Numbers Already Carry Their Own Embeddings

Suhyun Bae

Master's Capstone Project Presentation

Department of Mathematics

November 27, 2025

AIML@K

KOREA UNIVERSITY

Motivation

- LLMs like “ChatGPT” often fail at basic arithmetic
- Even they win gold-medal in IMO

DeepMind and OpenAI models solve maths problems at level of top students

For the first time, large language models performed on a par with gold medallists in the International Mathematical Olympiad.

By Davide Castelvecchi

Models from OpenAI and DeepMind achieved gold-medal scores in the International Mathematical Olympiad. Credit: MoiraM/Alamy

Solve: 5.9 = x + 5.11

Let's solve step by step:

$$5.9 = x + 5.11$$

Subtract 5.11 from both sides:

$$5.9 - 5.11 = x$$

Compute carefully:

$$5.90 - 5.11 = 0.79 - 0.00 = -0.21$$

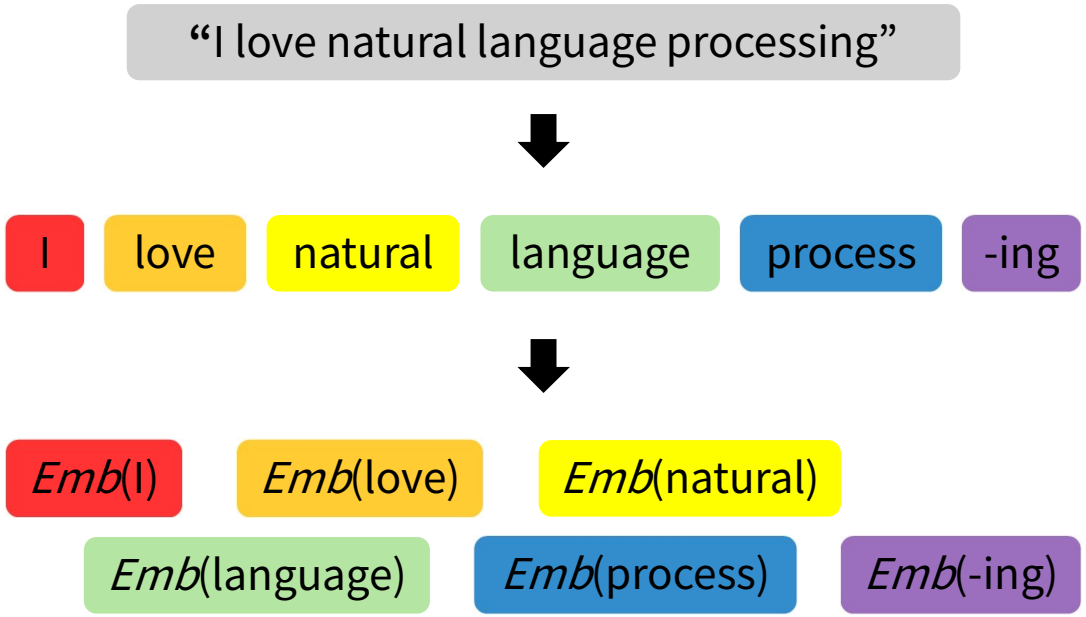
✅ **Answer:**

$$x = -0.21$$

AIML@K KOREA UNIVERSITY

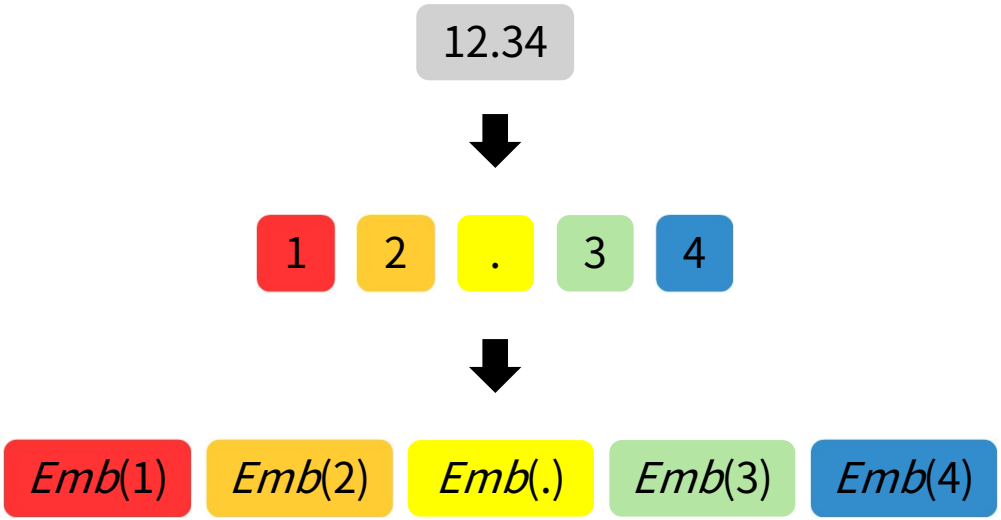
How AI See Language

- Tokenizer
  - Break down sentence into tokens
  - Each token is mapped to high-dim embeddings
    - Embeddings are randomly initialized
    - Adjusted through model training



Numbers as Token

- Numbers are treated as text
  - Lose original numeric meaning after tokenization and embedding
    - Arithmetic operations like addition and multiplication no longer hold



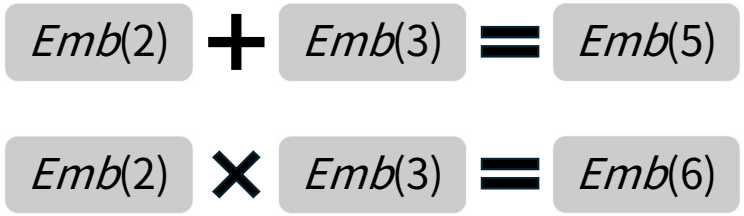
$$Emb(2) + Emb(3) \neq Emb(5)$$

$$Emb(2) \times Emb(3) \neq Emb(6)$$

Hypothesis

- Preserve algebraic structure in embedding space
  - Could fundamentally fix how LLMs interpret numbers
    - May eliminate the problematic behaviors of LLMs

→**How can we achieve this?**



What Embedding can satisfy this?

Adele Ring

Unify all perspective of number theory within single structure

- Let $\mathbb{Q}_p$ p-adic numbers and $\mathbb{Z}_p$ p-adic integers for prime p

  Then, "**Adele Ring**" $\mathbb{A}_{\mathbb{Q}}$ is defined as

$$\mathbb{A}_{\mathbb{Q}} \stackrel{\text{def}}{=} \mathbb{R} \times \prod_{p} (\mathbb{Q}_p, \mathbb{Z}_p)$$

- Let $A: \mathbb{Q} \to \mathbb{A}_{\mathbb{Q}}$. Then, for any rational q

$$A(q) = (q, q_2, q_3, \dots, q_p, \dots)$$

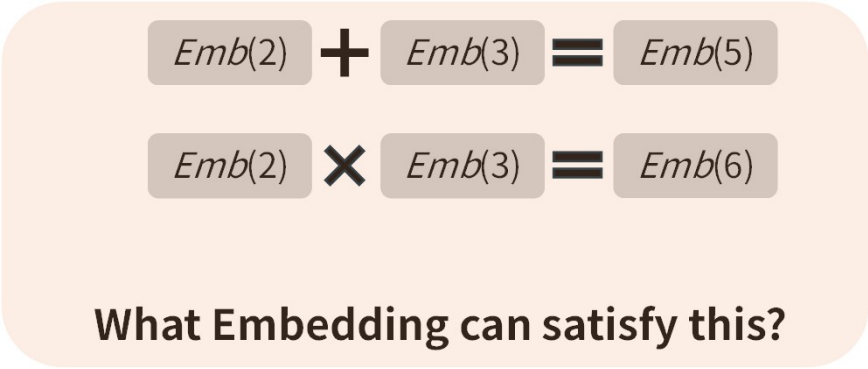
Idea

- Key features of Adele Ring
  - Additive and multiplicative

$$A(q_1 + q_2) = A(q_1) + A(q_2)$$

$$A(q_1 * q_2) = A(q_1) * A(q_2)$$

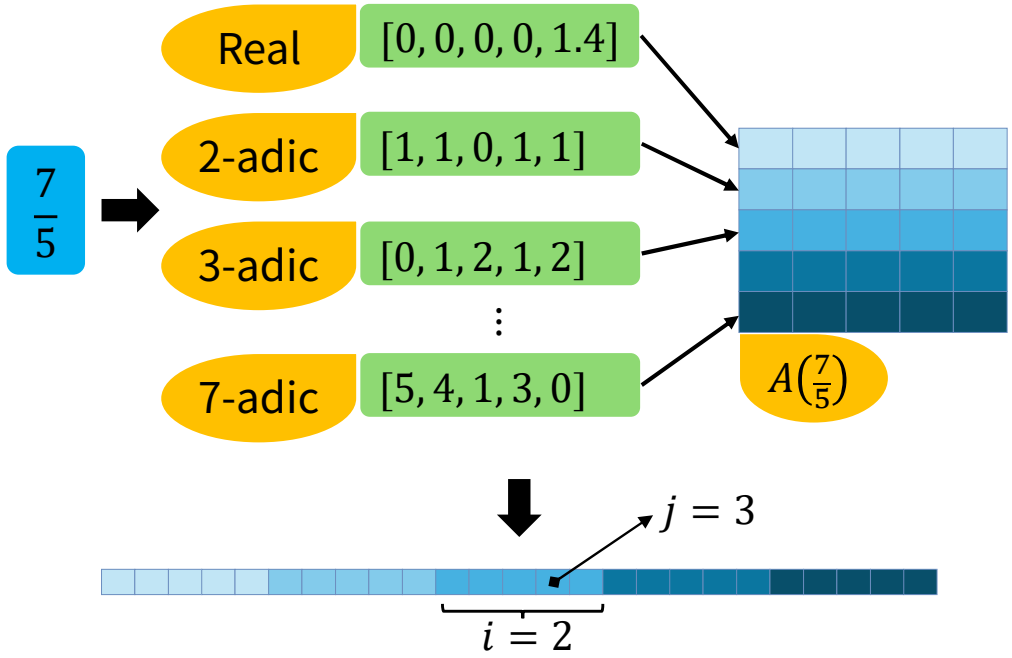
→Use Adele Ring for embedding



Look familiar?

Implementation

- AOE(Adelic Operation-preserved Embeddings)
  - Embedding module for inputs
    - N-digit precision for practical use
- 2D PE(Positional Encoding)
  - Encodes spatial information
    - Index i: Prime number assignment
    - Index j: Digit position in p-adic

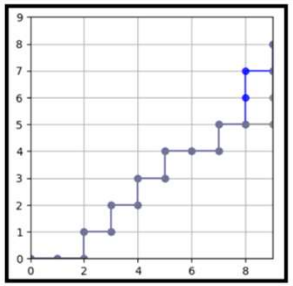


Experiment Setup

- Model
  - Architecture: 6-layer Transformer Encoder
  - **Baseline**: Trainable `nn.Embedding` + 1D PE module by Pytorch
  - **Ours**: Train-free AOE + 2D PE module
    - Embedding dimension size and total parameter number are same
- Every training applied auto hyperparameter search by `Optuna`
  - Ensures optimal output for each task

Experiment Setup

- **ACD(Algebraic Combinatorics Datasets)**
  - Research-level problem datasets in algebraic combinatorics
  - Task example
    - **Lattice Paths**: Predict whether a covering pair belongs to the matching order or Lagrange order



"The lattice path pair is a cover in the *Lagrange partial order*"

$$x = [1,1,0,1,0,1,0,1,0,1,1,0,1,0,0,1,0,0]$$
$$y = 1$$

Result

- AOE outperforms the baseline across all tasks
- Achieved 100% accuracy on “Weaving patterns”
  - Remarkable, considering this remains an open problem

Dataset		Accuracy(%) ↑	
Task	**n**	**Baseline**	**AOE(Ours)**
Lattice paths	10	66.19	**70.10**
	11	66.30	**66.30**
	12	66.50	**78.44**
Weaving patterns	6	53.02	**100.00**
	7	51.53	**100.00**
Quiver mutation classes		45.13	**91.11**
Grassmannian cluster algebras		91.27	**97.39**
Schubert polynomials	4	50.59	**87.89**
	5	49.83	**96.87**
	6	50.00	**99.96**
mHeight	8	91.42	**96.94**
	9	93.20	**99.17**
	10	94.15	**99.66**

Discussion

1. AOE requires dataset pre-processing, significantly increasing overall training time
   - 2-5 times slower per epoch than baseline
   - Techniques to reduce computational complexity is necessary
2. Implication of 100% accuracy
   - Investigate model's capacity to assist in formulating answers for open problems

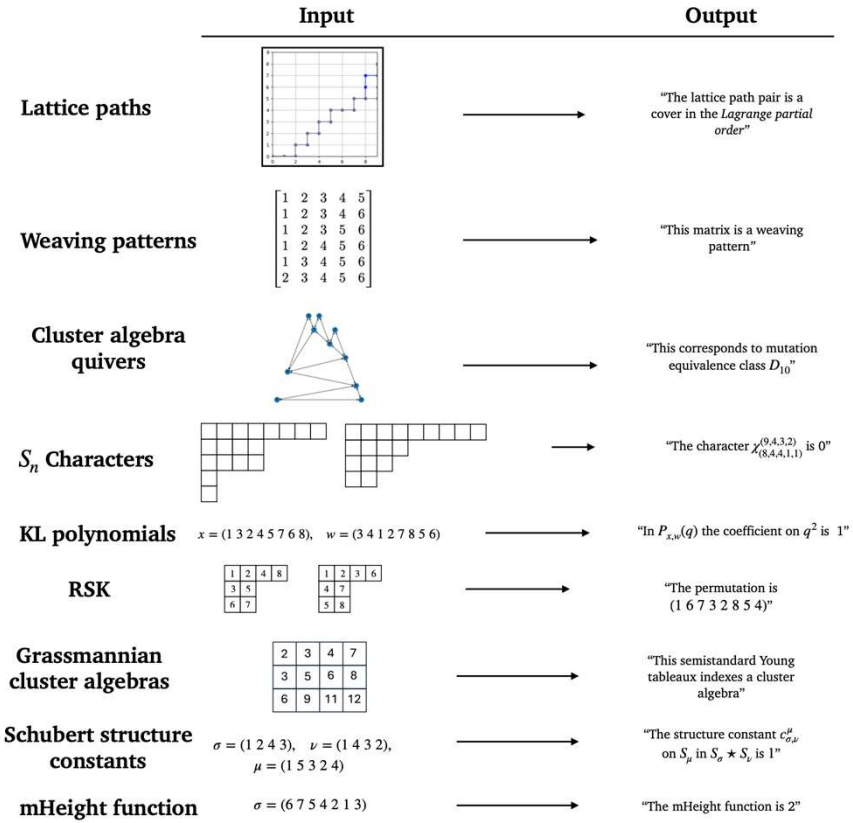
Future Direction

1. **Theoretical Verification**
   - Mathematical proofs for AOE's consistent superiority
2. **Methodological Expansion**
   - From rationals to number fields
3. **Integration Strategy**
   - Interoperability with existing text embeddings

Thank you for listening

Feel free to ask any question

Algebraic Combinatorics Datasets



AIML@K KOREA UNIVERSITY

Other Results

Dataset		Accuracy ↑						
Task	**n**	**Logistic regression**	**MLP**	**Transformer (Decoder)**	**Claude 3.5 Sonnet**	**GPT-4o Mini**	**GPT-4o**	**o1-Mini**
Lattice paths	10	66.2	**90.6**	65.3				
	11	66.3	**95.8**	69.4				
	12	66.5	**98.6**	86.2				
Weaving patterns	6	70.4	**86.1**	85.9				
	7	85.8	99.3	**99.9**				
Quiver mutation classes		40.3	86.5	**92.9**				
Grassmannian cluster algebras		65.7	99.3	**99.5**				
Schubert polynomials	4	88.8	93.1	**94.6**	59.5	53.5	57.0	
	5	90.6	**97.5**	96.2	58.5	51.5	57.0	
	6	89.7	**99.8**	91.3				
mHeight	8	91.4	99.4	**99.7**				
	9	93.2	99.8	**99.9**				
	10	94.2	**99.9**	**99.9**		89.5	95.5	95.4

Test Loss Table

Dataset		Loss ↓	
Task	**n**	**Baseline**	**AOE(Ours)**
Lattice paths	10	0.7045	**0.5957**
	11	0.6893	**0.6454**
	12	0.6845	**0.4658**
Weaving patterns	6	0.6910	**0.6897**
	7	0.6933	**0.0001**
Quiver mutation classes		0.9873	**0.2247**
Grassmannian cluster algebras		0.2422	**0.0827**
Schubert polynomials	4	0.6937	**0.3052**
	5	1.0910	**0.1093**
	6	0.8031	**0.0020**
mHeight	8	0.4740	**0.1190**
	9	0.4329	**0.0230**
	10	0.3998	**0.0105**

Training Details

- Optimizer and Objective
  - Adam optimizer with β values $(0.9, 0.999)$ and weight decay 0
  - Cross-Entropy Loss
- Batching and Sampling
  - All experiments use batch size 2048
  - `WeightedRandomSampler` to address class imbalance

Dataset			
Task	**n**	**Learning Rate**	**Epochs**
Lattice paths	10	7.7e-5	100
	11	1.0e-3	100
	12	1.0e-3	20
Weaving patterns	6	2.0e-5	100
	7	1.0e-4	100
Quiver mutation classes		8.8e-5	100
Grassmannian cluster algebras		6.0e-5	100
Schubert polynomials	4	1.7e-5	200
	5	8.0e-5	100
	6	5.0e-5	50
mHeight	8	3.0e-4	100
	9	6.0e-4	100
	10	7.3e-5	30