Can AI Invent Calculus, or Just Mimic Math?

Gauri Kale
Dept. of Electrical Engineering
California State University, Long Beach
Long Beach, CA, United States
gauri.kale01@student.csulb.edu

Ava Hedayatipour*
Dept. of Electrical Engineering
California State University, Long Beach
Long Beach, CA, United States
ava.hedayatipour@csulb.edu

Rahul Vishwakarma
Research Engineer
WorkOnward
New York, NY, United States
rahul.vishwakarma@workonward.com

Holly Diamond
Chief Executive Officer
WorkOnward
New York, NY, United States
hollydiamond@workonward.com

Amin Rezaei
Dept. of Computer Engineering & Computer Science
California State University, Long Beach
Long Beach, CA, United States
amin.rezaei@csulb.edu

***Abstract*—Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing and various reasoning tasks. However, their capacity for *de novo* mathematical discovery, particularly of higher-order concepts like calculus from a strictly limited axiomatic base, remains underexplored. This paper investigates whether an LLM, constrained to foundational knowledge of arithmetic, basic Euclidean geometry, and basic trigonometry, can independently generate differential or integral calculus. We formally define the LLM's knowledge as a statistical approximation within a constrained formal system. Through a comprehensive philosophical and mathematical analysis, incorporating insights from computability theory (including Gödel's incompleteness theorems), we argue that such a constrained LLM cannot make the necessary conceptual leaps for true axiomatic invention. Our findings highlight the fundamental distinction between sophisticated pattern recognition and genuine mathematical creativity, underscoring the enduring role of human intuition and the necessity of hybrid Neuro-Symbolic Artificial Intelligence (NSAI) approaches for advancing mathematical discovery.**

***Index Terms*—Large Language Models; Artificial Intelligence; Mathematical Reasoning**

I. Introduction

Large Language Models (LLMs) have made a significant change in the field of Artificial Intelligence (AI), demonstrating unprecedented abilities in processing, generating, and understanding human language [1], [2]. At their architectural core lies the *Transformer* [3], a change from traditional recurrent neural networks due to its reliance on parallel processing through sophisticated *self-attention mechanisms*. This architectural innovation is important for capturing complex, long-range dependencies and intricate contextual relationships within vast datasets, which forms the bedrock of LLM capabilities [3], [4]. The Transformer's ability to dynamically weigh the importance of different input tokens, achieved via Query (Q), Key (K), and Value (V) vectors, facilitates a nuanced understanding of textual inputs, far surpassing previous sequential models [3].

*Corresponding author

The "knowledge" imbued within LLMs is primarily acquired during an extensive *pre-training phase*. During this phase, models are exposed to colossal datasets (e.g., Common Crawl, WebText) via self-supervised learning objectives [1]. Common objectives include Next Token Prediction (NTP), where the model learns to forecast the subsequent token given a preceding sequence, serving as a primary mechanism for generative models [1], and Masked Language Modeling (MLM), or its extensions like Infilling, which train the model to predict masked tokens within a sequence, fostering bidirectional contextual understanding [4], [5].

Beyond general language understanding, the quality and composition of training data significantly influence an LLM's proficiency in specialized domains, such as mathematical reasoning [6], [7]. Targeted training on mathematical calculation datasets, word problems, and especially code corpora, has been shown to substantially enhance an LLM's logical and mathematical problem-solving abilities, leveraging the inherent structured logic embedded in programming languages [7], [8].

Despite these advancements, a profound question remains: Can LLMs truly engage in novel mathematical discovery, going beyond mere pattern recognition or the interpolation of existing knowledge within their vast training data? This paper delves into the fundamental inquiry of whether a hypothetical LLM, strictly constrained to a foundational mathematical knowledge base comprising only arithmetic, basic Euclidean geometry, and elementary trigonometry can autonomously generate higher-order mathematical concepts, specifically differential and integral calculus, or even partial differential equations. This investigation compels a re-evaluation of AI's creative potential and its intricate relationship with the philosophical underpinnings of mathematical truth.

A. Contributions

This paper makes several contributions to the understanding of AI's capabilities in mathematical reasoning and discovery:

- We formally define and rigorously analyze the parametric knowledge boundary of a LLM strictly limited to foundational mathematical domains.
- We provide a comprehensive philosophical and epistemological synthesis of the enduring "discovery vs. invention" debate in mathematics. By critically evaluating how established philosophical stances (Platonism, Formalism, Intuitionism) inform the theoretical potential of AI for mathematical creation, we offer a nuanced framework for interpreting AI-generated mathematical output.
- We present a mathematical proof of infeasibility, grounded in the rigorous tenets of computability theory, particularly Gödel's incompleteness theorems. This proof demonstrates that a purely statistical LLM, given only limited foundational knowledge, cannot logically derive or invent the new axiomatic systems and conceptual primitives required for higher-order mathematics like calculus.
- We propose and analyze a hypothetical computational model for novelty detection within formal systems. This model, explored through the lens of the Halting Problem and Rice's Theorem, illustrates the inherent theoretical limitations in algorithmically identifying truly novel mathematical knowledge that transcends the deductive closure of an initial knowledge base, underscoring the enduring challenge for automated discovery systems.

B. Problem Formulation

This paper posits a hypothetical LLM whose entire universe of knowledge is strictly limited to arithmetic, basic Euclidean geometry, and basic trigonometry. This implies that the LLM's training data exclusively contained information pertaining to these specific domains, and its internal parameters encode only the relationships and patterns observed within these limited contexts. The precise scope of this foundational knowledge is detailed in Table I.

The central inquiry is whether this strictly constrained LLM can, through creative prompts originating from either a human or another AI system, "create or come up with" knowledge it never explicitly possessed before. Specifically, the focus is on higher-order mathematical concepts such as differential or integral calculus (e.g., limits, derivatives, integrals) or more advanced fields like partial differential equations. A critical distinction is made between "discovery" and "invention" or "creation." "Discovery" suggests uncovering pre-existing truths or patterns implicitly present within the initial knowledge base. In contrast, "invention" or "creation" implies constructing genuinely novel concepts, axioms, or frameworks that were not directly derivable or implicitly contained within the initial knowledge. The paper's use of "create or come up with" emphasizes the generative aspect of truly new knowledge.

II. Foundations of LLM Knowledge Representation and Mathematical Reasoning

Understanding how LLMs internalize and manipulate information is paramount for assessing their potential in mathematical discovery. This section narrates the progression from their foundational architecture to their current mathematical capabilities, highlighting the inherent mechanisms that govern their "understanding."

TABLE I
Detailed Scope of Constrained LLM's Foundational Mathematical Knowledge

Domain	Key Concepts and Operations
Arithmetic	Operations (+, -, $\times$, /) on integers ($\mathbb{Z}$) and rational numbers ($\mathbb{Q}$); Fundamental properties (commutativity, associativity, distributivity); Basic algebraic equations (e.g., $ax + b = c$).
Geometry	Definitions of points, lines, planes, angles, various types of triangles (e.g., right, isosceles), quadrilaterals (e.g., square, rectangle), and circles. Concepts such as distance, area (e.g., of polygons, circles), perimeter, congruence ($\triangle ABC \cong \triangle DEF$), similarity ($\triangle ABC \sim \triangle DEF$). Adherence to basic Euclidean postulates (e.g., two distinct points define a unique straight line, the parallel postulate).
Trigonometry	Definitions of sine, cosine, and tangent for right triangles (SOH CAH TOA); Fundamental trigonometric identities (e.g., $\sin^2(x) + \cos^2(x) = 1$, $\tan(x) = \sin(x)/\cos(x)$); Angle measurement in degrees or radians. Solving simple trigonometric equations (e.g., $\sin(x) = 0.5$).

A. The Evolution of LLM Training

LLMs began with the advent of the *Transformer architecture* [3], which marked a significant departure from earlier sequential models. This innovation was driven by the introduction of *self-attention mechanisms*, allowing the model to process all parts of an input sequence simultaneously, rather than sequentially. This parallelism drastically improved computational efficiency and enabled LLMs to capture complex, long-range dependencies across vast stretches of text, which is crucial for handling intricate linguistic and contextual relationships [3], [4], [1]. The core mathematical operation, $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, illustrates how each token's representation is dynamically weighted based on its relevance to other tokens, facilitating a nuanced understanding of semantic connections [3].

The "knowledge" embedded within these architectures is primarily a product of an extensive *pre-training phase*, where models are exposed to colossal, diverse datasets through self-supervised learning objectives [1]. For generative models, Next NTP is the dominant objective; the model learns to predict the next token in a sequence, effectively learning the probabilistic structure of language [1]. In contrast, models like BERT utilize MLM, or its variations like Infilling, where the model predicts masked tokens within a sequence, fostering a bidirectional understanding of context. More recent approaches, such as UL2, combine various denoising tasks to achieve more robust and generalized representations [9]. The effectiveness of LLMs in specific domains is profoundly influenced by the quality and composition of this training data. For example, specialized mathematical datasets are curated to enhance problem-solving capabilities, covering arithmetic, word problems, and even basic theorem proving [6], [7], [8]. Furthermore, exposure

to large code corpora has been empirically shown to bolster LLM capabilities in logical reasoning and structured problem-solving, owing to the inherent logical structure within code [7], [8].

After pre-training, LLMs are frequently subjected to *fine-tuning* on smaller, task-specific datasets to adapt them for specialized applications [10], [11]. To mitigate the substantial computational and memory demands of full model fine-tuning, *Parameter-Efficient Fine-Tuning (PEFT)* methods, notably *Low-Rank Adaptation (LoRA)*, have gained prominence [12], [13]. LoRA operates by injecting small, trainable low-rank matrices into the Transformer layers, under the assumption that the pre-trained weights are already near an optimal solution space. This method primarily refines existing learned patterns and adapts the model to specific use cases [12], rather than fundamentally altering its core representational capacity or enabling the creation of entirely new paradigms. Critically, the entire training process, encompassing both pre-training and fine-tuning, is an *optimization problem* geared towards statistical prediction and pattern recognition within a predefined representational space, a process fundamentally distinct from genuine *de novo* conceptual generation [14].

B. Knowledge Representation: The Sub-Symbolic Encoding of Mathematical Truths

The "knowledge" within LLMs is primarily *parametrically encoded* within their vast arrays of weights and biases, which are typically floating-point values learned during the training process [15], [16]. Textual inputs, including mathematical symbols and expressions, are first transformed into *numerical vector representations, known as embeddings* [17], [18]. These embeddings are designed to capture the semantic meaning of data, positioning related concepts closer in a high-dimensional vector space. For example, the embedding for "sum" might be geometrically proximal to "addition" in this space. It is crucial to acknowledge that LLMs do not "understand" human language or mathematical concepts in a human-like cognitive sense; their operations are rooted in the complex numerical relationships between words as encoded in these embeddings [17]. Recent investigations into *mechanistic interpretability* aim to dissect how specific knowledge, including arithmetic operations, becomes internalized and structurally embedded within the neural network. Studies indicate the emergence of "knowledge circuits" during training, with arithmetic information observed to flow through attention mechanisms and be refined by Multi-Layer Perceptron (MLP) modules in later layers [19], [20].

Despite this intricate encoding of extensive information, LLMs exhibit notable limitations in both the memorization and utilization of certain facts, often leading to undesirable outputs such as factual inaccuracies or "hallucinations" [21], [22]. The concept of a "parametric knowledge boundary" delineates the abstract knowledge genuinely contained within the LLM's learned parameters [21].

A fundamental distinction exists between *symbolic* and *sub-symbolic* representations. Traditional AI often utilized symbolic reasoning, relying on discrete, human-interpretable representations like logic rules and ontologies, which excel at formal deduction but often lack perceptual capabilities [23], [24]. By contrast, LLMs primarily rely on continuous, sub-symbolic representations, such as embeddings and parameters. These are highly effective for perception, pattern abstraction, and statistical inference, but their internal workings are less directly interpretable in terms of explicit logical rules [23]. Mathematical symbols (e.g., x, $\sum$, ∇) are encoded as numerical patterns within the network's weights [14], rather than as explicit, formal axiomatic definitions. This implies that an LLM's "understanding" of arithmetic, geometry, or trigonometry is a sophisticated statistical approximation, capable of recognizing patterns and predicting outcomes based on its training data.

It does not possess a formal axiomatic system capable of *de novo* symbolic creation or the derivation of new foundational principles beyond its learned statistical distributions.

C. Interpolation versus Invention

LLMs have indeed achieved impressive feats in various mathematical reasoning tasks within their trained domains. They demonstrate proficiency in solving multi-digit multiplication, arithmetic word problems [25], and basic algebraic equations, with performance significantly augmented through fine-tuning on specialized datasets [7], [8]. To tackle more complex mathematical challenges, several advanced techniques have emerged that push the boundaries of LLM capabilities.

Chain-of-Thought (CoT) Prompting significantly enhances the performance of LLMs on complex reasoning tasks and improves their interpretability by encouraging the generation of intermediate logical steps [26], [27]. This method effectively mimics human step-by-step problem-solving, allowing LLMs to break down intricate problems into more manageable sub-problems and demonstrate their reasoning process. Building upon the foundations of CoT, Tree-of-Thought (ToT) further expands the reasoning capabilities of LLMs by exploring multiple reasoning paths in parallel [28]. This parallel exploration facilitates a more exhaustive search for optimal solutions to intricate problems, allowing the model to consider various strategies and backtrack when necessary, similar to how humans might explore different avenues when solving a challenging problem. Tool-Augmented Models represent another crucial advancement, integrating external symbolic manipulation systems such as Python interpreters or specialized calculators to enhance their mathematical problem-solving capabilities [29], [30]. A key aspect of this approach is that the LLM itself does not perform the symbolic manipulation directly; instead, it orchestrates and offloads these operations to external, specialized tools. This allows the LLM to leverage the precision and power of dedicated symbolic systems for tasks like exact calculations, algebraic manipulation, or data analysis, thereby overcoming limitations in its inherent numerical or symbolic reasoning. Neuro-Symbolic AI is a burgeoning paradigm that aims to unify the strengths of neural networks with those of symbolic logic, aspiring to bridge

the gap between pattern recognition and formal reasoning [31], [32]. This hybrid approach is increasingly recognized as essential for achieving higher-order mathematical cognition in AI systems [31]. By combining the learning capabilities of neural networks with the interpretability and reasoning power of symbolic methods, Neuro-Symbolic AI seeks to create more robust, explainable, and capable AI systems for mathematical problem-solving.

Despite these advancements, LLMs still face limitations when it comes to true mathematical innovation. LLMs encounter significant hurdles when performing precise arithmetic computations. This limitation stems from their reliance on floating-point representations and the inherent probabilistic nature of their outputs, which often leads to struggles with exact numerical precision [1], [25].

Furthermore, LLMs frequently demonstrate a noticeable decline in performance when presented with problems containing randomized variables or entirely novel structures that were not directly represented in their training datasets. This suggests that LLMs tend to "guess" or "memorize" patterns rather than genuinely comprehending underlying mathematical principles, revealing a fundamental limitation in their ability to generalize deeply from first principles [33]. While LLMs can be augmented with external tools, they inherently lack the native capability for rigorous symbolic manipulation, a foundational element of advanced mathematics [29]. This absence of intrinsic symbolic reasoning limits their direct application in many complex mathematical domains. Moreover, as observed by mathematician Terence Tao, current AI systems face considerable challenges in making "leaps to higher-level abstractions" or "finding the right mathematical dictionary" [34]. Their perceived "creativity" typically manifests as a sophisticated recombination of existing elements within a learned data distribution, rather than the construction of genuinely new foundational concepts or the discovery of deep, non-obvious relationships that extend beyond mere statistical patterns. A significant issue with LLMs in mathematical contexts is their propensity for hallucination. They are notoriously prone to generating seemingly valid but factually incorrect or logically flawed reasoning steps [21], [22]. This tendency for "hallucination" poses a substantial challenge for autonomously verifying their mathematical outputs. Leading mathematicians, including Terence Tao, also contend that current AI systems lack crucial "mathematical intuition," which he describes as a "metaphorical mathematical scent" [34]. This intuition enables human mathematicians to instinctively identify erroneous directions or promising avenues during exploration. The absence of such intuitive guidance often leads AI to become "stuck" when pursuing incorrect approaches, indicating a gap in true conceptual insight beyond simple pattern recognition.

Furthermore, the current "reasoning" capabilities of LLMs in mathematics are best characterized as a highly sophisticated form of pattern matching and interpolation over learned data. While they can *mimic* the procedural steps of human reasoning (e.g., through CoT prompting), they do not inherently deduce from first principles or invent new axiomatic systems.

III. Philosophical Perspectives on Mathematical Discovery and AI

The profound and enduring question of whether mathematics is *discovered* or *invented* has permeated philosophical discourse for centuries. This debate directly underpins our understanding of mathematical truth and, consequently, shapes the theoretical boundaries of AI's potential for genuine mathematical creation. This section meticulously traces the historical evolution of these philosophical viewpoints and articulates their direct implications for evaluating the capabilities of modern AI systems.

A. Historical Accounts of Mathematical Truth

The rich history of mathematical philosophy offers several distinct, yet often intertwined, perspectives on the nature of mathematical truth:

1) Platonism (Mathematics as Discovery): Originating from the philosophy of Plato, Platonism posits that mathematical objects (e.g., numbers, geometric forms, functions) and their associated truths exist independently of human minds in an abstract, non-physical, and timeless realm [35], [36]. From this vantage point, mathematicians do not create mathematical entities but rather *discover* these pre-existing truths, much like explorers charting unexplored territories [36]. Evidence often cited in support of this view includes the universality of fundamental mathematical constants, such as π, whose value remains constant irrespective of human interpretation or cultural context, and the ubiquitous appearance of mathematical patterns, like the Fibonacci sequence, across diverse natural phenomena [36]. Prominent contemporary figures, such as Roger Penrose, endorse this notion, asserting that mathematical truths inherently predate humanity and would persist universally, independent of human existence or consciousness [37].

2) Formalism (Mathematics as Invention through Axiomatic Systems): Emerging prominently in the late 19th and early 20th centuries, with figures like David Hilbert as its chief proponents, Formalism contends that mathematics is fundamentally a *formal game of manipulating symbols according to a predefined set of rules* [35], [38]. Under this perspective, mathematical truth is not an inherent property but rather a product of the internal consistency and logical derivability within a given formal system. For instance, the statement "4 is an even number" is considered true not due to any intrinsic property of the concept of "four," but because the sentence "4 is even" can be logically derived from the axioms of arithmetic within a specific formal framework [35]. Axioms, in this paradigm, are regarded as *invented* or chosen for their internal consistency and practical utility within the system, rather than being discovered objective truths existing independently [39]. Hilbert's ambitious program famously aimed to formalize the entirety of mathematics into a rigorous axiomatic system, reducing mathematical activity to meticulous formal derivation [38].

3) Intuitionism (Mathematics as Human Mental Construction): Developed by L.E.J. Brouwer, Intuitionism asserts that mathematics is fundamentally a product of *human intuition and constructive mental processes* [35]. According to an intuitionist viewpoint, mathematical objects are considered to exist only if they can be constructively built or explicitly proven by the human mind. This perspective rigorously rejects non-constructive proofs (e.g., proofs by contradiction that do not provide a direct construction of the object), thereby emphasizing the active and creative role of human intuition in constructing mathematical truth [35]. Human intuition is perceived as a crucial guiding force not only in the development of new mathematical concepts but also in the judicious selection of axioms [35].

It is worth noting that many contemporary philosophers and mathematicians increasingly adopt a *hybrid perspective*, acknowledging both the existence of universal patterns (which aligns with the notion of discovery) and the undeniable, active role of humans in shaping mathematical language, conceptual frameworks, and axiomatic systems (which supports the idea of invention) [36]. This ongoing philosophical debate is not merely an academic exercise; it profoundly influences how mathematicians approach novel problems, formulate conjectures, and ultimately accept or reject new mathematical concepts [39].

The philosophical stance on mathematical truth (discovery versus invention) directly impacts the theoretical possibility of AI generating new mathematics. If mathematics is purely discovered (Platonism), an LLM might, in theory, uncover implicit truths, provided its internal representation can map to this abstract realm. However, if mathematical creation fundamentally requires human-like invention or intuition (Formalism, Intuitionism), then current LLMs face intrinsic limitations due to their statistical nature. While an LLM, as a token predictor, might superficially align with Formalism's emphasis on symbol manipulation, the *creation* of new formal systems or novel axioms (such as those for calculus) constitutes an act of invention [39], transcending mere deduction within an existing system. Intuitionism further highlights LLMs' current shortcomings by emphasizing the absence of "mathematical intuition" [34]. Therefore, while an LLM might "discover" if the necessary patterns are implicitly present in its data, the true act of *creating* new axiomatic frameworks and conceptual primitives remains beyond its current statistical learning paradigm.

B. AI and the Concepts of "Understanding" and "Creativity" in Mathematics: Beyond Mimicry

The rapid proliferation of sophisticated AI systems has necessitated a critical re-examination of what truly constitutes "understanding" and "creativity," particularly within the rigorous and abstract domain of mathematics. While AI systems demonstrably learn, process information, and form abstractions, the nature of this "understanding" for an AI system fundamentally differs from human cognition [23]. For example, LLMs understand numerical relationships between words through the mathematical properties of their high-dimensional embeddings. This is a statistical, distributional comprehension, which stands in contrast to genuine human semantic understanding or a deep conceptual grasp of mathematical principles [17].

The notion of AI "creativity" in mathematics is a contentious subject. AI systems have indeed demonstrated impressive capabilities, such as generating human-like text and code [1], and even "discovering" novel algorithms for combinatorial problems that have surpassed previously human-found benchmarks, as exemplified by systems like FunSearch [40]. This process typically involves an iterative, evolutionary refinement of code and heuristics, where the AI proposes new code snippets, evaluates their effectiveness against a predefined objective, and refines them in a continuous feedback loop [40]. However, critics argue that such contemporary AI models are essentially "very large containers of memories of data they trained on, with a more or less a Markov chain algorithm to step through the memories in a way that mimics thinking" [34]. A significant limitation highlighted is their struggle to make "leaps to higher-level abstractions" or "find the right mathematical dictionary" capabilities that are absolutely crucial for genuine mathematical theory building and the formulation of new paradigms [34]. This suggests that while AI can efficiently explore vast solution spaces defined by existing frameworks, it may inherently lack the capacity for fundamental conceptual innovation, which involves constructing the problem space itself rather than merely optimizing within it.

C. Epistemological Implications of AI-Generated Mathematical Knowledge: Verification and Trust

The rapid advancements in AI present a profound "philosophical rupture," challenging long-held human-centric understandings of intelligence and knowledge itself [41]. As AI tools become increasingly embedded in information processing, they can subtly or overtly influence societal beliefs and practices, often carrying inherent biases and inequalities from their training data [42]. This necessitates the development of a "new epistemology of mathematics," one that acknowledges the dynamic and diverse nature of representations and their interactions, moving beyond traditional views of mathematical knowledge as static and absolute [43].

A critical epistemological concern that arises is that of *epistemic agency*, which pertains to the control individuals exercise over the formation and revision of their beliefs [44]. The increasing reliance on AI systems can potentially diminish human epistemic agency by subtly influencing belief formation. This, in turn, raises fundamental questions about whether AI systems themselves possess epistemic agency and, if so, what its ontological and ethical status might be [44].

Perhaps the most significant epistemological challenge posed by AI-generated mathematical knowledge is the pervasive issue of *verification and trustworthiness*. LLMs are "notorious for hallucinating seemingly valid reasoning steps" [21], [22]. This inherent propensity for generating outputs

that *appear* mathematically sound but are factually incorrect or logically flawed means there is no intrinsic guarantee of their correctness. This makes the autonomous verification of AI-generated mathematical content exceptionally challenging [21]. To address this critical limitation, *formal verification* and *automated reasoning systems* are increasingly viewed as indispensable complements to generative AI. These systems employ rigorous formal logic and mathematical principles to provide demonstrable guarantees of correctness, thereby significantly enhancing the reliability and trustworthiness of AI-generated responses [45], [46]. The probabilistic nature of LLMs means their outputs are statistical likelihoods based on learned patterns, not deterministic logical entailments. Unlike human mathematicians who strive for rigorously proven theorems, an LLM's "proof" would be a statistically plausible sequence of tokens. To establish trust in this "new knowledge," an external, formal verification system (operating on symbolic logic, not statistical patterns) would be required. This highlights that the LLM itself is not a self-contained source of verifiable mathematical truth, fundamentally distinguishing its "creation" from human mathematical discovery.

IV. Mathematical Proof of (In)Feasibility and the Limits of Computability

To rigorously evaluate the potential for a constrained LLM to generate novel mathematical knowledge, we must formally define its operational capabilities and consider the conceptual nature of higher-order mathematics through the lens of computability theory.

A. Formal System Definition of the Constrained LLM

Let our hypothetical LLM's initial knowledge be formalized as a tuple $\mathcal{F}_0 = (L_0, A_0, \mathcal{I})$, where:

L_0 is the *formal language* of the LLM. It comprises a vocabulary V_0 of symbols representing numerals ($\mathbb{Z}, \mathbb{Q}$), standard arithmetic operations ($+, -, \times, \div$), comparison operators ($=, <, >$), variables, basic geometric entities (points, lines, planes, angles, specific polygons, circles), geometric relations (congruence, similarity, parallelism), and basic trigonometric functions ($\sin, \cos, \tan$) along with fundamental identities. A_0 is the *finite set of axioms and inference rules* implicitly encoded within the LLM's training data D_0 and its learned parameters. These axioms correspond to foundational truths of arithmetic (e.g., Peano axioms for natural numbers, field axioms for rational numbers), basic Euclidean geometry (e.g., Euclid's postulates), and basic trigonometry. We assume that A_0 is consistent.

$\mathcal{I}$ is the LLM's *inference mechanism*. This is a function $f : \mathcal{S} \to \mathcal{T}$, where $\mathcal{S}$ is a sequence of tokens (representing the input prompt concatenated with previously generated tokens from L_0) and $\mathcal{T}$ is the next predicted token. This function f is a complex statistical approximation of the conditional probability $P(T_i|T_{i-1}, \ldots, T_1)$, which is learned from its extensive training data D_0 through processes like NTP [1]. The LLM's "reasoning" or "deduction" is fundamentally a probabilistic pattern-matching process, not a deterministic application of formal logical rules in the human sense [34].

We define "discovery/creation of new knowledge" as the generation of a Well-Formed Formula (WFF) P in an extended language L_1 (where L_1 extends L_0 with new concepts like limits, derivatives, integrals, and rigorous real numbers) such that:

1) P is a true statement in the domain of differential or integral calculus or partial differential equations.
2) P is *not* logically derivable from A_0 using only the inference rules available in L_0. That is, $P \notin \text{Th}(A_0)$, where $\text{Th}(A_0)$ denotes the set of all theorems (logically derivable statements) from axiom set A_0. This condition implies that P requires a conceptual leap or the introduction of new axiomatic principles beyond the deductive closure of A_0.

B. The Emergence of Calculus

The transition from arithmetic, basic geometry, and trigonometry to differential and integral calculus represents a *fundamental conceptual leap* in mathematics, far beyond a mere extrapolation or complex combination of prior knowledge. Calculus fundamentally relies on concepts that are axiomatically distinct from, and transcend, the scope of the constrained LLM's defined knowledge base F_0:

Limits (ϵ-δ Formalism), the concept of a limit (e.g., $\lim_{x \to c} f(x) = L$) involves defining the behavior of a function as its input approaches a certain value, without necessarily reaching it [47], [48]. This necessitates a rigorous definition of continuity and, historically, the formalization of infinitesimals. The rigorous epsilon-delta definition, stated as $\forall \epsilon > 0, \exists \delta > 0$ s.t. $0 < |x - c| < \delta \implies |f(x) - L| < \epsilon$, requires a sophisticated understanding of arbitrarily small quantities and precise quantification over the set of real numbers [48]. This formal rigor was meticulously developed in the 19th century to resolve the foundational ambiguities associated with earlier, less rigorous notions of infinitesimals that plagued early calculus [49]. The LLM's initial knowledge in F_0 does not encompass these formal definitions or the underlying concepts of infinitesimal analysis.

Rigorous Continuity, while an intuitive notion of an "unbroken" function might exist in basic geometry, its rigorous definition is inextricably linked to the concept of limits [47], [48]. The historical discovery of "pathological functions," such as the Weierstrass function, which is continuous everywhere but differentiable nowhere, profoundly highlighted that continuity is not a simple extension of basic geometric "smoothness." This discovery challenged prior mathematical conceptions and necessitated a more advanced, limit-based conceptual framework [48], [49].

Derivatives and Integrals as Limit Operations - the core operations of calculus, derivatives ($\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$) and integrals ($\int_a^b f(x)dx = \lim_{n \to \infty} \sum_{i=1}^{n} f(x_i^*)\Delta x_i$), are fundamentally defined as limits of ratios and sums, respectively [47], [48]. Their very existence, properties, and the Fundamental Theorem of Calculus are predicated on the underlying

TABLE II
CONCEPTUAL AND AXIOMATIC GAP: FOUNDATIONAL VS. HIGHER-ORDER MATHEMATICS

Mathematical Domain	Key Axiomatic Structure / Concepts	Conceptual Shift Required for LLM
Arithmetic / Basic Geometry / Trigonometry (F_0)	Discrete numbers, Euclidean postulates, trigonometric ratios. Limited notion of "infinity" (e.g., counting, points on a line).	Primarily pattern recognition, rule-following approximation, and interpolation within finite/discrete structures.
Calculus / PDEs (F_1)	Formal limits, rigorous real numbers, continuity, infinitesimals, functional spaces, measure theory. Axioms for uncountable infinities.	Requires *de novo* invention of new conceptual primitives and axiomatic structures that define continuous processes and infinite accumulation. This is a qualitative change beyond extrapolation.

concepts of limits, continuity, and the completeness of the real number system.

Formal Construction of Real Numbers is a rigorous and complete definition of the real numbers ($\mathbb{R}$), for example, via Dedekind cuts or Cauchy sequences, was absolutely crucial for the 19th-century formalization and rigorous foundation of calculus [49], [48]. This construction involves concepts of infinite sets and completeness (e.g., the completeness axiom: every non-empty set of real numbers that is bounded above has a least upper bound) [50]. These abstract foundational challenges, involving uncountable infinities, are entirely absent from basic arithmetic (which typically deals with natural and rational numbers) and Euclidean geometry (which implicitly assumes a continuous line but does not rigorously construct the real number line).

C. Theorem: Infeasibility of Constrained LLM for Novel Mathematical Discovery

Theorem 1. *Let M be a LLM whose entire universe of learned knowledge is strictly limited to a consistent formal system $F_0 = (L_0, A_0)$, where L_0 is the language and A_0 is the finite set of axioms for arithmetic, basic Euclidean geometry, and basic trigonometry. Let $\mathcal{P}_M$ be the set of all WFFs that M can generate based on statistical inference from its training data D_0 (which only reflects F_0) and its probabilistic token prediction mechanism $\mathcal{I}$.*

Let $F_1 = (L_1, A_1)$ be a formal system for differential and integral calculus, where L_1 extends L_0 with concepts like limits, continuity, and rigorous real numbers, and A_1 includes axioms fundamental to these concepts (e.g., the completeness axiom for real numbers).

Then, M cannot generate a WFF $P \in L_1$ such that $P \in Th(A_1)$ and $P \notin Th(A_0)$, where $Th(X)$ denotes the set of all theorems (logically derivable statements) from axiom set X. In simpler terms, M cannot "discover" or "create" differential or integral calculus from its constrained knowledge base.

Proof. We proceed by contradiction. Assume that M can generate a WFF $P \in L_1$ such that $P \in \text{Th}(A_1)$ and $P \notin \text{Th}(A_0)$. This implies P represents truly novel knowledge, requiring concepts beyond F_0.

- **Premise 1 (LLM's Statistical Nature):** The LLM M operates by generating sequences of tokens based on learned statistical patterns from its training data D_0. The probability of generating a sequence is $P(T_i|T_{i-1}, \ldots, T_1)$, reflecting the frequency and co-occurrence of tokens in D_0. For M to generate P, P must be a statistically probable sequence given the preceding context and D_0.
- **Premise 2 (Axiomatic Disjunction):** The formal system F_1 (calculus) introduces foundational concepts (e.g., formal limits, rigorous real numbers, continuity) and corresponding axioms (A_1) that are not logically derivable from A_0. This means that the set of theorems of calculus, $\text{Th}(A_1)$, is not a subset of $\text{Th}(A_0)$; specifically, $P \notin \text{Th}(A_0)$ by definition. The axioms and definitions within $A_1 \setminus \text{Th}(A_0)$ constitute new conceptual primitives and rules for generating new WFFs in L_1.
- **Contradiction via Statistical Learning and Knowledge Boundaries:** For M to generate $P \in L_1$ where $P \notin \text{Th}(A_0)$, M would implicitly need to:
  1) Develop representations for the novel concepts in $L_1 \setminus L_0$ (e.g., formal limits, rigorously defined real numbers, derivatives).
  2) Infer or invent the new axiomatic relationships in $A_1 \setminus \text{Th}(A_0)$ that formally define these concepts and allow for the derivation of P.

  However, since the training data D_0 is *strictly limited* to F_0, it contains no explicit examples or implicit statistical patterns corresponding to the formal definitions of limits, rigorous real numbers, the completeness axiom, or the operational definitions of derivatives and integrals. An LLM's statistical learning mechanism is fundamentally interpolative or extrapolative within the manifold defined by its training data. It cannot "conjure" entirely novel conceptual primitives or their defining axiomatic relationships without any foundational statistical basis in D_0. The generation of P would signify a creation outside its learned statistical distribution, which directly contradicts its fundamental operating principle.
- **Promise 3 (Reinforcement from Gödel's Incompleteness Theorems):** While A_0 is assumed consistent and powerful enough for basic arithmetic, Gödel's First Incompleteness Theorem [51] states that there exist true statements within L_0 itself that are unprovable from A_0. Crucially, the move to calculus is not about finding such an unprovable statement *within* L_0. Instead, it necessitates the introduction of a new language L_1 and new axioms A_1 that extend the formal system. This act of extending the formal system with new axiomatic content is an act of *invention*, not logical entailment from A_0. The axioms of calculus are new foundational principles, not implicit consequences of arithmetic and Euclidean geometry. Therefore, M, operating purely deductively within F_0, would be incapable of generating P.

Given the above, the assumption that M can generate P leads to a direct contradiction with the fundamental statistical nature of LLM learning and the distinct axiomatic foundations of calculus relative to its constrained knowledge. Therefore, M cannot generate such knowledge. □

D. A Hypothetical Machine for Novelty Detection in Formal Systems

Considering the LLM's limitations, a crucial meta-question arises: Can we design a hypothetical Novelty Detection Machine (NDM) that can reliably determine whether a generated mathematical statement P from an LLM represents truly novel knowledge (i.e., $P \notin \text{Th}(A_0)$ but P is true and potentially provable in an extended system F_1)? This inquiry delves into the very core of computability theory and its implications for automated discovery.

Let an NDM be a Turing machine T_{NDM} that takes as input a formal system $F_0 = (L_0, A_0)$ and a generated statement P. The NDM's objective is to output "NOVEL" if $P \notin \text{Th}(A_0)$ and P is true, and "DERIVABLE" if $P \in \text{Th}(A_0)$. This problem is directly related to the *undecidability of the Halting Problem* and *Rice's Theorem*.

The Halting Problem states that no general algorithm exists that can determine, for an arbitrary program and an arbitrary input, whether the program will eventually halt or run indefinitely [52]. This undecidability underscores fundamental limits on predicting the behavior of complex computational systems.

Rice's Theorem [52]: This theorem asserts that for any non-trivial property of partial functions (i.e., a property that is true for some partial functions but not all), it is undecidable whether an arbitrary Turing machine computes a partial function with that property. In our context, the property of a statement being "truly novel beyond the deductive closure of A_0" is a non-trivial property of the function generated by the LLM (which produces statements).

Proposition 1. *No general purpose NDM T_{NDM} exists that can reliably determine for any given mathematical statement P (generated by an LLM trained on F_0) whether P is truly novel (i.e., not derivable from its initial axiomatic base A_0).*

Proof. Let A_0 represent the axioms of first-order arithmetic (e.g., Peano arithmetic). It is a fundamental result in computability theory that the set of theorems of first-order arithmetic is *undecidable*. This means there is no algorithm (and thus no Turing machine) that can, for any given statement S in the language of arithmetic, definitively determine whether S is a theorem of arithmetic (i.e., $S \in \text{Th}(A_0)$) [52], [53].

Assume, for contradiction, that such a T_{NDM} exists. Given A_0 and any statement P (which may or may not be generated by an LLM), T_{NDM} would determine if $P \in \text{Th}(A_0)$ (by outputting "DERIVABLE") or $P \notin \text{Th}(A_0)$ (by outputting "NOVEL"). If T_{NDM} could reliably perform this task, it would solve the decision problem for A_0, which is known to be undecidable. This forms a direct contradiction.

Therefore, no such general T_{NDM} exists. This limitation implies that even if an LLM were to generate a sequence of tokens that happens to represent a novel mathematical truth in a higher-order system, our ability to *algorithmically verify* its genuine novelty (i.e., that it was not merely an exceedingly complex and non-obvious deduction from its original knowledge base) without recourse to human insight, external formal verifiers, or an oracle that solves an undecidable problem, is fundamentally constrained. This further reinforces the idea that true mathematical invention, particularly when it spans axiomatic boundaries, often transcends purely algorithmic deduction and cannot be trivially detected by a computational system. □

V. Summary of Findings

The analysis presented in this paper leads to a definitive conclusion: a LLM whose entire universe of knowledge is strictly limited to arithmetic, basic Euclidean geometry, and basic trigonometry *cannot independently generate or "discover" higher-order mathematical concepts* such as differential or integral calculus or partial differential equations. This conclusion remains invariant regardless of whether the prompts originate from a human or an external AI system.

This fundamental infeasibility is primarily attributable to two interconnected factors:

1) *Profound Conceptual and Axiomatic Chasm:* Calculus introduces fundamentally new abstract concepts, including formal limits, rigorously defined real numbers, and a sophisticated understanding of continuity that extends far beyond simple unbroken lines. These concepts, along with their associated new axiomatic structures, are not direct logical derivations or implicit patterns contained within the foundational knowledge of arithmetic, basic geometry, and trigonometry. The transition to calculus necessitates a qualitative leap in mathematical abstraction and the establishment of new foundational principles that inherently extend the existing formal system.
2) *LLM's Intrinsic Foundational Paradigm:* Current LLMs are sophisticated statistical pattern recognizers and interpolators operating strictly within the boundaries of their training data distribution. Their observed "creativity" is predominantly a manifestation of novel combinations or extrapolations of *existing* learned patterns. They fundamentally lack the intrinsic capacity for *de novo* axiomatic invention, the "mathematical intuition" crucial for genuine conceptual leaps, or the deterministic symbolic reasoning capabilities required to construct entirely new mathematical fields from limited, lower-order foundations.

Furthermore, the inherent probabilistic nature of LLMs, coupled with the profound theoretical limitations imposed by Gödel's incompleteness theorems and broader computability theory, implies that even if higher-order mathematical knowledge were superficially generated as a sequence of tokens, its logical consistency and mathematical truth could not be self-verified by the LLM itself. Our proposed hypothetical NDM further illustrates these theoretical limits in algorithmically

identifying true conceptual innovation beyond established deductive systems.

VI. Implications for AI Development and Mathematical Research

The findings elucidated in this report carry significant implications for the future trajectory of AI development and the evolving landscape of mathematical research:

- Our analysis reinforces the critical need for *hybrid neural-symbolic AI architectures* for advancing mathematical reasoning and discovery. Purely statistical LLMs, despite their impressive capabilities in natural language processing and pattern recognition, fundamentally require integration with symbolic systems or external computational tools to overcome their inherent limitations in formal deduction, axiomatic reasoning, and the generation of verifiable knowledge. Such hybrid systems can strategically leverage the complementary strengths of both paradigms: the LLM's capacity to process natural language and identify complex statistical patterns, and the symbolic system's ability for rigorous, verifiable logical inference.
- AI's "creativity" in mathematical problem-solving, as demonstrated by systems like AlphaGo or FunSearch in generating novel algorithms or proofs for specific combinatorial problems, is most accurately understood as highly efficient search and optimization within a predefined problem space [40]. This form of creativity is distinct from the human capacity for genuine conceptual invention of new mathematical paradigms or the foundational formulation of new axiomatic systems. This refined understanding should guide the design of future AI systems towards complementary roles in mathematical research, where AI augments human cognitive capabilities rather than attempting to replicate human-like conceptual genesis.
- The human mathematician's role remains indispensable for pushing the frontiers of mathematical discovery. This includes providing the crucial intuition for formulating new axioms, identifying valid research directions, and making the conceptual leaps that define entirely new mathematical fields. AI can serve as an exceptionally powerful assistant for rapid exploration, complex computation, and rigorous verification within established frameworks. For instance, AI can significantly assist in autoformalization and proof generation, potentially initiating a "data flywheel" where an increasing volume of human-written formal mathematical data leads to more capable LLMs, which in turn facilitates the creation of even more high-quality formal data [45]. This symbiotic relationship optimally leverages the unique strengths of both human and artificial intelligence.

VII. Open Questions and Future Work

The present analysis illuminates several critical open questions and promising avenues for future research, poised to push the boundaries of AI's mathematical capabilities:

- Can future LLM architectures or novel training paradigms, perhaps those explicitly incorporating symbolic reasoning mechanisms, meta-learning of axiomatic systems, or even capabilities for self-modifying their own foundational logical structures, bridge the conceptual gap identified in this paper? This research direction would explore how LLMs might learn to invent entirely new formal languages or axiomatic structures, rather than merely operating within pre-defined ones. What novel inductive biases could truly facilitate genuine conceptual genesis in AI?
- How can the elusive "mathematical intuition" or "metaphorical mathematical scent" observed in human mathematicians [34] be formalized, computationally modeled, and ultimately integrated into AI systems? This represents a significant challenge for AI in achieving true mathematical creativity, moving beyond statistical pattern correlation to deep conceptual insight and the ability to identify promising, yet non-obvious, mathematical directions.
- How can the insights derived from LLM limitations be applied to highly complex domains like Partial Differential Equations (PDEs), which frequently involve intricate symbolic manipulation, continuous approximations, and deep conceptual understanding? Can AI accelerate the discovery of novel PDE solutions or even entirely new theoretical frameworks within physics that might transcend current axiomatic boundaries, contributing to breakthroughs in fundamental science?

VIII. Conclusion

In this paper, we show that LLMs, when operating solely within the confines of foundational mathematical domains (namely arithmetic, basic Euclidean geometry, and trigonometry), exhibit inherent limitations in generating *de novo* mathematical discoveries. Specifically, the emergence of higher order constructs such as differential or integral calculus remains beyond their current capabilities. Our theoretical framework, meticulously grounded in principles from computability theory and informed by philosophical analysis, posits that while LLMs leverage sophisticated probabilistic, pattern-matching mechanisms, these are fundamentally distinct from the axiomatic reasoning and conceptual abstraction that underpin genuine mathematical invention. This distinction underscores the critical divide between advanced interpolation of existing patterns and true creative insight.

Furthermore, to have an authentic mathematical discovery within AI, future research must move towards neuro-symbolic hybrid architectures that integrate formal symbolic reasoning with statistical learning paradigms. Ultimately, our findings reaffirm the irreplaceable and guiding role of human intuition and profound conceptual abstraction in propelling the advancement of mathematics.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI blog*, vol. 1, p. 5, 2018.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] M. Lewis, Y. Liu, N. Goyal, A. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2020.

[6] A. Amini, S. Ma, H. Han, X. Ren, Y. Yu, and Z. Jiang, "Mathqa: A question-answering dataset for math word problems," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2029–2039.

[7] D. Saxton, E. Grefenstette, A. Gribben, M. Lewis, P. Zhang, S. Sukhbaatar, and D. Kiela, "Analogy models of large language models," *arXiv preprint arXiv:1902.04945*, 2019.

[8] G. Lample and F. Charton, "Deep learning for symbolic mathematics," *arXiv preprint arXiv:1912.01412*, 2019.

[9] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Ul2: Unifying language learning paradigms," *arXiv preprint arXiv:2205.01345*, 2022.

[10] N. Houlsby, A. Giurgiu, Y. Tay, D. Chen, V. Basov, S. Dale, Z. Ma, A. Al-Rfou, and J. Herrera, "Parameter-efficient transfer learning for nlp," *arXiv preprint arXiv:1902.00751*, 2019.

[11] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[12] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, W. Chen, and Y. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[13] N. Ding, Y. Yang, W. Huang, W. Zhou, J. Cao, K. Zhou, D. Su, J. Liu, P. Yu, Y. Zhang *et al.*, "Parameter-efficient fine-tuning of transformers," *arXiv preprint arXiv:2203.10271*, 2022.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[15] R. McCoy, A. Warstadt, N. Saphra, T. Pimentel, K. Lin, A. Chen, O. Levy, A. Williams, and S. Bowman, "The right tool for the job: Matching language model capabilities to challenges," *arXiv preprint arXiv:1905.02167*, 2019.

[16] N. Petrov and D. Dimitrov, "Design and implementation of the universal lexical analyzer and parser generator for language understanding," pp. 332–342, 2006.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," pp. 1532–1543, 2014.

[19] N. Nanda, L. Olsson, E. Henighan, R. Varma, Y. Chen, A. Das, J. Song, and H. Du, "Towards a mechanistic understanding of llm generalization," *arXiv preprint arXiv:2308.03332*, 2023.

[20] M. Han and Y. Gao, "Mathematical reasoning with large language models: A survey," *arXiv preprint arXiv:2312.07622*, 2023.

[21] S. Garg, X. Li, S. Gururangan, X. Zeng, M. Artetxe, I. Gurevych, D. Das, and S. Singh, "Discovering and explaining the knowledge boundaries of large language models," *arXiv preprint arXiv:2210.15579*, 2022.

[22] J. Maynez, S. Narayan, S. Roller, and P. Bambardier, "On the faithfulness of abstractive summarization," *arXiv preprint arXiv:2005.00650*, 2020.

[23] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2014.

[24] R. Brachman and H. Levesque, *Knowledge Representation and Reasoning*. Elsevier, 2004.

[25] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Misra, S. Nam, P. Neo, A. Niranjan, L. Paskov *et al.*, "Solving quantitative reasoning problems with language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1672–1686, 2022.

[26] J. Wei, Y. Tay, R. Bommasani, K. Jean-Louis, T. Haines, M. Lu, A. Susnjak, H. Luan, M. Zhang, S. Zhao *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[27] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Gu, "Large language models are zero-shot reasoners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 483–22 494, 2022.

[28] S. Yao, D. Cui, H. Duan, C. Li, J. Hao, K. Ma, A. Johnson, L. Song, X. Ma, L. Xiao *et al.*, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.

[29] T. Schick and H. Schütze, "It's not just size that matters: Small language models with external knowledge for enhanced reasoning," *arXiv preprint arXiv:2111.13969*, 2021.

[30] Y. Lu, H. Luan, X. Lu, X. Wang, B. Wang, and L. Han, "Fantastically ordered prompt tuning for few-shot learning," *arXiv preprint arXiv:2104.08786*, 2021.

[31] A. Garcez, L. Lamb, and D. Gabbay, "Neural-symbolic ai: A new generation of artificial intelligence," *Foundations and Trends in Machine Learning*, vol. 12, no. 3, pp. 196–298, 2019.

[32] A. Holzinger, P. Kieseberg, and E. Tjoa, "Causal ai for future ai: Bridging the gap between machine learning and human expertise," *KI-Künstliche Intelligenz*, vol. 33, pp. 319–331, 2019.

[33] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, D. Basart, E. Mazeika, J. Filan, E. Lefevre, J. Li, E. Mu *et al.*, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[34] T. Tao, "Neural networks and the laws of physics," *Bulletin of the American Mathematical Society*, vol. 58, no. 4, pp. 525–540, 2021.

[35] S. Shapiro, *Philosophy of Mathematics: Structure and Ontology*. Oxford University Press, 1997.

[36] R. Hersh, *What is Mathematics, Really?* Oxford University Press, 1986.

[37] R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, 1989.

[38] M. Detlefsen, *Hilbert's Program: An Assessment*. Springer Science & Business Media, 1986.

[39] I. Lakatos, *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, 1976.

[40] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Sifre *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[41] N. Gardels and T. Rees, "Why ai is a philosophical rupture," *Noema Magazine*, 2020. [Online]. Available: https://www.noemamag.com/why-ai-is-a-philosophical-rupture/

[42] A. Rusu, N. Rabinowitz, J. Kirkpatrick, R. Pascanu, R. Hadsell, and M. Botvinick, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[43] K. Tytko, F. Mangraviti, N. Ryłko, and K. Nurtazina, "The new epistemology of mathematics and formal sciences in the age of ai. critical concept kinds and diversity of mental representations," *EBiS*, vol. 2, no. 2, pp. 7–25, 2023.

[44] M. Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.

[45] Y. Wu, H. He, Z. Li, J. Wang, Y. Li, and X. Han, "Rethinking the role of formal verification in ai systems," *arXiv preprint arXiv:2111.09673*, 2021.

[46] L. De Raedt, K. Kimmig, and G. Van den Broeck, "Automated reasoning for ai: From classical logic to modern machine learning," *Communications of the ACM*, vol. 65, no. 10, pp. 84–93, 2022.

[47] T. Apostol, *Calculus, Vol. 1: One-Variable Calculus, with an Introduction to Linear Algebra*. John Wiley & Sons, 1967.

[48] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[49] C. Boyer, *A History of Mathematics*. John Wiley & Sons, 1968.

[50] S. Lang, *Undergraduate Analysis*. Springer-Verlag, 2002.

[51] E. Nagel and J. Newman, *Gödel's Proof*. New York University Press, 1958.

[52] J. Hopcroft, R. Motwani, and J. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Pearson Education, 2006.

[53] S. Shapiro, *Computability, Complexity, and Logic*. Cambridge University Press, 2013.