From Text to Visual Images: A Scalable Multimodal LLM for Math Question Answering over Heterogeneous Data

Angel Crawford
Computer Science Department
Brigham Young University
Provo, Utah 84604
angelwhwrt@gmail.com

Yiu-Kai Ng
Computer Science Department
Brigham Young University
Provo, Utah 84604
ng@compsci.byu.edu

***Abstract*—Mathematical question answering (Math QA) poses a significant challenge for multimedia information retrieval due to the heterogeneous nature of mathematical content, including LaTeX expressions, images, diagrams, and natural language text. We present a multimodal LLM-based Math QA framework that unifies textual and visual mathematical representations within a scalable indexing and reasoning pipeline. Built on Llemma 7B, a math-specialized LLM, the framework employs a multimodal encoder-decoder architecture with cross-modal alignment to support coherent retrieval and reasoning across text, equations, and diagrams. To evaluate effectiveness at scale, we conduct systematic experiments on large Math QA benchmarks that include both textual and visual mathematical inputs, assessing formula interpretation, answer relevance, and reasoning accuracy. Results demonstrate state-of-the-art performance in multimodal math understanding and consistent gains over text-only baselines, highlighting the value of cross-modal evaluation for Math QA in multimedia retrieval settings.**

***Index Terms*—Multimodal LLMs, math QA, heterogeneous data, visual-textual math representation**

I. Introduction

Math is foundational to many areas of study, enhancing problem-solving abilities and providing skills transferable across subjects and professional roles. It serves as a universal language, conveying quantitative relationships and processes, and is integral to everyday activities such as scheduling, cooking, financial planning, measurements, and organization. Despite its importance, math proficiency has declined worldwide, particularly in the United States. According to the National Assessment of Educational Progress (NAEP) [5], 12th graders in the USA are considered proficient with a score of 176 or higher, yet the 2019 average was only 150. Similarly, only 38% of US public school students demonstrated math proficiency in 2023 [12], and global PISA scores fell from 489 in 2018 to 472 in 2022 [6]. Compounding this challenge, locating reliable math resources for learning or research remains difficult, especially for those unfamiliar with the subject area [13]. These trends highlight the need for systems that can efficiently retrieve, extract, and interpret math resources from large, heterogeneous datasets containing both textual and visual math notations captured in images.

To address these challenges in practice, we focus on leveraging large language models (LLMs) for math question answering that incorporates both textual and visual content, including formulas embedded in text or presented as isolated images. Our goal is to extract accurate archived answers from such questions while maximizing the utility of large-scale heterogeneous datasets. The system emphasizes flexible and scalable retrieval from visually structured math content and incorporates self-verification of answers, ensuring reliability and consistency across diverse input types.

LLMs have historically struggled with mathematical content, often producing hallucinations or incomplete reasoning due to limited understanding of the steps and background knowledge required in math, science, and programming. Recent math-specialized LLMs improve reasoning capabilities [14], but most focus primarily on textual data, and some rely on code execution, leaving visually formatted math content, such as math formulas captured as images, underutilized. This gap underscores the need for a scalable multimodal approach to process heterogeneous inputs effectively.

To address these limitations systematically, we propose a scalable multimodal LLM framework for Math question answering (Math QA) that integrates textual and visual math inputs. The framework features a multimodal encoder-decoder architecture capable of fusing heterogeneous data, along with a cross-modal alignment module to ensure coherent interpretation across text, equations, and visual math. This design enables visual-symbolic formal understanding at scale while preserving the precision of textual math LLMs. Its modular and extensible structure allows efficient retrieval, interpretation, and utilization of math resources in large-scale educational and research repositories.

II. Related Work

Recent progress in math-oriented and multimodal LLMs is reviewed below.

A. Math LLMs

Recent LLMs have advanced mathematical reasoning, including PaLM 2 [2], GPT-4 [1], and Minerva. PaLM 2, trained on diverse textual objectives, excels across downstream tasks. GPT-4's large-scale multimodal design processes text and images, achieving human-level performance on academic and professional benchmarks, including math. Minerva, a math-specialized PaLM variant, leverages scientific papers and web

math data to solve complex quantitative problems with step-by-step reasoning. These models demonstrate the role of scale, targeted pretraining, and domain adaptation in enabling near-human mathematical problem-solving.

Although modern LLMs can solve math problems, they still exhibit common limitations such as overconfidence, hallucinations, and inconsistent explanations [7]. This is particularly problematic for learners, who require not only correct answers but also clear, logical solution steps to build independent problem-solving skills [10]. Existing LLMs also primarily handle textual math, struggling with visual or audio representations, and can fail on niche problems where little prior documentation exists, as they rely on predicting probable outputs rather than possessing a foundational understanding of math.

B. Multimodal LLMs

Several math-focused LLMs have begun incorporating multimodal capabilities. GPT-4 and Kosmos [9] can process text and images, with GPT-4 emphasizing textual understanding and Kosmos prioritizing visual content. Benchmark datasets such as MathVista [8] and GeoEval [16] evaluate visually grounded math tasks, highlighting limitations in current foundation models and assessing performance on geometric problem-solving. However, these datasets are not fully open source, and their access is restricted.

Despite multimodal inputs, LLMs often struggle with non-textual math due to the strict syntactic and semantic structures of mathematical content. To address this, we design a new multimodal architecture that leverages the shared embedding concepts of NExT-GPT [15] but integrates Llemma, a math-specialized LLM, with advanced cross-modal alignment and reasoning modules. This allows the system to process and generate outputs across textual and visual formats, enabling robust interpretation and reasoning over heterogeneous math data beyond the capabilities of existing works.

III. Design Methodology

In this section, we present our multimodal LLM framework for processing mathematical formulas in both textual and visual formats. We detail the system architecture and emphasize the innovative integration strategies and cross-modal reasoning mechanisms that set our approach apart from existing Math QA models and other multimodal LLMs. Figure 1 illustrates the overall architecture of our multimodal math LLM.

A. A Multimodal LLM for Math QA

To enable robust Math QA with multimodal data, we introduce our LLM architecture that seamlessly processes both text and images. The system unifies a *NExT-GPT*-based multimodal framework with Llemma, employing cross-modal reasoning and alignment to handle heterogeneous math inputs beyond existing models.

Our multimodal architecture builds on the high-level design of *NExT-GPT* but introduces a key innovation by replacing the default Vicuna backbone with *Llemma*, an LLM optimized for reasoning over symbolic and textual mathematical

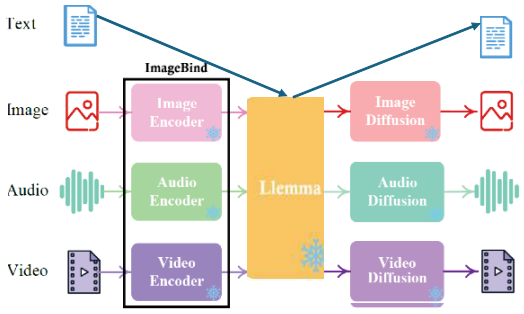

Fig. 1: Proposed multimodal math LLM

content (see Figure 1). The system follows three stages—encoding, LLM understanding and reasoning, and decoding. In the *encoding* stage, ImageBind projects inputs from diverse modalities into a unified embedding space, simplifying cross-modal integration compared to managing multiple independent encoders. A *projection* layer further transforms these embeddings into language-like representations interpretable by the central LLM. Unlike prior architectures that rely on general-purpose models, our proposed multimodal framework leverages Llemma's domain-specific mathematical expertise in the semantic understanding stage, enabling accurate reasoning across text, equations, and visual content. By integrating specialized symbolic reasoning into the broader NExT-GPT framework, the system advances multimodal Math QA beyond general-purpose designs.

The framework unifies heterogeneous data—text, symbolic expressions, and images—into a shared representation space while maintaining scalability and extensibility. Llemma's math specialization ensures precise interpretation of mathematical structures, and the modular architecture supports seamless integration of new modalities, such as images, without retraining the full system, reducing both computational and development overhead. This efficiency-driven design bridges modality gaps, achieves robust performance, and delivers personalized outputs across multiple forms, positioning our approach as an advancement in multimodal LLM research for education, analytics, and scientific discovery.

The central LLM performs semantic reasoning over inputs and generates both textual outputs and modality-specific signal tokens that guide the decoding layers in producing multimodal responses. These signals are projected through lightweight transformer layers into formats compatible with different decoders—text is returned directly, while an off-the-shelf latent diffusion model generates images. Importantly, only the small projection layers ($\sim$1% of parameters) are updated during training, while encoders and decoders remain frozen, enabling efficient adaptation and low-cost scaling across heterogeneous datasets. This design ensures rapid deployment and flexible integration of new data modalities.

1) Alignment Learning for Multimodal Framework: Our multimodal math LLM introduces an alignment framework that bridges modality gaps by updating only selected layers during encoding and decoding, enabling efficient integration of text and images for robust mathematical reasoning. During

encoding, learnable concept tokens aggregate grid-level image features into semantic representations that capture both local and global structure, and convert them into textual tokens compatible with Llemma, unifying multimodal inputs for precise problem solving. Alignment is trained via an image-to-text task on the ImageCaption dataset (CC3M, 3M+ images), providing robust bidirectional mapping between visual and textual content. During decoding, signal tokens (e.g., $[\mathrm{IMG}_i]$) guide pretrained diffusion models by indicating when and where to generate images, while concept-token-encoded images propagate alongside text. The LLaMA-based math LLM encodes textual input and restores it to natural language, with signal tokens specifying image placement and content to ensure accurate text-to-image generation. The framework is modular and extensible, supporting additional modalities, user-selected output formats, and efficient scaling across large, heterogeneous datasets, enabling flexible and high-performance multimodal reasoning.

2) Multimodal Signal Representation Training: To ensure accurate user instruction following and controllable multimodal outputs, we perform Modality-switching Instruction Tuning (MosIT). MosIT trains the LLM on input-output pairs, where inputs are user instructions and outputs represent the desired responses, including modality signal tokens. On the decoding side, signal token representations are aligned with gold multimodal captions via the diffusion condition encoder. This targeted tuning enables precise cross-modal reasoning and flexible multimodal content generation, supporting scalable, robust, and efficient handling of heterogeneous large-scale data.

3) The Llemma 7B Model: Our multimodal system leverages Llemma 7B, a math-specialized LLM based on a decoder-only transformer (Code Llama) pretrained on scientific papers, math-focused web data, and code, and fine-tuned to use Python interpreters and formal theorem provers to verify problem structure and solution logic [11]. Integrated within the NExT-GPT architecture, Llemma processes projected text inputs and produces outputs compatible with multimodal projection layers, solving math problems step by step while generating self-verifying code to detect and correct logical errors [17]. Conceptual tokens enable native handling of numerical values, formulas, and structured expressions, supporting seamless multimodal integration. This design ensures logically consistent reasoning through step-level verification and correction, with multimodal inputs mapped into a shared embedding space for coherent interpretation. Controlled generation via signal tokens and instruction-tuned outputs, together with fine-tuning limited to different layers, preserves model stability while enabling scalable and accurate reasoning across heterogeneous datasets.

Building on these capabilities, the following advances highlight how retraining Llemma 7B within our multimodal framework enables precise reasoning, cross-modal integration, and scalable deployment for large-scale math problem solving:

(i) *Retrained, step-by-step reasoning*: Fine-tuned on domain-specific math data with Python interpreters and theorem provers, Llemma 7B automatically detects and corrects logical errors, outperforming open-source math LLMs (see Section V-B3).

(ii) *Seamless multimodal integration*: Text, formulas, and images are projected into a shared embedding space, with specialized tokens representing numerical values and structured expressions. Aligned representations enable Llemma to process multimodal inputs and generate coherent reasoning across heterogeneous data.

(iii) *Stable, scalable, deployable*: Llemma 7B processes multimodal outputs, with instruction tuning constraining generation for stability. Its modular architecture scales to large datasets and new modalities without full retraining, supporting robust educational and scientific applications.

Together, these key features demonstrate that retraining Llemma 7B within a carefully designed multimodal framework produces a state-of-the-art, scalable, and verifiable system for large-scale mathematical reasoning.

B. The Overall System Process and Applications

The user interface of the proposed multimodal Math QA system is shown in Figure 2, where Llemma is embedded within NExT-GPT to combine the strengths of both models. The workflow begins with a user query, which may include both text and images. The text is fed directly into Llemma, while images are first encoded into text through ImageBind on a patch-wise basis, then projected into conceptual image tokens with signal markers (see Figure 1). This preserves modality information while converting images into a format compatible with text-based LLMs. Llemma then processes the full multimodal question, leveraging code-based solvers and theorem provers to generate mathematically accurate answers. Image tokens in the output are passed to an image projection block along with the text answer, which is further processed by a vision transformer for diffusion-based image generation [15]. The system can thus handle large-scale, mixed-modality math queries, producing integrated text-and-image answers that enhance both precision and interpretability.

With this framework, LLM-based Math QA tools gain significant improvements in flexibility, enabling them to process diverse data types and serve a broader range of learners. The integration of multimodal inputs also expands the knowledge sources available to LLMs, improving mathematical understanding beyond text alone. Training overhead remains minimal because the majority of pretrained backbone weights (from Llemma) remain frozen, while only lightweight adapters and projection layers are fine-tuned. This design reduces memory requirements by an order of magnitude and cuts GPU hours by more than 50%, consistent with findings from parameter-efficient fine-tuning approaches such as LoRA and recent surveys on PEFT methods [4]. The result is faster convergence, lower energy consumption, and improved scalability for data environments. Also, the conversational interface supports iterative question-answering, allowing users to refine queries, request clarifications, and engage in deeper exploration of complex problems.

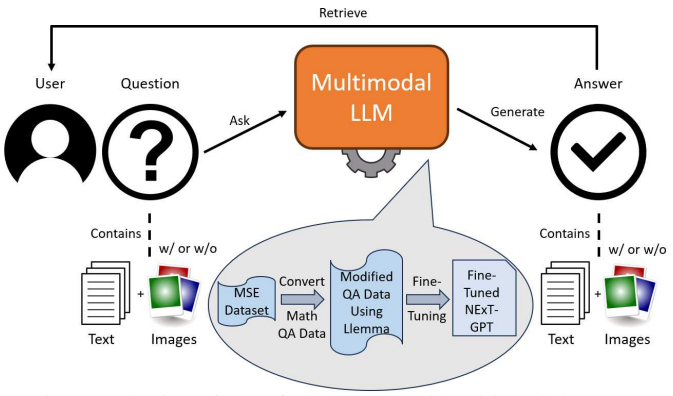

Fig. 2: The interface of the proposed multimodal system

IV. Advancing Multimodal Math Reasoning

Our framework delivers large-scale, multimodal mathematical reasoning with domain-specialized precision, efficiently integrating text, equations, and visual inputs to solve complex problems across diverse datasets. Unlike general-purpose models, it combines symbolic reasoning with flexible multimodal representation, enabling robust Math QA at scale. The framework's advances can be summarized in four key contributions:

(i) *Domain-Specialized Reasoning*: Embeds *Llemma* to interpret symbolic structures accurately, enabling robust problem-solving across millions of math problems.

(ii) *Unified Multimodal Representation and Scalability*: Aligns text, equations, and visual inputs in a shared embedding space while supporting seamless addition of new modalities, such as audio, 3D data, or point clouds, without retraining the full system.

(iii) *Efficient Generalization and Real-World Impact*: Leverages few-shot and fine-tuned learning to perform accurately on unseen problems, producing personalized outputs suitable for education, analytics, and scientific applications.

(iv) *Flexible Multimodal Input and Self-Verifying Reasoning*: Most existing math LLMs handle either text or a text-image blend and cannot natively leverage multiple modalities. The framework emphasizes self-verification and solution explanations, reducing hallucinations and incorrect answers while supporting learners in independently solving problems. With extension, our framework can also handle processing of diverse data sources, producing customizable outputs, e.g., videos or audio.

Together, these contributions establish a scalable, domain-specialized framework that sets a new benchmark for multimodal math reasoning at scale.

V. Experimental Results

In this section, we analyze the performance of the proposed multimodal math LLM. The codes of the models and experimental results gathered from the empirical study can be found at https://anonymous.4open.science/r/MathLLM-B32B, whereas the dataset used for performance evaluation is archived at https://huggingface.co/datasets/Zenos5/mse-text-img-dataset.

I'm delving into a linear algebra approach akin to Sheldon Axler's "Linear Algebra Done Right," which eschews traditional determinant expressions to improve conceptual clarity, particularly focusing on the challenging proof of volume properties via multilinear and alternating transformations. I aim to establish determinant properties using eigenvalues, specifically $\det(A) = \Pi_{i=1}^{N} \lambda_i$, and $\det(AB) = \det(A)\det(B)$, and explore whether understanding the product of eigenvalues or singular values for matrix combinations, especially with unitary and diagonal matrices, can offer new insights into these properties.

(a) A sample math question

Defining the determinant as the sole n-linear alternating map where $\det(\overrightarrow{e_1}, \ldots \overrightarrow{e_n}) = 1$ establishes foundational properties, including $\det(AB) = \det(A)\det(B)$. Demonstrations leverage matrix properties, illustrating that the determinant inversely relates to matrix inversion and directly correlates to the product of diagonal elements in diagonal and triangular matrices. Ultimately, the determinant of any matrix M, diagonalizable or in Jordan-Canonical form, equates to the product of its eigenvalues, $\det(M) = \lambda_1\lambda_2 \ldots \lambda_n$, providing a profound insight into its spectral characteristics through an algebraic expression.

(b) A sample math answer

Fig. 3: A math question and one of the potential answers from the Math Stack Exchange (MSE) dataset in which all of the common keywords are underlined and similar or same formulas are in red color

A. Dataset and Fine Tuning

To evaluate the performance of our multimodal math LLM, we constructed a custom text-image dataset derived from the Math Stack Exchange (MSE) corpus. The MSE dataset is designed specifically for Math QA, whereas most existing resources serve broader purposes. In contrast to ARQMath, which is restricted to text-only content, MSE includes both textual and visual mathematical information. To ensure quality, only relevant responses were retained, defined as either the highest-ranked answer or the one marked correct by the original asker. The dataset is partitioned into training, validation, and test sets to support model development, performance evaluation, and future research in multimodal math LLM.

Example 1: Figure 3 depicts a sample math question Q and one of its potential answers A, included in the MSE corpus, from which we produce our text-image dataset.□

The MSE dataset contains 64,860 questions paired with 117,380 answers (1.81 answers per question on average), including metadata such as answer scores and acceptance markings. Each question and answer was rendered into an image from the original LaTeX-embedded text, producing a multimodal representation comparable to PDF formatting. Data is stored in JSON format, with separate files for training, testing, and validation splits (90%, 5%, 5%) and corresponding image directories indexed by question and answer IDs. With over 180,000 text and image instances, this dataset is among the largest publicly available multimodal Math QA benchmarks, substantially exceeding the scale of MathQA (∼37K questions), Math23K (∼23K questions) and MathVista (∼6K questions). Its structured JSON and image organization supports efficient storage, parallel data loading, and scalable distributed training, enabling high-throughput experimentation with both NExT-GPT and Llemma. Llemma is evaluated using the text-only portion, while NExT-GPT processes the full multimodal data. This dataset provides a foundation for large-scale analytics, real-time QA systems, and benchmarking of next-generation multimodal models.

The training portion of the text-image dataset is designed

to fine-tune the multimodal Math QA model while keeping the internal Llemma weights frozen. This approach enables the model to process math questions incorporating textual, mathematical, and visual components, and to generate answers in the same multimodal formats. By leveraging cross-modal representations, the model can provide more comprehensive and flexible solutions, enhancing learners' understanding.

Llemma was initialized with Code Llama 7B weights and trained on the Proof-Pile-2 for 200B tokens. This method is used as the base textual model for our multimodal math LLM. While the original Llemma was trained on a wide variety of quality mathematical and programmatic text from the internet, this model was not specified for giving answers to math questions, and may struggle to provide answers to user queries in a comprehensible, responsive manner. To alleviate this, our version of the model is further fine-tuned using a different subset of the MSE dataset, all submitted and responded to by users of MSE. This dataset contains 395,869 math questions in LaTeX format with ranked answers.

B. Performance of the Modified Llemma Model

To evaluate the effectiveness of our modified Llemma model, we compare it against the original Llemma by examining the impact of n-shot learning and fine-tuning with samples from the MSE dataset. We report multiple LLM-centric performance measures across different model iterations. While zero-shot learning relies solely on pre-trained knowledge, n-shot learning enhances generalization to new tasks by providing n ($\geq$ 1) examples at inference time. In our experiments, we assess both the baseline Llemma and the proposed multimodal variant under n-shot (n = 1, 5, 10) learning and fine-tuning on the curated MSE dataset to determine which configuration achieves superior results.

1) Assessment Metrics: To evaluate the performance of our multimodal LLM, particularly in handling large-scale and complex data sources, we employ three essential metrics: *Answer Relevancy* (AR) Contextual Precision (CP), and Contextual Recall ($CREC$) AR uses LLM-as-a-judge to assess how well the generated output aligns with the input query, providing a direct measure of answer quality. CP examines whether relevant portions of the retrieval context are ranked higher than non-relevant ones, while $CREC$ measures how closely the retrieval context supports the expected output. Together, these metrics capture both the *quality* and *contextual relevance* of answers, allowing us to rigorously assess outputs that require advanced reasoning across text and mathematical content. Because the LLMs employed in evaluation (e.g., $Llemma$ and LLM-as-a-judge) are restricted to narrowly defined tasks, stochastic variability and flakiness in scoring are minimized. This metric suite is therefore critical for benchmarking reliability, scalability, and reasoning accuracy in multimodal LLMs designed for data applications.

The performance of the modified Llemma model under different training configurations, compared to the original Llemma, is summarized in Table I. Each metric—*Answer Relevancy* (AR), *Contextual Precision* (CP), and *Contextual Recall* ($CREC$)—produces a value between 0 and 1. A test case is considered successful if the metric value exceeds a predefined threshold, which is set based on baseline performance or domain-specific standards (e.g., 0.5 for relevance metrics). Values below the threshold are considered failures, and the reported metric represents the percentage of test cases that pass. Table I shows that the *modified* Llemma model achieves accuracy values *equal* to or *higher* than the *original* Llemma across all n-shot configurations and with fine-tuning. The table also depicts that increasing n in n-shot learning generally improves performance, as more examples allow the model to build a more robust task representation, leading to better generalization on unseen data. Accuracy continues to improve with higher n, and fine-tuning further enhances the model's effectiveness, as reflected in the reported metrics.

TABLE I: Assessment based on the MSE dataset, where O and M in each cell '*O//M*', such as 25%//30%, stand for the *Original* and *Modified* Llemma accuracy values, respectively for the N-shot training and fine-tuning (FT)

N-**Shot**	0	1	5	10	FT
Answer Relevancy (AR)	25% 30%	7% 9%	8% 11%	19% 23%	30% 36%
Contextual Precision (CP)	81% 92%	92% 97%	91% 97%	92% 98%	86% 98%
Contextual Recall ($CREC$)	8% 8%	7% 8%	6% 7%	13% 15%	27% 29%

TABLE II: Accuracy results on the MSE dataset, where O//M indicates *Original* vs. *Modified* Llemma accuracy (measured by %) at thresholds 0.5 and 0.7 , respectively

N-**Shot**	0	1	5	10
AR	75 86	90 100	86 80	78 87
CP	88 94	92 97	95 95	93 93
$CREC$	99 99	94 99	99 99	99 99

(a) Threshold set to be 0.5

N-**Shot**	0	1	5	10
AR	25 17	28 19	27 38	24 48
CP	90 92	98 97	95 95	93 94
$CREC$	6 8	7 8	9 9	9 9

(b) Threshold set to be 0.7

The default standard initial test case threshold of **0.5** proved too lenient. We used it to evaluate the baseline Llemma-7B model in 0-, 1-, 5-, and 10-shot settings (see Table IIa), but adding just one example at inference caused AR to reach 100%, limiting its usefulness for comparing performance at higher n-shot levels. To implement more conservative validation, we set the test case threshold to **0.7**, as this better constitutes a reasonable test case threshold for relevancy, precision, and recall. However, while CP still had a high pass rate, the other metrics were far too low, with AR hitting 28% passing for the test cases at one shot, as shown in Table IIb. CP with a threshold of 0.7 returned similar results to using a threshold of 0.5, and would benefit from using a higher threshold value. We settled on a threshold of **0.6** for AR and $CREC$ and a threshold of **0.8** for CP. All of the performance assessment results show that the modified Llemma outperforms the original Llemma in most cases.

2) Discussion on Threshold Setting: We note that the thresholds used for the test cases imply all model versions have relatively high CP, with 92-97% of test cases passing when a

score of 70% or higher was considered sufficient. This implies that the model architecture does a good job of identifying which parts of the retrieval context are relevant to the given input and ranking them higher within the neural network. AR and $CREC$ required lower test case thresholds to achieve comparable pass rates. This does not indicate that the LLMs failed on these tasks—0.5 is a common threshold standard, and pass rates remained high—but rather that performance was relatively weaker compared to other tasks. This suggests the models perform worse on AR, particularly in aligning outputs with inputs or matching retrieval context to expected answers. This may partly stem from the test case design, which lacked surrounding context and accepted only a single correct answer. To address this, we used the top-ranked answer (based on user validation) as the expected output, and either the next best answer or the highest ranked question's answer as the retrieval context when no other answer was available in the MSE dataset.

Inconsistent content within the retrieval context may have made it difficult for LLMs to align context with the expected output. Additionally, the varied nature of user-provided answers could hinder the models' ability to compare generated output with the input. While such variation is valuable for learning and evaluating real-world performance, future test cases—such as AR, CP, and $CREC$—would benefit from standardized context.

3) Performance Comparisons with Other Math LLMs: To measure how well our multimodal math LLM performs in comparison to other existing math LLM models, we have calculated the *accuracy* of our multimodal LLM on a number of math benchmark datasets and compared them to the results of other baseline math LLMs. Our multimodal math LLM-Llemma version employs 7 billion parameters and leverages Chain-of-Thought (**CoT**) prompting, enabling complex reasoning through intermediate steps. In order to have a fair comparison in regards of model performance we have identified *six*, including the original Llemma 7B, models. Out of these 6 baseline models, which also use CoT prompting, one contains 8 billion parameters, another one 11 billion parameters, and the remaining ones 7 billion parameters. To assess the accuracy of the models, including the modified Llemma LLM, we evaluate them on *four* different math task benchmark datasets commonly employed for Math QA validation, testing how well the LLMs perform on diverse unseen data and tasks.

We compare our multimodal Llemma 7B model against the six baseline math LLMs, Llama 3.2 11B, Code Llama, Minerva, Mistral, InternLM Math, and Llemma 7B. For the benchmark assessment, we use the following four datasets for solving math tasks: MATH, GSM8k, SAT, and MMLU-STEM.

The comparative performance of the different math LLM models on the benchmark datasets is shown in Table III. Overall, the comparison clearly shows that our modified Llemma model outperforms other baseline math LLMs on various datasets, and the results are statistically significant ($p < 0.04$) based on the Wilcoxon Signed-Rank test [3].

TABLE III: Assessment of our modified Llemma model and a collection of baseline math LLMs of billion parameters. The models' *accuracy* is measured by percentage (%) on the four baseline math task datasets used for comparison purpose

LLM Model	Size	Math	GSM 8K	SAT	MMLU-STEM	Average
Llama 3.2	11B	8.5	28.7	32.6	27.2	24.3
Code Llama	7B	4.5	10.5	9.4	25.1	12.4
Minerva	8B	14.1	16.2	-	35.6	22.0
Mistral	7B	11.6	41.2	59.4	**49.5**	40.4
Intern Math	7B	14.4	41.8	37.5	24.8	29.6
Llemma	7B	18.0	36.4	53.1	37.7	36.3
Our LLM	7B	**18.0**	**50.0**	**71.9**	46.1	**46.5**

4) Baseline Math QA Systems to be Compared: To further evaluate the effectiveness of the proposed multimodal LLM in retrieving relevant answers to math questions, we compare its performance against existing well-known Math IR/QA models designed for ranking and retrieving answers involving both textual content and mathematical expressions. We have selected *four* Math QA systems to use as *baseline* models for comparison purpose. Each of these models fulfills the same task as our multimodal LLM does, i.e., answering math questions based on textual content and mathematical notations (presented as plain text), but use different approaches to solve math problems. These Math QA models are $MIas$, $Tangent$-CFT, $Tangent$, and $MaRec$.

In short, $MIaS$ is a math-aware, full-text model that separates text and formulas, weighting expressions by their deviation from the original representation and using TF-IDF for scoring. $Tangent$-CFT embeds symbol-position sequences with fastText and uses cosine similarity, while $Tangent$ represents formulas as symbol layout trees for matching. $MaRec$ targets MathML-formatted text, ranking answers based on textual content and embedded formulas.

The experimental results of all of these methods conducted using the MSE dataset, along with the proposed multimodal LLM, are shown in Figure 4. Besides verifying the effectiveness of our multimodal LLM in combining the textual data and images specified in math questions for retrieving relevant answers to user questions, we also assessed the contribution of each modality by conducting an *ablation study* that evaluated the proposed multimodal LLM using text-only and image-only inputs, isolating the impact of each on overall performance. The results have verified (as shown in Figure 4) that the two subsystems contributes to the overall performance of our multimodal LLM instead of relying on individual one.

Observation. Compared with baseline Math QA models, the proposed multimodal LLM consistently achieves superior performance. It yields higher precision (P@1, P@3, P@5) and an nDCG@5 of 0.321, exceeding all baselines (<0.315). Its MRR of 0.398 also surpasses $MaRec$ (0.385), the strongest baseline. These improvements are statistically significant ($p < 0.05$) under the Wilcoxon Signed-Rank test. Standard IR metrics ensure fair comparison and support evaluating scalability and robustness in large-scale settings.

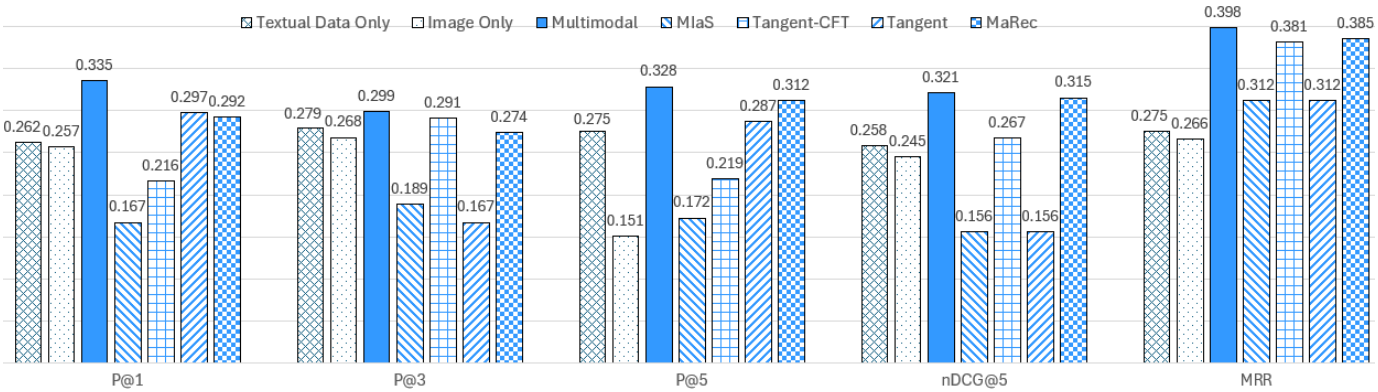

Fig. 4: Experimental results of the different baseline models and our proposed multimodal LLM on the MSE dataset

TABLE IV: Question processing time (in seconds) between different math IR models, the proposed multimodal, and $ChatGPT$

Models/Processing Time	MaRec	MIaS	Tangent-CFT	Tangent	Multimodal	ChatGPT
	0.037s	0.398s	0.405s	4.000s	0.631s	15.042s

5) The Lack of Multimodal Math Dataset: While Section III outlines the theoretical framework for our multimodal math LLM, building such a model in practice requires a multimodal math dataset that integrates multiple data sources, including text and images. Without such data, the model cannot be trained to handle heterogeneous mathematical inputs or evaluated against equivalent multimodal sources. However, few existing datasets incorporate both visual and textual math content, particularly for Math QA tasks involving text-image pairs. Consequently, we first focus on constructing a custom multimodal Math QA dataset before developing the multimodal projection system around our math LLM. Details of the proposed MSE dataset are provided in Section V-A, and the dataset constitutes a contribution to the Math QA community.

C. Question Processing Time

A key design goal of our multimodal LLM is to achieve question processing times comparable to web search engines. As shown in Table IV, our model retrieves and ranks answers significantly faster than ChatGPT. While it is slightly slower than some traditional IR-based baselines (except Tangent), the differences are modest, supporting our claim that the proposed model improves the efficiency of LLM-based Math QA.

VI. Conclusion

Extracting math equations from images enables access to a wide range of previously untapped resources, including electronic files, printed documents, and digital photos that conventional Math QA systems cannot process. Because manual transcription is slow and error-prone, automated extraction is essential for improving the efficiency of math IR and QA.

Our multimodal system integrates Llemma 7B, an LLM based on a decoder-only transformer and fine-tuned with Python interpreters and formal theorem provers to verify problem structure and solution logic. Embedded within the NExT-GPT architecture, the model processes projected text inputs, generates auto-checked outputs, and integrates multimodal inputs via conceptual tokens. Solution steps are verified and corrected, with text, formulas, and images mapped into a shared embedding space for coherent interpretation. Controlled generation with signal tokens, instruction-tuned outputs, and selective-layer fine-tuning preserves stability while enabling scalable, accurate reasoning across diverse datasets.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, and Others. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] R. Anil, A. Dai, O. Firat, M. Johnson, and Others. Palm 2 Technical Report. *arXiv preprint arXiv:2305.10403*, 2023.

[3] W. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, 2010.

[4] N. Ding, Y. Qin, G. Yang, F. Wei, and Others. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[5] National Center for Educational Statistics. The NAEP Mathematics Achievement Levels by Grade. https://nces.ed.gov/nationsreportcard/mathematics/achieve.aspx, October 2022.

[6] IES. Highlights of U.S. PISA 2022 Results Web Report (NCES 2023-115 and 2024-103). nces.ed.gov/surveys/pisa/pisa2022/index.asp, 2024.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, and Others. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[8] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K. Chang, M. Galley, and J. Gao. Mathvista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *Proceedings of the ICLR*, pages 1–116, 2024.

[9] T. Lv et al. Kosmos-2.5: A Multimodal Literate Model. *arXiv preprint arXiv:2309.11419*, 2023.

[10] B. Nye, A. Graesser, and X. Hu. AutoTutor and Family: A Review of 17 years of Natural Language Tutoring. *AIED*, 24(4):427–469, 2014.

[11] Z. Pan et al. Lemma: Learning from Errors for Mathematical Advancement in LLMs. *arXiv preprint arXiv:2503.17439*, 2025.

[12] Average Public School Math Proficiency. Public School Review. public schoolreview.com/average-math-proficiency-stats/national-data, 2023.

[13] Y. Tong, X. Zhang, R. Wang, R. Wu, and J. He. Dart-Math: Difficulty-Aware Rejection Tuning for Mathematical Problem-Solving. *Advances in Neural Information Processing Systems*, 37:7821–7846, 2024.

[14] K. Wang, H. Ren, A. Zhou, Z. Lu, S. Luo, W. Shi, R. Zhang, L. Song, M. Zhan, and H. Li. Mathcoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning. In *Proceedings of the ICLR*, 2023.

[15] S. Wu, H. Fei, L. Qu, W. Ji, and T. Chua. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the 41st ICML*, volume 235, pages 53366–53397, 2024. Article 2187.

[16] J. Zhang, Z. Li, M. Zhang, F. Yin, C. Liu, and Y. Moshfeghi. GeoEval: Benchmark for Evaluating LLMs and Multi-Modal Models on Geometry Problem-Solving. In *Proceedings of ACL 2024*, pages 1258–1276, 2024.

[17] A. Zhou et al. Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-Based Self-Verification. In *Proceedings of the ICLR*, pages 1–27, 2023.