GeoNum: Bridging Numerical Continuity and Language Semantics via Geometric Embedding

Shengkai Jin[1], **Tianyu Chen**[1]*, **Chonghan Gao**[1], **Jun Han**[1]

[1]SKLCCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
{jinshengkai, tianyuc, gaoch, jun_han}@buaa.edu.cn

Abstract

Large language models excel at semantic reasoning yet struggle with numerical tasks because tokenization disrupts geometric continuity. Traditional methods fragment numerically close values into inconsistent token sequences, severing the correspondence between numerical proximity and representational similarity, which is essential for numerical cognition. We introduce **GeoNum**, a geometrically coherent numerical embedding based on polar coordinate decomposition. By encoding integer magnitudes through classification and fractional components via trigonometric regression, GeoNum constructs a continuous manifold where numerical distance is preserved geometrically. A three-stage framework progressively integrates GeoNum into pretrained language models via self-supervised pretraining, projection alignment, and efficient adaptation. Experimental results across diverse arithmetic benchmarks demonstrate consistent gains in high-precision accuracy and improved interpolation and extrapolation, underscoring the promising benefits of geometric continuity for numerical modeling in large language models.

Introduction

Large language models have achieved remarkable success in natural language understanding and generation, demonstrating sophisticated reasoning capabilities across diverse textual tasks (Achiam et al. 2023; Anil et al. 2023). However, their proficiency in handling numerical information reveals a fundamental limitation rooted in representational misalignment. While LLMs excel at learning semantic relationships through high-dimensional embeddings that preserve similarity structure, they tend to fail at numerical reasoning because tokenization disrupts the geometric continuity inherent to numerical cognition. As illustrated in Figure 1, traditional tokenization fragments numerically close values like 2.9 and 3.1 into inconsistent token sequences [2, ., 9] and [3, ., 1], completely severing their distance relationships. This creates a fundamental bridging gap where numerical proximity in value space bears no correspondence to proximity in representation space, forcing models to treat numbers as arbitrary symbolic tokens rather than meaningful points on a continuous manifold.

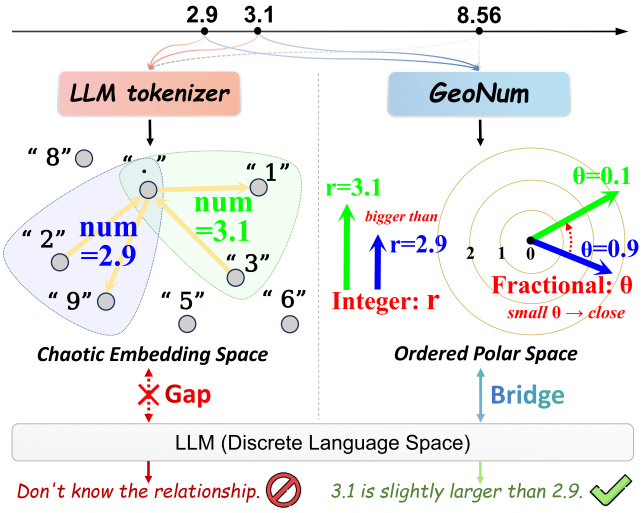

Figure 1: Comparison of tokenization approaches. Traditional methods disrupt geometric relationships between numerically close values. GeoNum preserves numerical continuity through polar coordinate decomposition.

Recent evaluations expose the severity of this representational mismatch. Even state-of-the-art models struggle with basic numerical tasks, often performing near chance levels on numerical comparisons (Li et al. 2025; Mirzadeh et al. 2024). Mathematics-specialized models solve complex word problems yet fail catastrophically on elementary arithmetic (Shao et al. 2024; Yu et al. 2024), revealing reliance on pattern matching rather than genuine numerical understanding (Hu et al. 2024). The root cause lies in tokenization's failure to preserve continuous structure. GPT-2's tokenizer assigns unique representations to only 916 of the first 10,000 integers (Singh and Strouse 2024), creating geometrically inconsistent embeddings where numerically close values scatter randomly in representation space. Without distance preservation, models cannot leverage their inductive bias for smoothness to interpolate between known values or extrapolate to unseen ranges.

Inspired by vision-language alignment that bridges continuous visual features with discrete text (Radford et al. 2021; Liu et al. 2023), we address the numerical-textual gap

*Corresponding author

through geometric coherence. We introduce **GeoNum**, a geometrically coherent numerical embedding based on polar coordinate decomposition. As Figure 1 shows, polar coordinates naturally preserve the dual structure that tokenization disrupts. Numbers like 2.9 and 3.1 map to similar radii with small angular differences in their fractional components, maintaining geometric proximity. GeoNum encodes integer magnitudes through classification learning while representing fractional components via trigonometric regression on the unit circle, constructing a continuous manifold where numerical relationships become geometric relationships.

Our three-stage framework progressively integrates GeoNum into pretrained language models. Stage I establishes the numerical manifold through self-supervised polar encoding. Stage II bridges discrepancies between numerical and textual embedding spaces via projection-based alignment. Stage III internalizes numerical capabilities through efficient adaptation. This progression enables models to develop geometric intuition for numerical relationships rather than relying on symbolic pattern matching, demonstrating promising capabilities in transforming language models toward principled numerical understanding (Bengio, Courville, and Vincent 2013; Fefferman, Mitter, and Narayanan 2016). Our contributions are the following:

- We propose GeoNum, a geometrically coherent numerical embedding that decouples ordinal structure from continuous metric relationships through polar coordinate decomposition, constructing a topologically consistent manifold for unified numerical learning.
- A three-stage framework progressively aligns continuous numerical representations with discrete textual semantics through self-supervised polar encoding, projection learning, and efficient adaptation.
- Experiments demonstrate substantial improvements in high-precision arithmetic tasks, with distance preservation analysis confirming that geometric continuity enables effective numerical interpolation and extrapolation.

Related Work

Numerical Capability Enhancement in LLMs. Recent benchmarks reveal significant deficiencies in numerical capabilities of advanced language models, particularly in arithmetic and magnitude comparison tasks (Yang et al. 2025; Li et al. 2025). While prompting techniques including Chain-of-Thought (Wei et al. 2022) and external tools such as PAL (Gao et al. 2023) and ReAct (Yao et al. 2022) have demonstrated improvements in arithmetic accuracy, these approaches do not fundamentally enhance models' internal numeric representations (Chen et al. 2022). Specialized training methods, including curriculum-based fine-tuning (Lewkowycz et al. 2022) and intermediate computation generation (Nye et al. 2021), yield notable improvements yet rely heavily on task-specific supervision, leaving inherent numeracy issues unresolved.

Structured Numerical Representations. Conventional subword tokenization severely disrupts numerical continuity, motivating development of structured numeric embeddings. While digit-level tokenization addresses representation completeness, it places the burden of digit aggregation entirely on models (Wallace et al. 2019). Deterministic embeddings like DICE (Sundararaman et al. 2020) explicitly encode numeric distances, preserving magnitude relations. Learned continuous embeddings, notably xVal (Golkar et al. 2023) and Fourier-based FoNE (Zhou et al. 2025), represent numbers through single-token vectors achieving high arithmetic precision. Methods including Abacus (McLeish et al. 2024) and NumeroLogic (Schwartz et al. 2024) similarly demonstrate superior numeric generalization, yet typically lack direct semantic alignment with linguistic embeddings.

Representational Alignment. Aligning continuous embeddings with discrete textual representations poses a persistent challenge within token-based large language models. Classical frameworks such as CLIP leverage contrastive pretraining to align continuous visual signals with language semantics (Radford et al. 2021), while subsequent approaches including LLaVA (Liu et al. 2023) and MiniGPT-4 (Zhu et al. 2024) employ lightweight projection layers to bridge pretrained vision encoders with language models. Flamingo and PalM-E further demonstrate that pretrained LLMs can efficiently integrate diverse continuous modalities through minimal adapters (Alayrac et al. 2022; Driess et al. 2023). Despite these advances in vision–language alignment, analogous efforts toward aligning numerical and textual representations remain largely unexplored.

Methodology

Figure 2 illustrates the GeoNum framework: polar encoding decomposes numbers into ordinal and metric components, projection alignment bridges numerical and textual spaces, and efficient adaptation integrates into pretrained LLMs.

Geometric Continuity Hypothesis. Standard tokenization functions as a discrete mapping $\tau : \mathbb{R} \to \mathcal{V}$ that violates local topology, where infinitesimal numerical perturbations δ often yield orthogonal representation vectors, i.e., $\langle \tau(x), \tau(x+\delta) \rangle \approx 0$. To bridge this representational gap, we posit that numerical embedding Φ must construct a smooth manifold satisfying the Lipschitz continuity condition:

$$\|\Phi(x) - \Phi(y)\|_2 \leq K \cdot |x - y|, \quad \forall x, y \in \mathbb{R}, \qquad (1)$$

where K bounds the geometric distortion. This constraint ensures that the induced metric aligns with the scalar field's Euclidean topology, allowing gradient descent to leverage numerical proximity as a valid geometric inductive bias.

Polar Decomposition

Traditional tokenization treats numerically similar values as unrelated symbol sequences, disrupting the geometric structure essential for mathematical reasoning. We address this through polar coordinate decomposition that naturally separates numbers into complementary components aligned with distinct learning paradigms. Formally, any real number admits a canonical decomposition:

$$x = s \cdot \left(\sum_{i=0}^{N-1} d_i \cdot 10^i + f \right), \qquad (2)$$

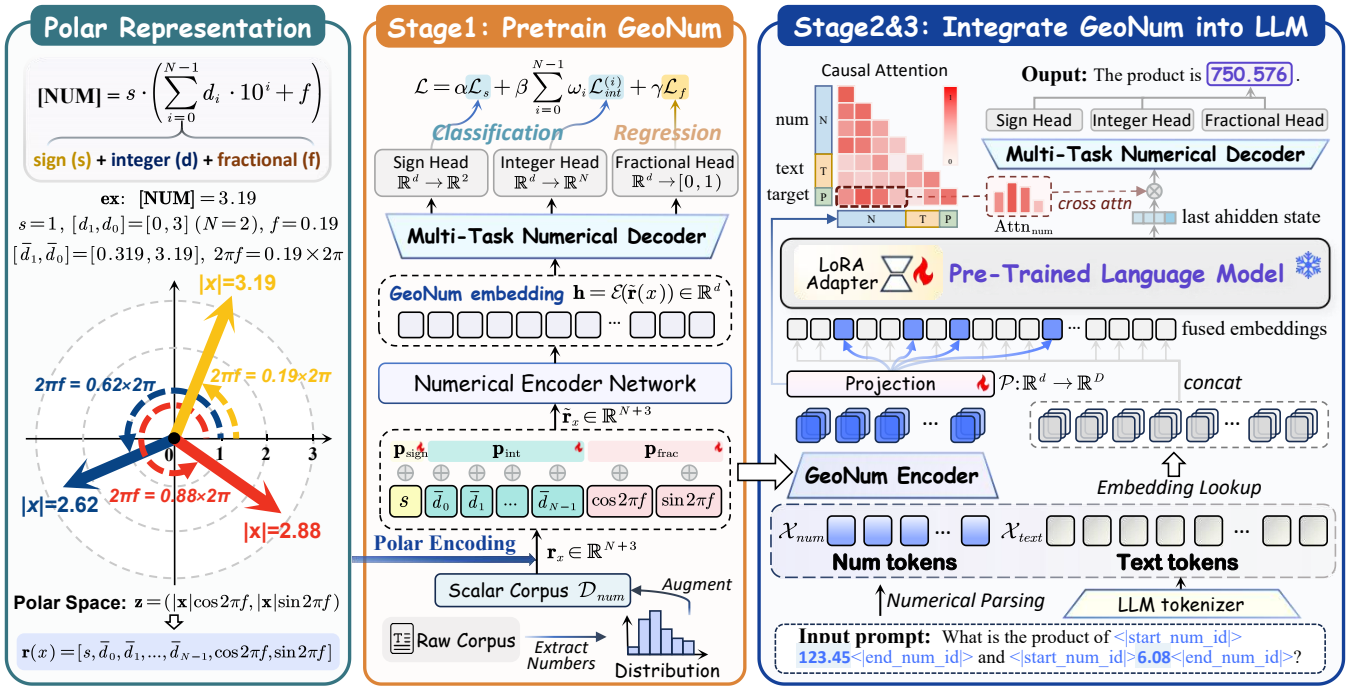

Figure 2: Overview of the GeoNum framework. **Left:** Polar decomposition separates numbers into hierarchical magnitudes and trigonometric fractions. **Middle:** Stage I pretrains the numerical encoder via joint classification-regression. **Right:** Stages II-III progressively integrate GeoNum into pretrained LLMs through projection alignment and efficient adaptation.

where $s \in \{-1, +1\}$ denotes sign, $d_i \in \{0, \ldots, 9\}$ are positional digits, and $f \in [0, 1)$ captures fractional precision. This decomposition exposes numerical cognition's dual nature where discrete digits exhibit ordinal structure suited for classification, while continuous fraction possesses metric properties amenable to regression. Standard tokenization conflates these fundamentally different aspects into arbitrary sub-tokens, preventing models from learning appropriate inductive biases for each component.

Trigonometric Fractional Encoding. The fractional component requires special treatment to preserve continuity across decimal boundaries. Linear encoding fails at boundary conditions where values like $f = 0.99$ and $f = 0.01$ are separated by Euclidean distance 0.98 despite being numerically adjacent. We resolve this through unit circle embedding:

$$\phi(f) = (\cos 2\pi f, \sin 2\pi f) \in \mathbb{S}^1. \tag{3}$$

This trigonometric representation maps $[0, 1)$ onto a continuous manifold where boundary discontinuities vanish. The circular distance between $\phi(0.99)$ and $\phi(0.01)$ becomes $2\sin(\pi \cdot 0.02) \approx 0.13$, accurately reflecting their true proximity. The unit circle naturally handles wraparound through its topological structure while maintaining differentiability everywhere, enabling gradient-based learning of periodic fractional patterns without artificial boundary artifacts.

Hierarchical Magnitude Encoding. We encode complete magnitude information at each digit position through contextual augmentation. Rather than treating digits as isolated symbols, we construct magnitude-aware representations:

$$\bar{d}_i = d_i + \frac{R_i + f}{10^i}, \tag{4}$$

where $R_i = \lfloor |x| \rfloor \bmod 10^i$ captures all digit positions below i. This creates a hierarchical encoding where each $\bar{d}_i$ simultaneously represents both its coarse positional value d_i and the complete fine-grained sub-magnitude beneath it through the fractional term $(R_i + f)/10^i$. For instance, consider $x = 3456.78$. The thousand-position encoding becomes $\bar{d}_3 = 3 + 456.78/1000 = 3.45678$, explicitly incorporating the full lower-order magnitude. Similarly, $\bar{d}_2 = 4.5678$, $\bar{d}_1 = 5.678$, and $\bar{d}_0 = 6.78$. Each $\bar{d}_i$ thus carries multi-scale information, enabling the model to understand about numerical quantities at varying granularities from individual digits to complete values. The complete polar representation integrates these components:

$$\mathbf{r}(x) = [s, \bar{d}_0, \bar{d}_1, \ldots, \bar{d}_{N-1}, \cos 2\pi f, \sin 2\pi f]^\top \in \mathbb{R}^{N+3}. \tag{5}$$

To distinguish semantic roles during learning, we augment with learnable type embeddings:

$$\tilde{\mathbf{r}}(x) = \mathbf{r}(x) + \mathbf{p}, \tag{6}$$

where $\mathbf{p} = [\mathbf{p}_{\text{sign}}, \mathbf{p}_{\text{int}}, \ldots, \mathbf{p}_{\text{int}}, \mathbf{p}_{\text{frac}}, \mathbf{p}_{\text{frac}}]^\top$. Since positional information is embedded in $\bar{d}_i$, all integer digits share $\mathbf{p}_{\text{int}}$, avoiding redundant position encoding.

Geometric Properties. The polar encoding satisfies the topological constraints defined in Eq. 1. Specifically, for any

local perturbation satisfying $|x-y| < \epsilon$, the feature distance is strictly bounded by $\|\mathbf{r}(x) - \mathbf{r}(y)\|_2 \leq \mathcal{O}(\sqrt{N} \cdot \epsilon)$. This continuity is underpinned by two structural mechanisms: the hierarchical digit encoding scales positional contributions by $\mathcal{O}(10^{-i})$, while the trigonometric component ϕ maps fractional values onto the unit circle $\mathbb{S}^1$, effectively eliminating boundary discontinuities via sinusoidal smoothness. Consequently, the embedding establishes a robust isometry between the scalar field and the feature manifold, guaranteeing that numerical proximity in $\mathbb{R}$ is faithfully preserved as geometric proximity in $\mathbb{R}^d$.

Three-Stage Integration

We integrate GeoNum into pre-trained language models via a progressive framework that establishes a continuous numerical manifold, aligns it with textual semantics, and enables efficient adaptation for high-precision arithmetic tasks.

Stage I: Pretraining GeoNum. The first stage initiates the learning of continuous numerical representations through self-supervised reconstruction. We utilize a scalar corpus $\mathcal{D}_{num}$ constructed by extracting and augmenting numerical distributions from raw text. Specifically, a numerical encoder $\mathcal{E}$ transforms polar inputs into latent embeddings $\mathbf{h} = \mathcal{E}(\tilde{\mathbf{r}}(x)) \in \mathbb{R}^d$, which concurrently drive three task-specific decoding heads:

$$\mathbf{p}_s(\mathbf{h}) = \text{softmax}(\mathbf{W}_s\mathbf{h} + \mathbf{b}_s) \in \mathbb{R}^2, \tag{7}$$

$$\mathbf{p}_{int}(\mathbf{h}) = \text{softmax}(\mathbf{W}_{int}\mathbf{h} + \mathbf{b}_{int}) \in \mathbb{R}^N, \tag{8}$$

$$\hat{f}(\mathbf{h}) = \sigma(\mathbf{W}_f\mathbf{h} + \mathbf{b}_f) \in [0, 1), \tag{9}$$

where σ denotes the sigmoid activation function. The projection matrices $\{\mathbf{W}_s, \mathbf{W}_{int}, \mathbf{W}_f\}$ and bias vectors $\{\mathbf{b}_s, \mathbf{b}_{int}, \mathbf{b}_f\}$ parameterize the binary sign classifier, the N-dimensional integer magnitude estimator, and the fractional regressor, respectively.

A notable feature of our framework is the encoding-decoding asymmetry. While the encoder projects inputs onto the unit circle via $(\cos 2\pi f, \sin 2\pi f)$ to enforce boundary continuity, the decoder regresses the scalar f to facilitate stable end-to-end gradient flow. Consequently, the circular topology is implicitly induced through representation learning rather than imposed by hard constraints. This design synergizes the geometric advantages of circular embeddings with the optimization stability of scalar regression.The unified training objective combines these components:

$$\mathcal{L} = \alpha\mathcal{L}_s + \beta \sum_{i=0}^{N-1} \omega_i \mathcal{L}_{int}^{(i)} + \gamma\mathcal{L}_f, \tag{10}$$

where $\mathcal{L}_s$ and $\mathcal{L}_{int}$ minimize cross-entropy loss for categorical predictions, while $\mathcal{L}_f$ utilizes mean squared error for fractional regression. Crucially, the positional weights $\omega_i = 1 + 0.2i$ explicitly amplify the significance of higher-order digits to capture magnitude hierarchy. This composite objective drives the simultaneous learning of ordinal topology via classification and metric continuity via regression, yielding an embedding manifold where numerical closeness is faithfully preserved as geometric proximity. Post-pretraining, the learned encoder $\mathcal{E}$ is frozen for alignment.

Stage II: Numerical-Textual Alignment. The second stage bridges numerical and textual representations through projection learning. A lightweight network $\mathcal{P} : \mathbb{R}^d \to \mathbb{R}^D$ maps frozen numerical embeddings into the LLM's semantic space where D denotes the model's hidden dimension. During processing, numerical placeholders $\langle\text{NUM}\rangle$ in input sequences are replaced with projected embeddings, constructing numerical–textual hybrid representations:

$$\mathbf{X} = [\mathbf{e}_1^{text}, \ldots, \mathcal{P}(\mathbf{h}_{num}), \ldots, \mathbf{e}_n^{text}], \tag{11}$$

where $\mathbf{e}_i^{text}$ denote standard textual embeddings. The LLM processes this unified sequence, producing contextualized representations. At numerical output positions, we extract hidden states $\mathbf{s}_{num}$ from the final layer and decode through a unified head:

$$[\hat{s}, \hat{d}_0, \ldots, \hat{d}_{N-1}, \hat{f}] = \mathcal{D}(\mathbf{s}_{num}), \tag{12}$$

where $\mathcal{D} : \mathbb{R}^D \to \mathbb{R}^{2+10N+1}$ produces logits for all components simultaneously. Training optimizes projection $\mathcal{P}$ and decoder $\mathcal{D}$ using the loss from Equation (5), maintaining geometric structure through the LLM's processing pipeline.

Unlike vision-language approaches that typically rely on contrastive learning to discover correspondences, our method leverages the inherent structure of polar decomposition. This allows the projection to directly preserve distance relationships without the need for implicit matching.

Stage III: End-to-End Fine-Tuning. The final stage internalizes numerical capabilities through efficient adaptation. We apply Low-Rank Adaptation to attention projection matrices via trainable decompositions $W = W_0 + BA$ where low-rank factors $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ adapt frozen weights W_0. This approach preserves existing textual capabilities while enabling numerical processing.

To integrate numerical embeddings with textual context, the decoder employs cross-attention over GeoNum representations. At output positions, queries $\mathbf{q}$ from the LLM attend to numerical embeddings:

$$\mathbf{s}_{num} = \sum_j \frac{\exp(\mathbf{q}^\top \mathbf{k}_j/\sqrt{d})}{\sum_{j'} \exp(\mathbf{q}^\top \mathbf{k}_{j'}/\sqrt{d})} \mathbf{v}_j, \tag{13}$$

where keys and values $\{\mathbf{k}_j, \mathbf{v}_j\}$ derive from projected embeddings $\mathcal{P}(\mathbf{h}_{num})$. This enables selective integration of numerical information conditioned on linguistic context.

Through joint optimization of projection $\mathcal{P}$ and LoRA parameters, the model learns to perform arithmetic and comparison via geometric transformations in the learned manifold rather than symbolic manipulation, effectively integrating geometric intuition into its operational logic.

Experiments

We evaluate GeoNum across complementary arithmetic benchmarks, highlighting significant gains in fine-grained numerical precision. Extensive analyses on ablation , distance preservation , and manifold evolution further validate the necessity of the proposed polar geometric continuity.

Method	NumericBench					NUPA					FERMAT				
	A@5	A@1		RE	MAE	A@5	A@1	A@0.1	RE	MAE	A@5	A@1	A@0.1	RE	MAE
LLaMA-3.2-1B															
Direct	0.85	0.62	0.32	0.24	52.25	0.72	0.37	0.09	0.88	93.23	0.50	0.32	0.22	3.86	333.86
CoT	0.87	0.72	0.35	0.15	38.07	0.78	0.49	0.15	0.65	49.74	0.31	0.23	0.18	0.90	155.11
Fine-tuning	0.94	0.88	0.60	0.04	6.78	0.86	0.40	0.29	0.52	56.12	**0.69**	**0.52**	0.29	0.93	183.25
xVal	0.89	0.73	0.47	0.09	17.65	0.82	0.36	0.22	1.13	94.83	0.37	0.16	0.12	2.06	256.34
NumeroLogic	0.97	0.90	0.61	**0.03**	2.55	0.87	0.33	0.21	0.77	96.36	0.61	0.42	0.28	1.02	168.77
FoNE	0.90	0.81	0.53	0.06	35.31	**0.97**	0.47	0.29	0.66	104.51	0.25	0.15	0.13	3.02	156.04
GeoNum	**0.99**	**0.94**	**0.68**	0.05	**0.79**	0.96	**0.50**	**0.35**	**0.51**	**42.45**	0.66	0.39	**0.32**	**0.87**	**89.07**
LLaMA-3.2-3B															
Direct	0.87	0.69	0.35	0.20	14.08	0.81	0.45	0.23	0.74	95.11	0.53	0.38	0.21	1.06	246.91
CoT	0.90	0.80	0.33	0.05	7.97	0.79	0.56	0.21	0.68	54.91	0.44	0.46	0.42	0.88	148.39
Fine-tuning	0.95	0.86	0.62	0.03	3.14	0.88	0.65	0.35	0.39	36.04	**0.82**	0.48	0.39	0.76	57.43
xVal	0.88	0.71	0.45	0.06	1.60	0.97	0.62	0.25	0.49	94.08	0.45	0.23	0.19	2.04	158.32
NumeroLogic	**0.99**	0.92	0.66	0.03	8.67	0.97	0.60	0.24	0.59	32.96	0.78	0.52	0.36	0.61	59.68
FoNE	0.93	0.84	0.59	0.04	8.52	**0.99**	0.64	0.32	**0.27**	37.11	0.38	0.19	0.15	2.16	196.12
GeoNum	**0.99**	**0.95**	**0.71**	**0.02**	**0.63**	0.98	**0.71**	**0.52**	0.32	**24.09**	0.74	**0.66**	**0.45**	**0.58**	**44.24**

Table 1: Main Results on Arithmetic Benchmarks. We report performance across three datasets using LLaMA-3.2 1B and 3B models. Best results in each category are bolded.

Experimental Setup

Datasets. We evaluate on three complementary arithmetic benchmarks assessing distinct dimensions of numerical capability. **NumericBench** (Li et al. 2025) assesses fundamental numerical abilities through basic four-operation calculations with decimal numbers. We utilize the Arithmetic Operation subset for core computational evaluation. **NUPA** (Yang et al. 2025) examines comprehensive numerical understanding across multiple representations. We select the Float-ADD subset for high-precision multi-digit decimal computations challenging numerical accuracy at scale. **FERMAT** (Sivakumar and Moosavi 2023) targets numerical precision through adversarial examples. We employ the 2dp-random processed decimal dataset to test generalization beyond training distributions. We filter samples to retain values within $(-10^5, 10^5)$ covering 99.8% of benchmark distributions while enabling fixed 5-digit representation. Final datasets comprise NumericBench (7.2k samples), NUPA (7.3k samples), and FERMAT (5.5k samples).

Baselines. We compare against representative approaches across three categories. **Prompting methods** include Direct Inference without additional training and Chain-of-Thought (Wei et al. 2022) with step-by-step prompts. **Fine-tuning approach** applies LoRA adaptation on target datasets without specialized numerical representations. **Numerical encoding techniques** include xVal (Golkar et al. 2023) employing multiplicative scaling for end-to-end continuity, NumeroLogic (Schwartz et al. 2024) prefixing digit counts for positional information, and FoNE (Zhou et al. 2025) leveraging Fourier features exploiting internal Fourier-like representations. All methods use LLaMA-3.2 1B and 3B models.

Input-Output Format. Numerical values are marked with special tokens ⟨start_num_id⟩ ⟨NUM⟩ ⟨end_num_id⟩ where ⟨NUM⟩ receives method-specific representations. Models generate single numerical answers via decoder-only architecture. During training, we extract final-layer hidden states from answer token positions and decode through numerical prediction heads. Inference follows identical processing.

Evaluation Metrics. We employ both threshold-based accuracy and regression metrics. **ACC@τ** measures predictions within relative error threshold $\tau \in \{5\%, 1\%, 0.1\%\}$. ACC@0.1 specifically tests geometric continuity by requiring high precision. **Relative Error (RE)** computes $|y - \hat{y}|/(|y| + \epsilon)$ with $\epsilon = 10^{-8}$ preventing division by zero, clipped at 10.0 for stability. **Mean Absolute Error (MAE)** quantifies absolute deviation, percentile-capped at the 95th to prevent outlier dominance while preserving error signal.

Main Results

Table 1 presents performance across three arithmetic benchmarks with LLaMA-3.2 models at 1B and 3B scales. GeoNum consistently achieves superior results, particularly on metrics requiring high numerical precision.

High-Precision Performance. GeoNum's strongest advantages emerge at the ACC@0.1 threshold. On NUPA with 3B parameters, GeoNum reaches 0.52 compared to fine-tuning's 0.35, representing 48.6% relative improvement. The pattern holds across datasets: NumericBench shows GeoNum at 0.71 versus NumeroLogic's 0.66 despite comparable coarse accuracy, while FERMAT demonstrates 0.45 versus fine-tuning's 0.39. This precision advantage reflects

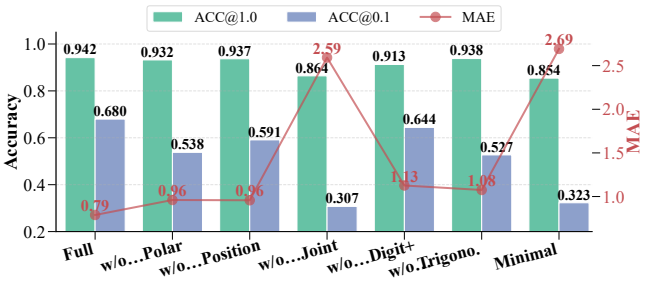

Figure 3: Ablation study on NumericBench. Bar plots show ACC@1.0 and ACC@0.1; line plot indicates MAE.

geometric continuity where polar decomposition preserves distance relationships, enabling accurate interpolation unavailable to discrete tokenization methods.

Dataset Characteristics. NumericBench favors explicit structural encoding, with both GeoNum and NumeroLogic achieving strong coarse accuracy through positional information. However, GeoNum maintains superiority as precision requirements increase, while NumeroLogic plateaus. NUPA's multi-digit decimal operations expose greater performance gaps. FoNE achieves 0.99 at ACC@5 on 3B through Fourier periodicity but drops to 0.32 at ACC@0.1, indicating local smoothness captures coarse patterns without fine-grained fidelity. FERMAT's adversarial scenarios prove most challenging for all baselines. Fine-tuning reaches 0.39 on ACC@0.1 with 3B parameters, whereas GeoNum achieves 0.45, demonstrating that geometric structure generalizes more effectively to distribution shifts.

Precision Degradation. Performance decay from ACC@5 to ACC@0.1 reveals representational differences. Traditional methods exhibit severe degradation. FoNE on NUPA-3B drops from 0.99 to 0.32, while fine-tuning on NumericBench-1B decreases from 0.94 to 0.60. In contrast, GeoNum demonstrates markedly better robustness, declining from 0.99 to 0.71 on NumericBench-3B and retaining substantially higher performance ratios across benchmarks. This pattern validates that continuous manifold representations preserve numerical fidelity under stringent precision requirements.

Scaling Effects. Increasing parameters from 1B to 3B benefits all methods while preserving GeoNum's advantages. On NumericBench ACC@0.1, the absolute gap versus fine-tuning remains stable around 0.08, while on NUPA, GeoNum's advantage amplifies from 0.06 to 0.17 as model size increases. MAE improvements further demonstrate scaling benefits. GeoNum reduces error from 0.79 to 0.63 on NumericBench, while fine-tuning improves from 6.78 to 3.14. This indicates that geometric representations may facilitate more effective utilization of additional model capacity for numerical precision tasks.

Numerical Range Generalization

We evaluate extrapolation by training on $[-50, 50]$ and testing on larger ranges. Table 2 shows systematic performance degradation across all methods as ranges expand, with GeoNum demonstrating superior robustness and maintaining consistent advantage. Polar decomposition enables structured extrapolation through explicit ordinal and continuous components, whereas tokenization fragments at unseen scales without interpolation guidance.

Range	Direct	FT	xVal	FoNE	GeoNum
[-200, 200]	0.147	0.218	0.284	0.309	**0.356**
[-1K, 1K]	0.082	0.139	0.192	0.227	**0.274**
[-5K, 5K]	0.041	0.073	0.118	0.164	**0.213**

Table 2: Numerical Range Generalization

Ablation Study

We dissect GeoNum's component contributions on NumericBench through ablation experiments shown in Figure 3.

Removing joint discrete-continuous training causes severe degradation, with ACC@0.1 dropping from 0.680 to 0.307 and MAE increasing from 0.79 to 2.59. This validates the necessity of unified modeling for numerical cognition's dual nature. Excluding polar encoding reduces ACC@0.1 to 0.538 with MAE rising to 0.96, while removing trigonometric fractional encoding decreases ACC@0.1 to 0.527. Both components prove essential for high-precision arithmetic, with trigonometric functions maintaining boundary continuity critical for fractional representations. While positional and enhanced digit encodings produce minor drops when removed individually, indicating they serve as secondary refinements rather than fundamental structural components. Their modest contributions nevertheless stabilize representations under stringent precision requirements.

Distance Preservation Analysis

We analyze how encoding methods preserve numerical distance relationships by training 10,000 samples from $[0, 10^4]$ and tracking embedding evolution across training epochs.

Geometric Structure. Figure 4 (left) reveals differences in learned representations. GeoNum achieves near-ideal preservation with Spearman correlation 0.97, exhibiting tight clustering along the diagonal where normalized numerical distances directly correspond to embedding distances. This geometric alignment confirms successful manifold learning where numerically proximate values maintain spatial proximity. FoNE exhibits moderate correlation 0.95 with visible scatter, capturing local smoothness through Fourier features while lacking hierarchical structure. LLM Digit produces chaotic distribution at 0.5, confirming discrete tokenization disrupts continuous relationships.

Training Dynamics. As illustrated in Figure 4 (Right), GeoNum exhibits efficient convergence, with distance MAE vanishing to near-zero and Spearman and Pearson coefficients simultaneously saturating (> 0.95) by epoch 10. This synchronization indicates that the embedding successfully unifies monotonic ordering with approximate linear mapping, effectively establishing a metric space isometric to the

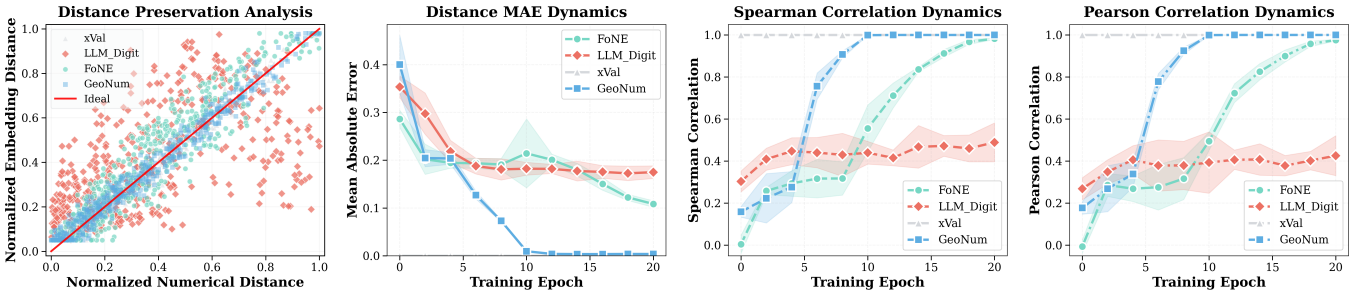

Figure 4: Distance preservation analysis. Left: Scatter plots visualizing the alignment between normalized numerical distances and embedding distances at convergence. Right: Temporal evolution of geometric fidelity, tracking the rapid convergence of Distance MAE and the saturation of both Spearman (ordinal) and Pearson (linear) correlations across training epochs.

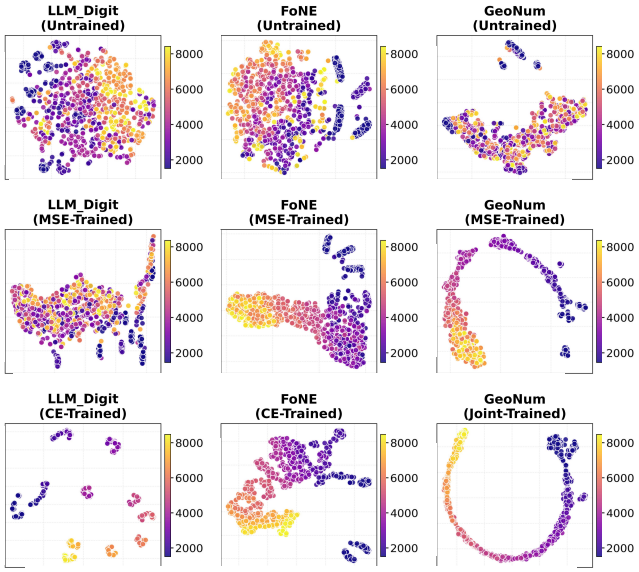

Figure 5: t-SNE visualization of embedding manifolds. Unlike the fragmented clusters in baselines, GeoNum evolves into a topologically continuous and ordinally structured manifold, validating its geometric inductive bias.

scalar field. In contrast, FoNE shows sluggish optimization due to Fourier learning complexity. While xVal maintains high correlation, this stems from trivial multiplicative scaling that compromises fractional topology, elucidating its inferior high-precision performance observed in Table 1.

Geometric Structure Validation

We validate geometric continuity through progressive training analysis using 20,000 samples across three training paradigms: *(i)* random initialization, *(ii)* MSE reconstruction, and *(iii)* classification-based learning with joint classification-regression for GeoNum.

Embedding Space Evolution. Figure 5 reveals how different architectures develop numerical structure under varying training objectives Post-MSE training, LLM Digit develops fragmented clusters reflecting tokenization boundaries rather than numerical relationships. FoNE creates inconsistent patterns without clear ordering. GeoNum establishes coherent geometric structures preserving numerical proximity. Under joint classification-regression training, GeoNum produces embeddings with systematic ordinal organization, confirming polar decomposition provides effective inductive bias, enabling structured manifold learning.

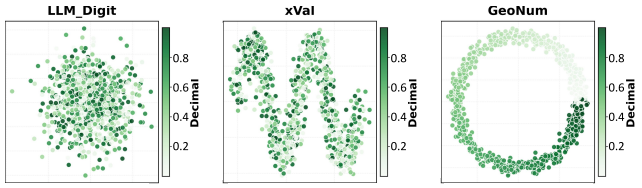

Figure 6: t-SNE visualization of fractional topology. GeoNum forms a closed circular manifold that preserves periodicity and boundary continuity

Fractional Periodicity Structure. Figure 6 visualizes fractional component organization via t-SNE projections colored by decimal magnitude. LLM Digit yields chaotic scatter, confirming that discrete tokenization disrupts intrinsic fractional structure. While xVal exhibits linear arrangements, it lacks the topological closure required for boundary continuity. In contrast, GeoNum establishes a coherent circular manifold, where the trigonometric representation naturally maps values onto the unit circle. This geometric structure explicitly preserves periodicity across decimal boundaries, as evidenced by the observed smooth color gradients confirming that numerical proximity is faithfully maintained for high-precision interpolation and numerical inference.

Conclusion

We introduce GeoNum to bridge the discrepancy between discrete tokenization and numerical continuity. By leveraging polar decomposition, GeoNum constructs a continuous manifold that harmonizes ordinal magnitude with periodic precision. Experiments demonstrate that this geometric coherence yields substantial gains in high-precision arithmetic, interpolation, and extrapolation. These findings validate geometric continuity as a critical inductive bias, offering a promising directions for reconciling symbolic semantics with continuous numerical cognition in future LLMs.

Acknowledgments

This work was supported by the National Science and Technology Major Project (No. 2022ZD0117800), the State Grid Corporation of China project on "Multi-Source Data Perception and Fusion Analysis Methods for Power System", and the CAAI-MindSpore Open Fund, with development on the OpenI Community platform.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *International Conference on Machine Learning*, 8469–8488. PMLR.

Fefferman, C.; Mitter, S.; and Narayanan, H. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049.

Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; and Neubig, G. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, 10764–10799. PMLR.

Golkar, S.; Pettee, M.; Eickenberg, M.; Bietti, A.; Cranmer, M.; Krawezik, G.; Lanusse, F.; McCabe, M.; Ohana, R.; Parker, L.; et al. 2023. xval: A continuous number encoding for large language models. *arXiv preprint arXiv:2310.02989*.

Hu, Y.; Tang, X.; Yang, H.; and Zhang, M. 2024. Case-Based or Rule-Based: How Do Transformers Do the Math? In *International Conference on Machine Learning*, 19438–19474. PMLR.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35: 3843–3857.

Li, H.; Chen, X.; Xu, Z.; Li, D.; Hu, N.; Teng, F.; Li, Y.; Qiu, L.; Zhang, C. J.; Qing, L.; et al. 2025. Exposing numeracy gaps: A benchmark to evaluate fundamental numerical abilities in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 20004–20026.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

McLeish, S.; Bansal, A.; Stein, A.; Jain, N.; Kirchenbauer, J.; Bartoldson, B.; Kailkhura, B.; Bhatele, A.; Geiping, J.; Schwarzschild, A.; et al. 2024. Transformers can do arithmetic with the right embeddings. *Advances in Neural Information Processing Systems*, 37: 108012–108041.

Mirzadeh, I.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; and Farajtabar, M. 2024. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *arXiv e-prints*, arXiv–2410.

Nye, M.; Andreassen, A. J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; et al. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. *arXiv preprint arXiv:2112.00114*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Schwartz, E.; Choshen, L.; Shtok, J.; Doveh, S.; Karlinsky, L.; and Arbelle, A. 2024. NumeroLogic: Number Encoding for Enhanced LLMs' Numerical Reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 206–212.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Singh, A. K.; and Strouse, D. 2024. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*.

Sivakumar, J.; and Moosavi, N. S. 2023. Fermat: An alternative to accuracy for numerical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15026–15043.

Sundararaman, D.; Si, S.; Subramanian, V.; Wang, G.; Hazarika, D.; and Carin, L. 2020. Methods for numeracy-preserving word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4742–4753.

Wallace, E.; Wang, Y.; Li, S.; Singh, S.; and Gardner, M. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5307–5315.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-

thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yang, H.; Hu, Y.; Kang, S.; Lin, Z.; and Zhang, M. 2025. Number Cookbook: Number Understanding of Language Models and How to Improve It. In *The Thirteenth International Conference on Learning Representations*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Yu, L.; Jiang, W.; Shi, H.; YU, J.; Liu, Z.; Zhang, Y.; Kwok, J.; Li, Z.; Weller, A.; and Liu, W. 2024. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Zhou, T.; Fu, D.; Soltanolkotabi, M.; Jia, R.; and Sharan, V. 2025. FoNE: Precise Single-Token Number Embeddings via Fourier Features. *arXiv preprint arXiv:2502.09741*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *ICLR*.