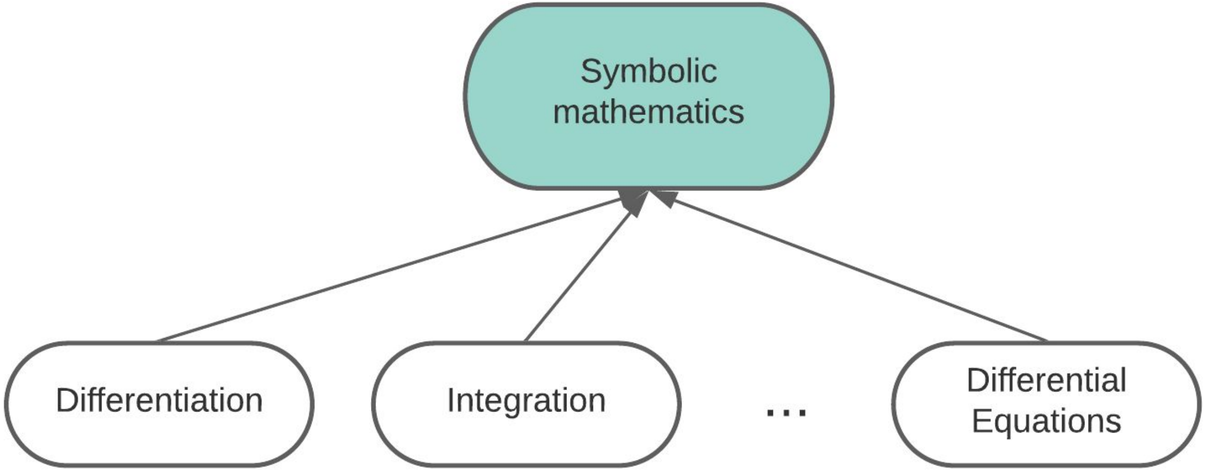
Research project

“Empirical Study of Transformers for Symbolic Mathematics”

Гельван Кирилл Павлович, гр. БПМИ171

Научный руководитель: Чиркова Надежда Александровна

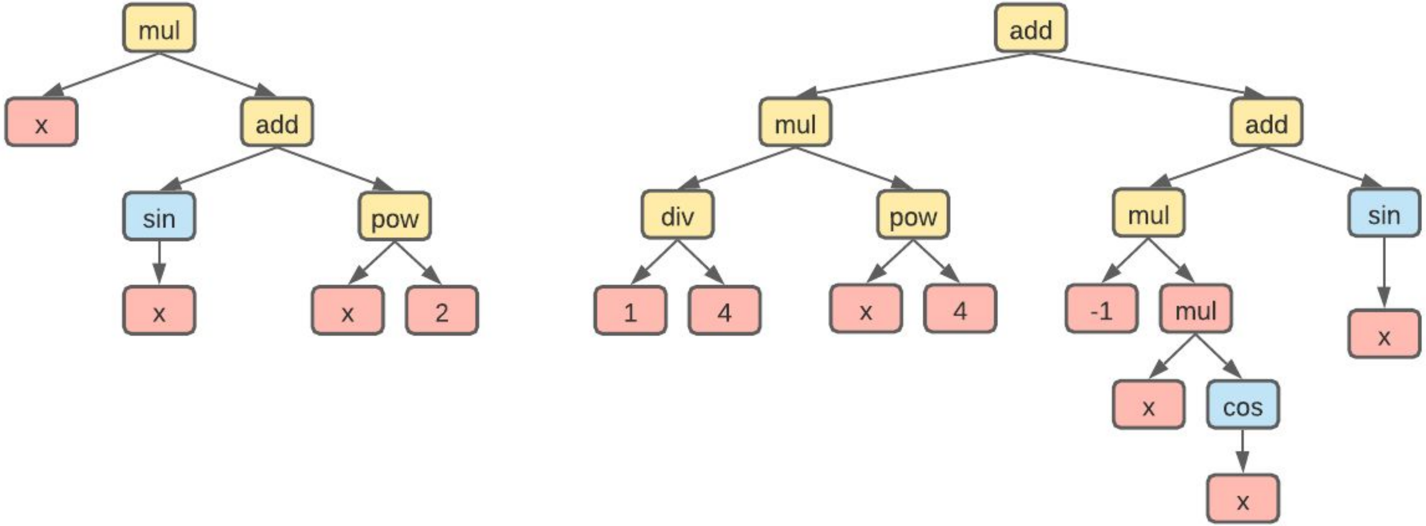
Problem statement



Passing structure to Transformers

Equation: $\int x(\sin x + x^2)\,\mathrm{dx}$ $\qquad$ $\dfrac{x^4}{4} + \sin x - x\cos x$

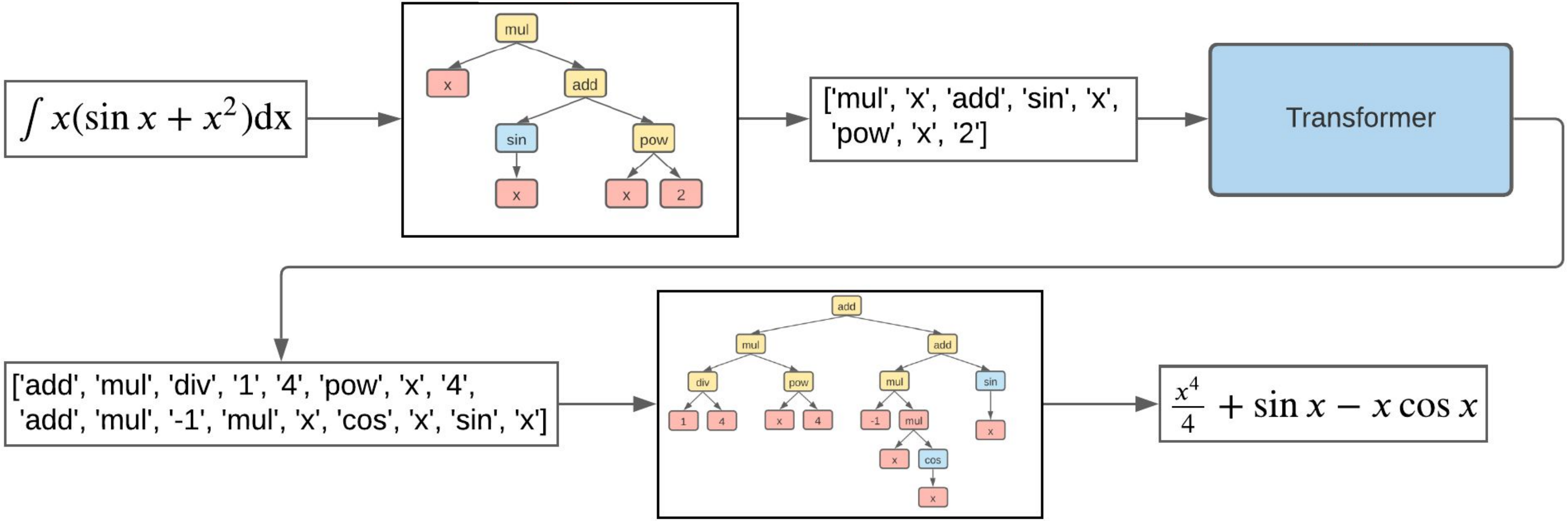
Tree:



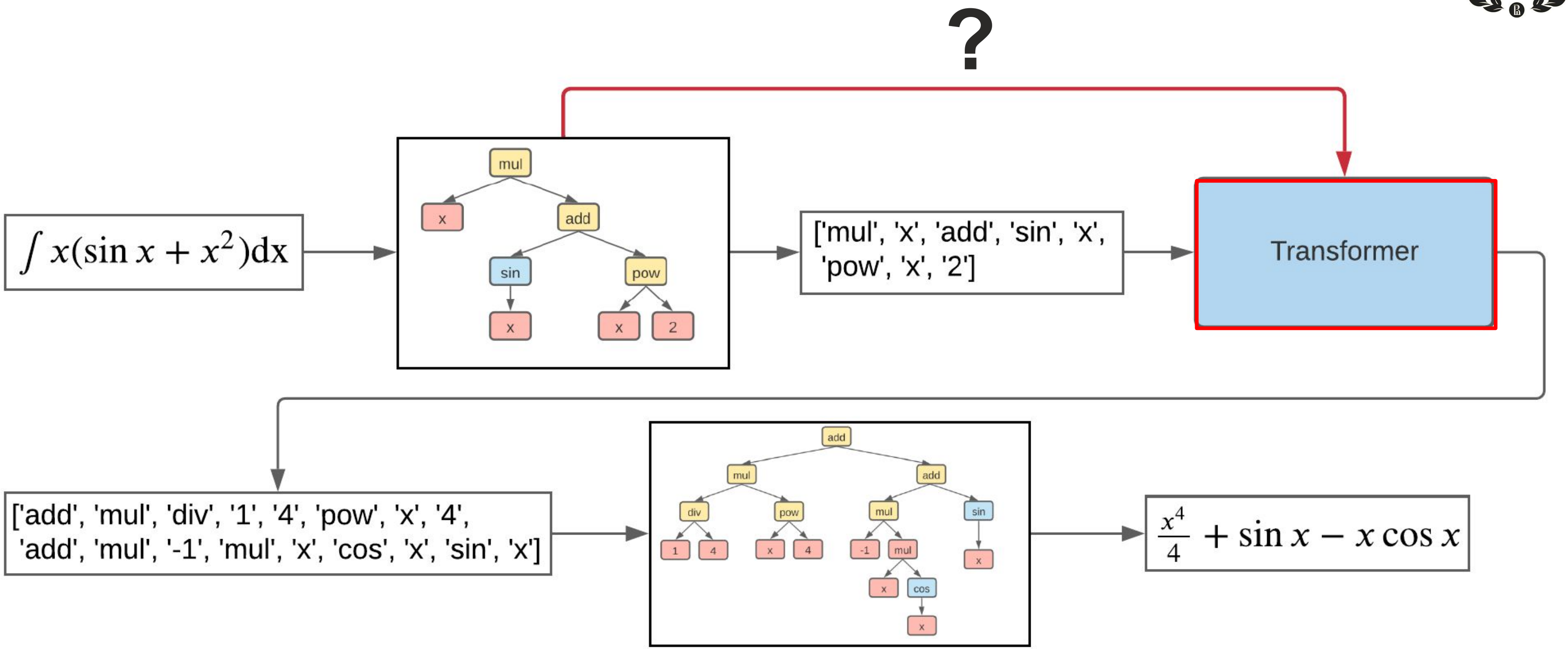
Sequence: ['mul', 'x', 'add', 'sin', 'x', 'pow', 'x', '2'] $\qquad$ ['add', 'mul', 'div', '1', '4', 'pow', 'x', '4', 'add', 'mul', '-1', 'mul', 'x', 'cos', 'x', 'sin', 'x']

Guillaume Lample and François Charton 2020

Problem statement and goal setting



Problem statement and goal setting



Problem statement and goal setting

Goal:

Investigate whether utilizing tree-based data structure in Transformer improves its performance on symbolic math tasks

Problem statement and goal setting

Goal:
Investigate whether utilizing tree-based data structure in Transformer improves its performance on symbolic math tasks

Tasks:

- Understand the specifics of the datasets
- Adapt different approaches to the task specifics
- Compare approaches empirically
- Analyze predictions of different approaches

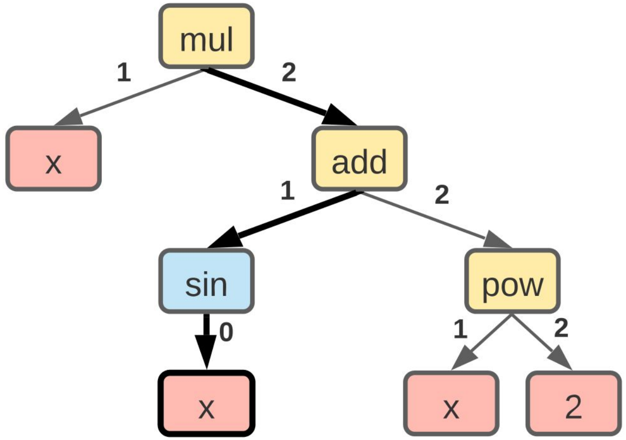
2 sym. math tasks: integration and ODEs. Solved with 4 different approaches

Positional embeddings

Input:	$x = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^{d_k}$
Positional embedding:	$p_i = \mathrm{nn.Embedding(i)}$
Embedded input:	$\hat{x}_i = x_i + p_i$

Tree positional encodings

Input:	$x = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^{d_k}$
Positional embedding:	$p_i = \text{nn.Embedding}(\text{"012"})$
Embedded input:	$\hat{x}_i = x_i + p_i$



(b) Tree positional encodings.
The node "x" is encoded as "012" (stack-like).

Shiv and Quirk 2019

Self-attention

Input:	$x = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^{d_k}$
Embedding size:	d_k
Learnable matrices:	W^Q, W^K, W^V

$$a_{ij} = \frac{x_i W^Q (x_j W^K)^T}{\sqrt{d_k}}$$

$$z_i = \sum_{j=1}^{n} \mathrm{softmax}(a_{ij})(x_j W^V)$$

Vaswani et al. 2017

Relative position representations

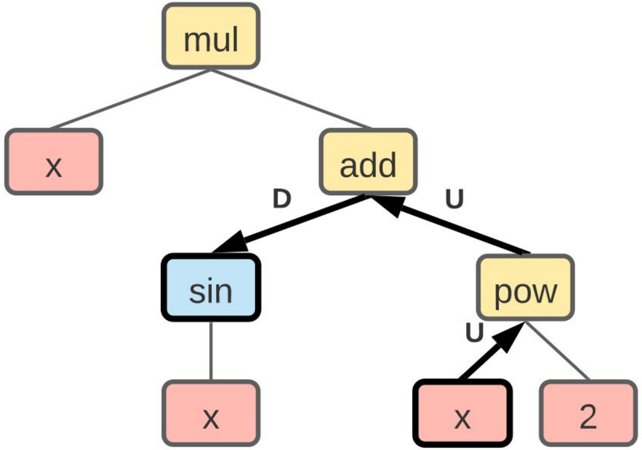
$$a_{ij} = \frac{x_i W^Q \left(x_j W^K + e_{ij}^K\right)^T}{\sqrt{d_k}}$$

$$z_i = \sum_{j=1}^{n} \mathrm{softmax}(a_{ij}) \left(x_j W^V + e_{ij}^V\right)$$

$$e_{ij}^V, e_{ij}^K \in \mathbb{R}^{d_k}$$

Shaw, Uszkoreit, and Vaswani 2018

Tree relative attention



(a) Tree relative attention.
The relation between "x" and "sin" is encoded as "UUD".

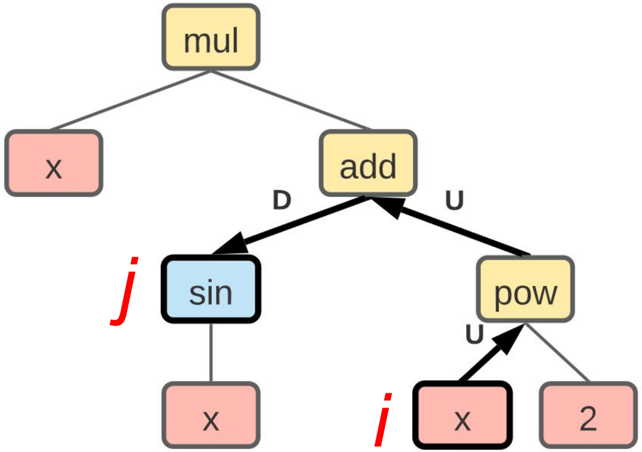
Kim et al. 2020

Tree relative attention

$$a_{ij} = \frac{x_i W^Q (x_j W^K)^T}{\sqrt{d_k}}$$

$$\tilde{a}_{ij} = \frac{\exp(a_{ij} \cdot r_{ij})}{\sum_j \exp(a_{ij} \cdot r_{ij})}$$

$$z_i = \sum_{j=1}^{n} \tilde{a}_{ij}(x_j W^V)$$



(a) Tree relative attention.
The relation between "x" and "sin" is encoded as "UUD".

Kim et al. 2020

Deduplication effect

Equation-wise accuracy (%)

	Initial test set	Deduplicated test set	Difference
Integration	*87.70*	*80.17*	**-7.53**
First order diff. eq.	*90.59*	*89.18*	**-1.41**

Preliminary experiments

- *Positional embeddings:*

Baseline

Preliminary experiments

- *Positional embeddings:* Baseline
- *Tree positional encodings:* Hyperparameter search + utilized structure in self-attention in decoder

Preliminary experiments

- *Positional embeddings:* Baseline
- *Tree positional encodings:* Hyperparameter search + utilized structure in self-attention in decoder
- *Relative position representations:* Hyperparameter search + utilized structure in encoder-decoder attention

Preliminary experiments

Positional embeddings:	Baseline
Tree positional encodings:	Hyperparameter search + utilized structure in self-attention in decoder
Relative position representations:	Hyperparameter search + utilized structure in encoder-decoder attention
Tree relative attention:	Utilized structure in self-attention in decoder

Models comparison

Equation-wise accuracy (%). Mean and standard deviation over 3 runs.

	Integration	First order diff. eq.
Positional Embeddings	80.17 ± 0.40	89.18 ± 0.35
Tree Positional Encodings	79.88 ± 0.45	$77.79 \pm 0.40^*$
Relative Position Representations	80.10 ± 0.15	89.53 ± 0.27
Tree Relative Position Representations	79.76 ± 0.12	$82.75 \pm 0.25^*$

*Runs were clipped by 5 days of training time.

Prediction analysis

Simple example

Equation:	$\int x(\sin x + x^2)dx$
1. Positional embedding:	$\frac{x^4}{4} - x\cos x + \sin x$
2. Tree positional encoding:	$\frac{x^4}{4} - x\cos x + \sin x$
3. Relative position representations:	$\frac{x^4}{4} - x\cos x + \sin x$
4. Tree relative attention:	$\frac{x^4}{4} - x\cos x + \sin x$

beam width 1

Gold:	$\frac{x^4}{4} - x\cos x + \sin x$

Prediction analysis

	Simple example	*One sample finding*
Equation:	$\int x(\sin x + x^2)dx$	$\int -48x^5 + \frac{x \cos x}{20} + \frac{\sin x}{20} + 1\ dx$
1. Positional embedding:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x - \frac{\cos(x)}{20}$
2. Tree positional encoding:	$\frac{x^4}{4} - x\cos x + \sin x$	$x^3(x+1)^{\frac{2}{3}} + \frac{x^3}{3}$
3. Relative position representations:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x$
4. Tree relative attention:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x - \frac{\cos(x)}{20}$
	beam width 1	*beam width 1*
Gold:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x$

Prediction analysis

	Simple example	*One sample finding*	*Different number of hypotheses required*
Equation:	$\int x(\sin x + x^2)dx$	$\int -48x^5 + \frac{x \cos x}{20} + \frac{\sin x}{20} + 1\ dx$	$\int - \sin^2 x + \cos^2 x\ dx$
1. Positional embedding:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x \sin(x)}{20} + x - \frac{\cos(x)}{20}$	$\sin x \cos x$ (hyp=86)
2. Tree positional encoding:	$\frac{x^4}{4} - x\cos x + \sin x$	$x^3(x+1)^{\frac{2}{3}} + \frac{x^3}{3}$	$\sin x \cos x$ (hyp=89)
3. Relative position representations:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x \sin(x)}{20} + x$	$\sin x \cos x$ **(hyp=1)**
4. Tree relative attention:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x \sin(x)}{20} + x - \frac{\cos(x)}{20}$	$\sin x \cos x$ (hyp=9)
	beam width 1	*beam width 1*	*beam width 100*
Gold:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x \sin(x)}{20} + x$	$\sin x \cos x$

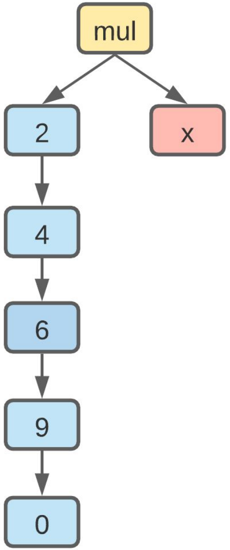
Prediction analysis

	Simple example	One sample finding	Different number of hypotheses required	Most cases with long numbers differ significantly
Equation:	$\int x(\sin x + x^2)dx$	$\int -48x^5 + \frac{x\cos x}{20} + \frac{\sin x}{20} + 1\,dx$	$\int -\sin^2 x + \cos^2 x\,dx$	$\int 24690x + \frac{6789}{x}dx$
1. Positional embedding:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x - \frac{\cos(x)}{20}$	$\sin x \cos x$ (hyp=86)	$12435x^2 + 6789\log x$
2. Tree positional encoding:	$\frac{x^4}{4} - x\cos x + \sin x$	$x^3(x+1)^{\frac{2}{3}} + \frac{x^3}{3}$	$\sin x \cos x$ (hyp=89)	$\frac{243x^2}{2} + 12\log x$
3. Relative position representations:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x$	$\sin x \cos x$ **(hyp=1)**	$12345x^2 + 6789\log x$
4. Tree relative attention:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x - \frac{\cos(x)}{20}$	$\sin x \cos x$ (hyp=9)	$12345x^2 + 6789\log x$
	beam width 1	*beam width 1*	*beam width 100*	*beam width 10-100*
Gold:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x$	$\sin x \cos x$	$12345x^2 + 6789\log x$

Prediction analysis

Sample: $24690x$

Tree:



Sequence: ['2', '4', '6', '9', '0', 'x']

Most cases with long numbers differ significantly

$$\int 24690x + \frac{6789}{x} dx$$

$$12435x^2 + 6789 \log x$$

$$\frac{243x^2}{2} + 12 \log x$$

$$12345x^2 + 6789 \log x$$

$$12345x^2 + 6789 \log x$$

beam width 10-100

$$12345x^2 + 6789 \log x$$

Prediction analysis

	Simple example	*One sample finding*	*Different number of hypotheses required*	*Most cases with long numbers differ significantly*
Equation:	$\int x(\sin x + x^2)dx$	$\int -48x^5 + \frac{x\cos x}{20} + \frac{\sin x}{20} + 1\,dx$	$\int -\sin^2 x + \cos^2 x\,dx$	$\int 24690x + \frac{6789}{x}dx$
1. Positional embedding:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x - \frac{\cos(x)}{20}$	$\sin x \cos x$ (hyp=86)	$12435x^2 + 6789\log x$
2. Tree positional encoding:	$\frac{x^4}{4} - x\cos x + \sin x$	$x^3(x+1)^{\frac{2}{3}} + \frac{x^3}{3}$	$\sin x \cos x$ (hyp=89)	$\frac{243x^2}{2} + 12\log x$
3. Relative position representations:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x$	$\sin x \cos x$ **(hyp=1)**	$12345x^2 + 6789\log x$
4. Tree relative attention:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x - \frac{\cos(x)}{20}$	$\sin x \cos x$ (hyp=9)	$12345x^2 + 6789\log x$
	beam width 1	*beam width 1*	*beam width 100*	*beam width 10-100*
Gold:	$\frac{x^4}{4} - x\cos x + \sin x$	$-8x^6 + \frac{x\sin(x)}{20} + x$	$\sin x \cos x$	$12345x^2 + 6789\log x$

Summary

- Investigated the deduplication effect and prepared the deduplicated dataset
- Adapted the modifications according to the task specifics
- Conducted extensive experiments on two symbolic mathematics tasks and empirically observed the absence of a statistically significant difference between the base Transformer architecture and advanced versions
- Performed a qualitative analysis of the trained models and found out that tree-structure-based approaches process long numbers better