suited for multimodal numerical reasoning tasks, which often involve combining information from different sources to make inferences and solve problems. However, there are also challenges associated with using LLMs for multimodal numerical reasoning. For example, they may struggle with certain types of reasoning tasks that require more abstract or symbolic reasoning. Additionally, the complexity of LLMs can make them difficult to interpret and diagnose when errors occur.The solution of geometric problems [7, 54] in multimodal numerical reasoning is representative, and various neural network methods for numerical reasoning have been effectively applied to solve this complex task. In addition, researchers are gradually exploring multimodal datasets and pre-training methods involving specific domain knowledge for scientific question-answering [34] in other STEM fields, such as chemistry and physics.

Mathematical reasoning ability is an important reflection of Numerical reasoning ability to some extent.In this survey paper,we review the progress of deep learning in the areas of numerical and mathematical reasoning.Specifically, we analyze various datasets and evaluation metrics in Section 2, and summarize the different methods used in multimodal numerical reasoning in a timeline, integrate the limitations they involve, and present the latest models and techniques for multimodal numerical and mathematical reasoning.In addition, we designed corresponding prompts to test part of the data in the GeoQA [7] and LILA [38] datasets on GPT series models such as GPT3.5, hugginggpt [53], etc. Through further comparison with the existing benchmarks, we found that the models has insufficient arithmetic and geometric theorem reasoning ability for geometrically related mathematical problems expressed in single or multiple mode. According to this, we put forward some possible development directions in the future.

## 2. Datasets,Tasks and Evaluations

In this section, we present a collection of recent datasets on multimodal numerical reasoning, ranging from geometric problem-solving to general numerical reasoning questions, covering fields such as natural science, social science and so on.We have summarized the specific information of the datasets in Table 1.The tasks of multimodal numerical reasoning are mainly divided into scientific visual reasoning question answering and geometric problem solving. We will explain and categorize the collected datasets into the two tasks mentioned above.The evaluation metrics also vary depending on the characteristics of different datasets and their corresponding solutions. Therefore, we summarized the commonly used evaluation metrics on these datasets.

### 2.1. Geometric problem datasets

Due to its data characteristics and strong correlation between text and graphics, geometry problem can serve as one of the benchmarks for multimodal numerical reasoning. Although automatic geometric problem solving has been a long-standing benchmark in the AI field, there are few suitable datasets available due to the complexity and diversity of the information contained in geometric problems.The format of a typical geometry dataset is shown in Figure 1.

Early on, there were small-scale datasets that relied heavily on manual annotation to assist geometric problem-solving models [52] and GEOS [51] is a dataset of SAT plane geometry questions with 186 questions. Based on GEOS, the GEOS+ [47] add some more entities, functions and predicates with 1406 questions.Then the Geometry3K [33] was proposed, which not only increased the quantity of data but also enriched the geometric problem types in terms of geometric shapes and variable operators. In the same time, a larger and more diverse dataset, GeoQA [7], included clear annotations of the problem-solving process. It improved the universality and interpretability of multimodal numerical reasoning. Later on, Cao [3] annotate 2,518 geometric problems with richer types and greater difficulty to form an augmented benchmark dataset GeoQA+, containing 7,528 questions totally. UniGeo [6], based on the calculation problems in GeoQA, collected and supplemented 9,543 geometric proving problems, defined 37 reason proof explanations, multiple operators and geometry elements, using concise annotations to analyze the proof process. As an expansion of Geometry3K, PGDP5K [65] have more complex layouts such as multiple classes of primitives and complicated primitive relations. Newly, [66] build a new largescale GPS dataset called PGPS9K labeled both fine-grained diagram annotation and interpretable solution program. It is the largest and the most complete annotation dataset for GPS up to now.

### 2.2. Scientific visual reasoning question answering

Visual Question Answering (VQA) [1]is one of the most widely studied visual language tasks, and most visual reasoning tasks are formulated as VQA tasks, aiming to evaluate the specific reasoning abilities of visual language models. Many works have been done to study and summarize it [15]. If the image in visual question answering has numerical relationships or requires numerical calculations combined with text explicitly or implicitly, we can also include it in the benchmark of multimodal numerical reasoning. Currently, visual question answering datasets related to multimodal numerical reasoning are all included datasets with broader domains, and there is no separate visual question answering dataset specifically for numerical reasoning. The general format of scientific question answering and rea-

| | #Q | #I | Grades | Subject | Evaluation Metric | Data Composition | An | In |
|---|---|---|---|---|---|---|---|---|
| GEOS | 186 | 186 | 6-10 | Math(geometry) | Acc,Pre,Rec,F1 | t,d,c,a | ✗ | ✓ |
| GEOS++ | 1406 | 1406 | 6-10 | Math(geometry) | Pre,Rec,F1,NMI | t,d,c,a | ✗ | ✓ |
| Geometry3K | 3002 | 2342 | 6-12 | Math(geometry) | Acc | t,d,c,l,a | ✗ | ✓ |
| GeoQA | 4998 | 4998 | 6-12 | Math(geometry) | Acc | t, d, c, a, e, pt, k, p | ✓ | ✗ |
| GeoQA+ | 7528 | 7528 | 6-12 | Math(geometry) | Acc | t, d, c, k, a, e, p | ✓ | ✗ |
| Unigeo | 9543 | 9543 | 6-12 | Math(geometry) | Acc-top1,10 | t, d, c, a, e, pt, k, p | ✓ | ✗ |
| PGDP5K | 5000 | 5000 | 6-12 | Math(geometry) | Pre,Rec,F1 | t,d,c,l,a | ✗ | ✓ |
| PGPS9K | 9022 | 4000 | 6-12 | Math(geometry) | Acc | t,d,c,l,p | ✓ | ✓ |
| Nuts&Bolts | 4941 | 1019 | 10-12 | Physics | Jaccard Similarity,F1 | t,d,l,a | ✓ | ✓ |
| IconQA | 107439 | 96817 | Pre-K-3 | Math | Acc | t,d,c,a | ✗ | ✗ |
| ScienceQA | 21208 | 10332 | 1-12 | Na, So, La | Acc/BLEU/ROUGE | t,d,c,a,e | ✓ | ✗ |

Table 1. Datasets concerning multimodal numerical reasoning. #Q: number of questions, #I: number of images, Data Composition: the main components of data (t: question text, d: diagram, c: choices, l: logic_form(translate text or diagram into formatted text), a: answer, e: solving explanation, pt: problem type, k: knowledge points, p: annotated programs), An: annotation (annotate the problem-solving process with standardized language), In: interpretation (interpret the charts and text into formal language). In ScienceQA, Na, So, La means Natural, Social, Language Science, respectively.

soning datasets is shown in Figure 2.

The ScienceQA [34] proposed in has more diverse domains with corresponding lectures and explanations, covering 3 subjects, 26 topics, 127 categories, and 379 skills, including physics, chemistry, geography, measurement, and other fields involving simple or complex numerical reasoning skills. It is designed to study the understanding of chain of thought (CoT) under different modal information. In addition, there is a dataset called IconQA [37], which focuses on the inference and understanding of abstract graphs with rich semantics. The questions involve 13 skills, and several skills such as geometry, algebra, measurement, and probability require the model to have some numerical reasoning ability. Besides, Nuts&Bolts [48] is a dataset of physics questions taken from three popular pre-university physics textbooks with 4941 questions and annotated ground truth logical forms for most data. But this dataset is not open source.

### 2.3. Evaluations

Since most of the datasets are in the form of multiple-choice questions, most datasets adopt answer accuracy as the evaluation metric, such as GeoQA [7], GeoQA+ [3],Geometry3K [33], It is worth noting that in the case where Inter-GPS [33] fails to output the numerical value of the problem objective within the allowed steps, it will randomly select one from the four candidates. UniGeo [6] follows IsarStep [29] and adopts top-1 accuracy and top-10 accuracy to evaluate proofs. PGPS9K [66] evaluates performance based on the accuracy of geometric solvers at two levels: numerical answer and solution program. At each level, there are three evaluation patterns: completion, choice, and top-3.

Furthermore, GeoS++ [47] utilizes two commonly used machine translation evaluation metrics: METEOR [11] and MAXSIM [4], and incorporates the evaluation scores as features. While METEOR computes n-gram overlaps with precision and recall control, MAXSIM performs bipartite graph matching and maps each word in one axiom to at most one word in the other. Additionally, they employ Rouge-S [30] as a text summarization metric. On the other hand, PGDP5K [65] has made more nuanced distinctions in their evaluation metrics. For geometric primitive extraction, there are two types of evaluations: parsing position evaluation for the Hough transform approach, and mask evaluation designed for the instance segmentation approach. Regarding relation parsing, they divide a multivariate relation into multiple binary relations, and evaluate the precision, recall, and F1 of binary relation terms. The results are evaluated on four indicators based on the F1 score.

For scientific visual reasoning question answering datasets, Nuts&Bolts [48] use Jaccard similarity [27]and F1 score. In ScienceQA [34], they use accuracy metrics for multi-class classification problem with multiple options. For generated lectures and explanations, they use automatic metrics Bleu [39],Rouge [30],Sentence-bert [46], and human scores. In IconQA [37], the metric of Top-5 accuracy is used to evaluate.

## 3. Deep Learning Models for Multimodal Numerical Reasoning

Here we review earlier work on deep learning methods associated with numerical or mathematical reasoning.

### 3.1. Exploring reasoning ability

As early as the 2000s, machine learning and mathematical reasoning were explored by researchers. Stefan Schultz was the first to use machine learning to standardize the search process [50]. By the 2010s, the rapid development of big data and computer hardware makes solvers related to numerical reasoning play a role in education, finance, medical and other fields. The end-to-end deep learning model has been successfully applied to multimodal mathematical reasoning tasks [22, 54].

Multimodal reasoning tasks are modeled as visual question answering(VQA) tasks. Multimodal reasoning and fusion are the core components of current VQA models. A simple network architecture is used to deal with images of linear algebraic equations and a natural language problem based on variables in the equation. This system provides a foundation for visual understanding, recognition of numbers, variables, operators, and conceptual understanding of coefficients, constants, and variables, as well as including higher levels of comprehension. Mathematical reasoning is formalized as a sequence generation task using common encoders, as well as decoders, which handle the representation and interaction of images, questions, and answers.Visual encoding is typically performed using CNN, ResNet, or Faster-RCNN, while text representation is obtained using GPU or LSTM. Multimodal fusion models, such as BAN [24], FiLM [41], and DAFA [16], are employed to learn a joint representation from different modalities.

### 3.2. Neural networks for numerical reasoning

Similar to early versions of these simple network architectures, these models have many limitations. They lack interpretability and are still insufficient for reasoning and fusing multimodal clues. Furthermore, in multi-modal reasoning tasks modeled as VQA, abstract diagrams, such as IconQA [37], which contain more semantic information related to different visual reasoning skills than natural images do in real-world scenes. Such datasets and their baseline models, such as Patch-TRM [37], have demonstrated superior performance on reasoning tasks related to numerical skills. By utilizing large-scale corpora and the Transformer model [9, 12, 25, 32], pre-trained language models have shown great potential in solving various mathematical problems.

However, these pre-trained language models or visual models based on ViT architecture [12,37,63] are not specifically trained on mathematical data or diagrams.Compared to natural language tasks, they tend to perform relatively worse on tasks related to numerical reasoning. In human cognition, a single sensory input is often insufficient to support high-level reasoning and decision-making. Instead, it is necessary to combine multiple sensory inputs to perform reasoning. Similarly, multimodal reasoning abilities must be built upon unimodal reasoning abilities. In previous studies by scholars in related fields on math word problems(MWPs) [8], it can be observed that NLP models have limitations in understanding mathematical or scientific data, such as shallow heuristics [40].

### 3.3. Reasoning methods for geometry problems

One common way to assess the capacity of deep learning models for advanced multi-modal reasoning is through the use of geometry problem solving as a testing ground. The method of automatically solving geometric problems can be traced back to a stage that was highly dependent on a limited set of manually crafted rules [49, 51], and its solving performance was not ideal. In recent years, some methods for solving geometric problems have been proposed to promote research in this field. An interpretable framework for solving geometric problems has been proposed [33], which is characterized by a symbolic geometric solver that parses diagrams and texts into a unified formal language [33, 51, 65] and iteratively updates the condition status through symbolic reasoning until the target is searched. Chen et al.propose Neural Geometric Solver (NGS) [7] to address geometric problems by jointly understanding text, diagram, and then generating explainable programs.Then also present a unified multitask Geometric Transformer framework, Geoformer [6], to tackle calculation and proving problems simultaneously in the form of sequence. Cao and Xiao proposed a dual-parallel text encoding method, DPE-NGS, based on the issue that NGS lacks the ability to extract features from long texts [3]. It should be noted that existing large pre-trained language models (LMs) can encode a vast amount of linguistic information, but advanced reasoning skills such as numerical reasoning are difficult to learn solely from language modeling objectives [17],and often requires additional fine-tuning and modification for optimal performance. Pre-training methods proposed by relevant researchers, such as MEP [6] and the inferable number prediction task [42], as well as skill injection methods [17], have significantly improved the ability of pre-trained language models to generalize to mathematical problems.

Currently, there is still a significant performance gap in solving geometry problems using neural solvers [3,6,7,22]. The main reason is the inadequate representation of diagrams and the difficulty of cross-modal fusion. Existing neural solvers adopt a multimodal framework similar to processing natural images. However, in geometric diagrams, the spatial relationships are complex, and the primitives are thin and overlapping, making it difficult to extract fine-grained features, and even disrupting the structure and semantic information.PGPSNet [66]utilizes CNN to learn the coarse-grained global features of diagrams, and en-

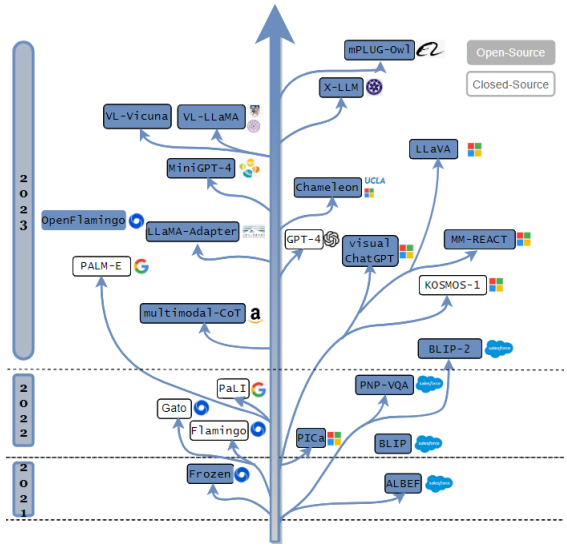



Figure 2. A timeline of representative multimodal large models in the past three years, primarily based on the publication dates of their technical papers.

codes sub-sentences of text to describe finer-grained structures and semantic information within the diagrams. This integrated approach further enhances the performance of geometric solvers.The analysis of geometric diagrams and cross-modal fusion will remain a challenging yet promising area for exploration.

# 4. LLMs for Numerical Reasoning

Language ability does not equal to ”thinking” or ”reasoning” in LLMs. One of the long-term goals of artificial intelligence is to develop machines with the ability to reason mathematically. The N2Formal team at Google Research Center, for instance, envisions creating an automated mathematician [43]. Large Language Models(LLMs) has already beaten a lot of expectations as to what is possible in automated Numerical reasoning and have recently revolutionized the field of natural language processing(NLP).

## 4.1. Reasoning ability of LLMs

An ability is emergent if it is not present in smaller models but is present in larger models [57]. In mathematical problem-solving, logical reasoning, and mathematical symbol manipulation, emergent abilities often appear suddenly and unpredictably at a particular model size. But the model size highly correlated with the performance of larger models is not the only scale to measure emergent abilities. Emergent abilities are an interesting and important phenomenon. However, existing LLMs have weaknesses in complex reasoning. For example, Hendrycks et al. have shown that pre-training GPT series models with smaller parameter sizes on mathematical corpora results in higher accuracy than GPT-3 175B without pre-training [19]. Bang et al. conducted a technical evaluation of ChatGPT's strengths and limitations in reasoning across 10 different categories, including numerical reasoning. They identified weaknesses in its ability to perform inductive reasoning, mathematical reasoning, and multi-hop reasoning, among others [2].

## 4.2. LLMs prompting methods

Especially when the language model is large enough (>100B parameters), it exhibits emergent abilities to perform complex multi-step reasoning with only a few examples provided [56, 58]. By combining some advanced prompting strategies, the understanding ability of LLMs can be further improved.Different prompting strategies, such as zero-shot, few-shot, and CoT [58], have been successfully applied to numerical reasoning tasks. The reasoning ability of LLMs is influenced by the complexity of the prompt, and extending the idea of self-consistency to complexity consistency shows that complex prompts achieve better performance than simple prompts [14]. MathPrompter utilizes zero-shot prompting technology to generate multiple algebraic expressions or Python functions that solve the same mathematical problem in different ways, improving the confidence level of the output results [21]. Prompting examples for more complex problems involving heterogeneous information, such as mathematical reasoning with tabular data, have also been proposed [36]. Progressive-Hint Prompting (PHP) [68] has proposed a new approach for LLMs to reconsider problems, rather than just focusing on the hand-designed prompts for the problems and answers. This approach has achieved significant performance improvements in mathematical reasoning tasks and has outperformed state-of-the-art results on multiple reasoning benchmarks. As LLMs have developed, prompts have been able to achieve performance comparable to or even better than full fine-tuning on the training set [14, 26, 28].

## 4.3. Multimodal numerical reasoning

Many researchers have explored various methods of interacting with LMs to input or output data in multiple modes, as displayed in Figure2. For example,the initial and most primitive way was to use programming code as an intermediate medium between visual and linguistic representations [10, 45]. However, this approach can only generate symbolic images, and its quality cannot compare to the images generated by modern text-to-image models.The current main methods involve prompting language models (LLMs), fine-tuning small-scale models, or training a visual module, connecting the language model with Visual Prompt Generator(VPG). The most direct way to utilize LLMs for reasoning is to convert multimodal information input into a single modality, namely textual language.Like X-LLM [5],

| Format(prompt) | Method | Dataset | Total(%) | Angle(%) | Length(%) | Other(%) |
|---|---|---|---|---|---|---|
| W/O Program | GPT-3.5 Turbo | GeoQA | 30.27 | 31.75 | 24.07 | 10.71 |
| W/O Diagram | | | 35.35 | 40.24 | 31.40 | 32.76 |
| Text-Diagram | | | 36.22 | 41.57 | 33.62 | 23.91 |
| Text-Only | | LīLA | 30.17(IID)↑1.87 \| 20.62(OOD)↑8.62 | - | - | - |

Table 2. Results on the GeoQA and LīLA dataset by GPT-3.5 Turbo.It should be noted that GeoQA uses accuracy as its evaluation metric, while LīLA uses F1 score as its evaluation metric.And the results of LīLA are compared with the results obtained using GPT-3 in the original paper. Besides, we evaluate both in-distribution (IID) and out-of-distribution (OOD) performance for datasets in LīLA.

X2L interface is utilized to convert multimodal inputs (images, speech, videos) into foreign languages and input them into a large language model (ChatGLM). X-LLM aligns multiple frozen unimodal encoders and a frozen LLM using the X2L interface.

LLMs have sparked a transformation in the field of multimodal understanding, shifting from traditional pre-trained visual language models (VLM) to large language model-based visual language models (VL-LLM). By integrating a visual module into LLM, VL-LLM can inherit the existing knowledge, zero-shot generalization ability, reasoning capability, and planning ability of LLM. Extracting visual features and image captions using a captioning model or frozen pre-trained image encoders and then combining the captions with the original information to input into LLMs [18, 20, 34, 60, 61]. A two-stage framework called multimodal-CoT [67] has been proposed to address the significant information loss and lack of spatial coordination during the captioning process. In this framework, the T5 model [44] is fine-tuned to integrate visual and language representations for performing multimodal CoT. The PNP-VQA approach [55], aiming to address the issue of missing image information in PICa [60], incorporates an Image-Question Matching module. This module selects relevant patches from the image that are most related to the question. Captions are generated specifically for these patches, which are then fed into UnifiedQAv2 as context. This process ensures that the generated captions are question-specific. Moreover, UnifiedQAv2 [23] is utilized for PLM selection, enabling zero-shot VQA capability. Visual ChatGPT [59] extends the advantages and potential of multimodal-CoT in a specific domain, such as ScienceQA [34], to a wide range of tasks. It is based on specific principles that govern how ChatGPT outputs calls and how to invoke the Visual Foundation Models (VFMs). This includes defining the input and output formats. Subsequently, an executor parses and executes the instructions, returning the results back to ChatGPT. Additionally, HuggingGPT [53] combines hundreds or even thousands of models from HuggingFace and GPT, allowing it to tackle 24 different tasks.

The general multimodal reasoning model does indeed focus on the breadth of tasks, but in the domain of numerical reasoning, it is important to pay more attention to accuracy and precision.Chameleon [35] enhances LLMs to help address inherent limitations such as the inability to access up-to-date information, utilize external tools, or perform precise. mathematical reasoning.With the introduction of the plugin feature in GPT-4, previous constraints have been overcome. Furthermore, models like MiniGPT-4 [69], which freeze the visual and language modules during pre-training and instruction fine-tuning stages while adjusting a limited number of parameters, or models like Kosmos-1 [20], which freeze the visual module and train the language module, or models like LLaVA [31], which freeze the visual module during instruction fine-tuning and train the language module, all limit the alignment between different modalities and lack joint training on single-modal and multi-modal data, making it difficult to effectively unleash the various potentials of large-scale models. In contrast, mPLUG-Owl [62] trains the visual module using multi-modal data while freezing the language module, allowing the visual features to align with the language features. However, the initial pre-training of the visual module and the loading of the base LLM incur significant computational costs. VPGTrans framework [64] can greatly reduce the computational overhead and required training data for training VL-LLM, achieving comparable or better results with just around 10% of the data and computation time. This provides greater possibilities for most researchers to customize their own VL-LLM.

### 4.4. Experimental setting

We tested the ability of GPT-3.5 Turbo to solve geometry problems using some data from GeoQA and LīLA [38] . Due to interface limitations, we input samples containing images into the model in a language form that the model could understand through designed prompts.

Firstly, an Inter-GPS diagram parser is used to extract the structural and semantic information of certain images in GeoQA, which serves as additional context for paired question text. Considering the ChatGPT API's good support for LaTeX interpretation, the extracted information is further converted to LaTeX format and included as part of the prompts. Additionally, the problem-solving process and so-

| Classification of Question Types | Number of Question Types | Acc(%) IID | Acc(%) OOD |
|---|---|---|---|
| Comparison Calculation Problem | 129 | 15.50 | 12.40 |
| Coordinate Calculation Problem | 27 | 18.52 | 7.40 |
| Length Calculation Problem | 300 | 25 | 9.67 |
| Area Calculation Problem | 538 | 18.77 | 8.56 |
| Volume Calculation Problem | 78 | 25.64 | 7.69 |
| Others | 312 | 22.76 | 13.14 |

Table 3. Classification of Question Types and Accuracy in the LīLA dataset

lution programs from the original dataset are incorporated as the thinking chain and answer references, respectively, in the prompt examples. The correctness of multiple-choice questions in the context of few-shot learning can be tested by conducting In-context Learning tests. In this approach, the model is provided with a small number of examples (e.g., two) for each task or question type, allowing it to learn

**Problem**: Given that $\angle 1 = \angle B$ and $\angle 2 = N_0 = 25^\circ$, we need to find the measure of $\angle D$.
**Description of Diagram**:
A diagram with $6$ vertices has the coordinates $\\left(\n\\begin{array}{cc}\n 122.66666666666666 & 22.0 \\\\\n 43.84615384615385 & 22.0 \\\\\n 70.125 & 45.5 \\\\\n 5.5 & 42.0 \\\\\n 86.0 & 0.0 \\\\\n 139.0 & 45.0 \\\\\n\\end{array}\n\\right)$. Each row of the matrix represents the coordinates of D, A, C, B, E and F. And the diagram with $8$ lines [\overline{\rm DA},\overline{\rm DC},\overline{\rm AB},\overline{\rm AE},\overline{\rm CB},\overline{\rm CF},\overline{\rm BE},\overline{\rm BF}].
**Additional conditions**:
\begin{aligned} & \text{m}\angle DCF = \text{m}\angle DAE \ & \text{m}\angle EAD = \text{m}\angle 1 \ & \text{m}\angle DCF = \text{m}\angle 2 \ & A \in \text{Line}(B,E) \ & C \in \text{Line}(B,F) \end{aligned}
**Choices**: ['$25^\circ$', '$45^\circ$', '$50^\circ$', '$65^\circ$']
**Constant**: [$30(C_0)$, $ 60(C_1)$, $90(C_2)$, $180(C_3)$, $ 360(C_4)$,, $\pi(C_5)$, $0.618(C_6)$]
**Solution**:
Since $\angle 1=\angle B$(from the Problem), we have $AD\parallel BC$ (line{AD} and line{BC} from Description of Diagram). Therefore, $\angle D=\angle 2=25^\circ$. Therefore, the answer is $\boxed{25^\circ}$. Therefore choose A.
**Program**:
The sequence is obtained according to Solution: ['g_equal', 'N_0']
**Final answer**: The final answer is A. I hope it is correct.

Figure 3. An example of a prompt about a GeoQA sample.

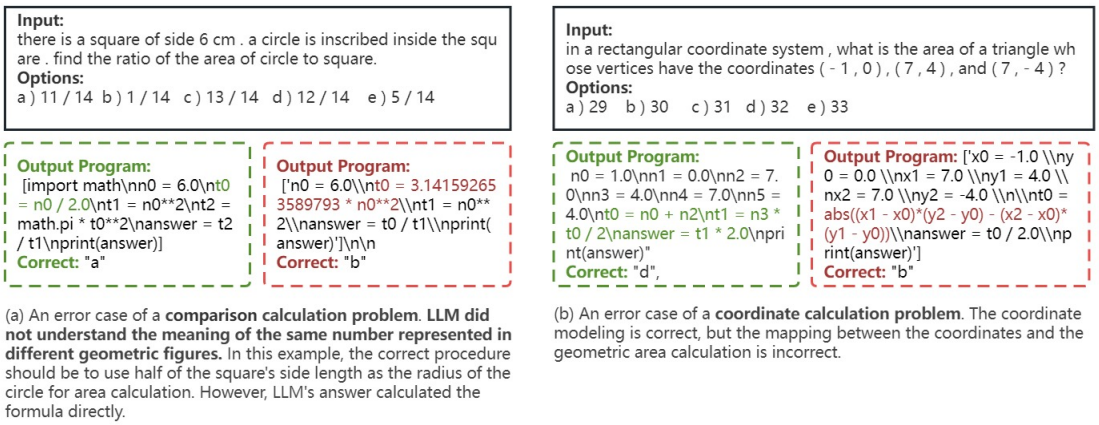
the patterns and context necessary for accurate answering. By evaluating the model's performance on a separate set of single-choice questions, we can assess its ability to generalize and provide correct answers in similar contexts.For the geometry questions in the LīLA dataset, randomly select three examples using both the IID and OOD strategies (3-shot). Additionally, ablation experiments were conducted on the prompts in the GeoQA dataset,as shown in the Figure3.

Our experimental results on LīLA show overall improvements compared to GPT3, but there is not much difference on Geoqa compared to earlier old methods. The main reason for this could be the uncertainty caused by random selection, leading to an increase in test variance. ChatGPT has high instability in selecting this type of contextual examples. Using policy gradient descent [36] may lead to better results. Another reason is to consider how to effectively convert visual information into language that ChatGPT can understand.In comparison to the short problem texts of Geometry3K, GeoQA's question text contains all the variable information required in the problem-solving program. Therefore, the inclusion of image information does not provide a significant advantage.

**Result Analysis.** The results of our experiment are shown in Table 2. Furthermore, we classify the data in the LīLA dataset into the following types and calculate their accuracy in Table 3:

- **Comparison calculation problems**, refer to a type of geometry problems that involve comparing the shapes and sizes of two or more geometric figures to solve for an unknown value.
- **Coordinate calculation problems**, refer to the type of geometry problems that require calculating coordinates in two or three dimensions to solve problems related to the area, perimeter, distance, and other properties of geometric figures.
- **Length calculation problems**, refer to the type of geometry problems that require calculating the length of geometric figures such as line segments, polylines, and polygons to solve problems.
- **Area calculation problems**, refer to the type of geometry problems that require calculating the area of geometric figures to solve problems.
- **Volume calculation problems**, refer to the type of geometry problems that require calculating the volume of geometric figures to solve problems.

**Error Analysis.** Even though we employed a 3-shot prompt scheme, the performance of GPT-3.5 Turbo is still relatively weak due to the need for improvement in its mathematical reasoning ability. Therefore, there are not many questions that can be correctly answered. To facilitate further research, we selected one incorrect question from each of the two types of questions: comparative calculation and coordinate calculation. Please refer to the failure cases in Figure 4.The failure cases can be divided into two types: a) The model fails to understand the different meanings of numerical values in geometric problems and applies them to different formulas. b)The model cannot perform calculations correctly by combining geometric space with graphical formulas.




(a) An error case of a **comparison calculation problem**. **LLM did not understand the meaning of the same number represented in different geometric figures.** In this example, the correct procedure should be to use half of the square's side length as the radius of the circle for area calculation. However, LLM's answer calculated the formula directly.

(b) An error case of a **coordinate calculation problem**. The coordinate modeling is correct, but the mapping between the coordinates and the geometric area calculation is incorrect.

Figure 4. Two examples with predictions from GPT-3.5 Turbo.

# 5. Future Work

We identified the limitations of existing work, which served as inspiration for our reflections on future research directions. There is significant room for exploration in the field of multimodal numerical reasoning.

## 5.1. Specialized datasets

Although we listed some datasets related to multimodal numerical reasoning in Chapter 2, most of them are focused on geometry calculations in mathematics and some general science-related question-answering tasks involving numerical reasoning [3, 6, 7, 33, 47, 51, 52, 65, 66]. Geometry calculation tasks are relatively clear examples of multimodal numerical reasoning related datasets. However, in the fields of science and engineering, there are also many applications that require the integration of multimodal information to infer the final result, such as geographic surveying, physical motion distance-speed information, chemical formula formulations, mathematical reasoning with image data, etc. However, on datasets such as ScienceQA [34] and IconQA [37], most questions are still focused on general science question-answering tasks related to natural images, and there are only a small number of questions that involve in-depth numerical reasoning. Therefore, there is still a lack of specialized datasets for science and engineering fields in the research of multimodal numerical reasoning. In future research, collecting and annotating datasets specifically for these fields can help promote the training of models and exploration of their capabilities. In addition, finer-grained classification and analysis of the possible numerical capabilities in the data set is also necessary to explore the problem-solving ability of the model, such as the analysis of numerical certainty, and the analysis of the consistent ability of images and text.

## 5.2. More effective prompts

Considering the limitations of current models in problem-solving, it is possible to design more appropriate and effective prompts.Prompt engineering is becoming more and more important in the era of large models, especially in mining the emerging capabilities of large language models. The thought chain prompt method shows good performance, and we have also made some discussions in the paper, such as [61] using the prompt method call tool to mine and expand the multimodal reasoning ability of LLMs. In our final experiments, we also adopted fixed-designed template hints to guide LLMs to generate answers in the format we need. Sometimes an effective prompt may come from a sentence [26], or it may be a well-designed template. In future work, we can think more about the techniques of prompt engineering and analyze the characteristics of multimodal numerical reasoning. For example, the prompt examples given can be how images and texts express the human way of thinking consistently.

## 5.3. Instruction fine-tuning applied to multi-modal problem solving

Researchers explored the method of LLM instruction tuning to make LLM follow natural language instructions and complete real-world tasks, and the results showed that its zero-shot generalization ability was significantly improved. In addition to its applications in NLP tasks, this method can also be applied to the multi-modal field of computer vision [31]. Futhermore, in future work on multimodal numerical reasoning, instruction fine-tuning can be considered by combining existing multi-modal reasoning datasets for manual annotation or using LLM to generate some multi-modal reasoning instructions to follow the data, and then selecting some evaluation metrics or scoring models for multi-modal reasoning. Instruction fine-tuning can

be used to fine-tune LLM to make it more suitable for multimodal reasoning tasks.

### 5.4. Generating questions

In the field of multimodal numerical reasoning, generating questions is a viable direction for education, testing, and data analysis. It can generate suitable questions that involve multimodal numerical reasoning skills, such as mathematical and geometric problems. This can be helpful for dataset expansion and model inference capability improvement.But the difficulty levels of generating questions and the reliability of generalizing to a wider range of numerical values still need to be examined. And there are some challenges that need to be addressed: One of the challenges is the difficulty level of the generated questions, which should be appropriate for different target audiences and maintain consistent difficulty levels, another challenge is the semantic consistency of the multimodal information in the generated questions. Generating suitable images relies on specific methods, and abstract image generation is different from natural image generation. Currently, there is limited research in this area, and maintaining consistency between different modalities is also a research issue that needs to be considered.

### 5.5. Complex abstract graphs processing

There is room for improvement in the fine-grained processing and modal fusion capabilities when dealing with complex abstract graphs.In multimodal numerical reasoning, the data images are often abstract images, which require analysis and processing of multiple layers of information and details. Current models, whether neural network models for image processing or large language models, still lack the ability to handle abstract images effectively. In Patch-TRM [37], experiments on the abstract dataset of IconQA showed good performance, but there is still a need to develop new methods for processing the information in mathematical and geometric images. Additionally, the ability to align abstract images with numerical text also requires improvement in model capabilities.

### 5.6. Enhance LLMs' reasoning abilities

Lastly, injecting domain-specific knowledge from existing research areas into LLMs to enhance their reasoning abilities, rather than solely relying on external tools for assisted computation, or exploring untapped potentials when conditions permit, could yield significant advancements.LLMs have shown excellent performance in language reasoning expression, and the emergence of generative capabilities provides possibilities for exploring the capabilities of large language models. However, given the current limitations of large language models in multimodal numerical reasoning, many studies have attempted to call upon external tools for assistance, such as the recent collaboration with Wolfram, which has impressed many researchers. However, given the complexity and computational cost of these external tools, it is still necessary to enhance the reasoning and multimodal processing capabilities of LLMs themselves. This requires injecting more domain-specific knowledge into the models to better integrate and process this information.

## 6. Conclusion

In this paper, we conducted a comprehensive survey into multimodal numerical reasoning. We reviewed various datasets and their evaluation criteria that have been employed, and discussed the methods that have been employed, including early symbolic solvers, subsequent neural networks, and more recently, large language models.We also identified deficiencies and limitations in the existing datasets and methods for multimodal numerical reasoning. Finally, we outlined directions for future research and emphasized the potential for further exploration in this field. Through our investigation and summary of multimodal numerical reasoning, we hope to provide interested researchers and practitioners with useful information and inspiration.

## References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[2] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 1, 5

[3] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, 2022. 2, 3, 4, 8

[4] Yee Seng Chan and Hwee Tou Ng. Maxsim: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, 2008. 3

[5] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 5

[6] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 2, 3, 4, 8

[7] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric

question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021. 2, 3, 4, 8

[8] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022. 1, 4

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 4

[10] Denis. Drawing mona lisa with chatgpt. 5

[11] Michael Denkowski and Alon Lavie. Extending the meteor machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, 2010. 3

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[13] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023. 1

[14] Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022. 5

[15] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022. 2

[16] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648, 2019. 4

[17] Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models. *arXiv preprint arXiv:2004.04487*, 2020. 4

[18] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*, 2022. 6

[19] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 5

[20] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 6

[21] Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023. 5

[22] Pengpeng Jian, Fucheng Guo, Yanli Wang, and Yang Li. Solving geometry problems via feature learning and contrastive learning of multimodal data. *CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES*, 136(2):1707–1728, 2023. 4

[23] Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*, 2022. 6

[24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018. 4

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 4

[26] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022. 5, 8

[27] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971. 3

[28] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022. 5

[29] Wenda Li, Lei Yu, Yuhuai Wu, and Lawrence C Paulson. Isarstep: a benchmark for high-level mathematical reasoning. *arXiv preprint arXiv:2006.09265*, 2020. 3

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 3

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 6, 8

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 4

[33] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 2, 3, 4, 8

[34] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2, 3, 6, 8

[35] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023. 6

[36] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 5, 7

[37] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 3, 4, 8, 9

[38] Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*, 2022. 2, 6

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 3

[40] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021. 4

[41] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 4

[42] Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. Improving the numerical reasoning skills of pretrained language models. *arXiv preprint arXiv:2205.06733*, 2022. 4

[43] Markus N Rabe and Christian Szegedy. Towards the automatic mathematician. In *Automated Deduction–CADE 28: 28th International Conference on Automated Deduction, Virtual Event, July 12–15, 2021, Proceedings 28*, pages 25–37. Springer International Publishing, 2021. 5

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 6

[45] Fabin Rasheed. Gpt3 sees, 202. 5

[46] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[47] Mrinmaya Sachan, Kumar Dubey, and Eric Xing. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 773–784, 2017. 2, 3, 8

[48] Mrinmaya Sachan, Kumar Avinava Dubey, Tom M Mitchell, Dan Roth, and Eric P Xing. Learning pipelines with limited data and domain knowledge: A study in parsing physics problems. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[49] Mrinmaya Sachan and Eric Xing. Learning to solve geometry problems from natural language demonstrations in textbooks. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 251–261, 2017. 4

[50] Stephan Schulz. Learning search control knowledge for equational theorem proving. *Lecture notes in computer science*, pages 320–334, 2001. 4

[51] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015. 2, 4, 8

[52] Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, and Oren Etzioni. Diagram understanding in geometry questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014. 2, 8

[53] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 2, 6

[54] A. Sinha and K. Ayush. Towards mathematical reasoning: A multimodal deep learning approach. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018. 1, 2, 4

[55] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 6

[56] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 5

[57] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1, 5

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 5

[59] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1, 6

[60] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022. 6

[61] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 6, 8

[62] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 6

[63] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 4

[64] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023. 6

[65] Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. Plane geometry diagram parsing. *arXiv preprint arXiv:2205.09363*, 2022. 2, 3, 4, 8

[66] Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097*, 2023. 2, 3, 4, 8

[67] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1, 6

[68] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*, 2023. 5

[69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 6