Embedding Trajectory for Out-of-Distribution Detection in Mathematical Reasoning

Yiming Wang[α] Pei Zhang[β,γ] Baosong Yang[β,✉] Derek F. Wong[γ]
Zhuosheng Zhang[α] Rui Wang[α,✉]

[α]Shanghai Jiao Tong University [β]Tongyi Lab [γ]NLP[2]CT Lab, University of Macau

✉: Corresponding Author

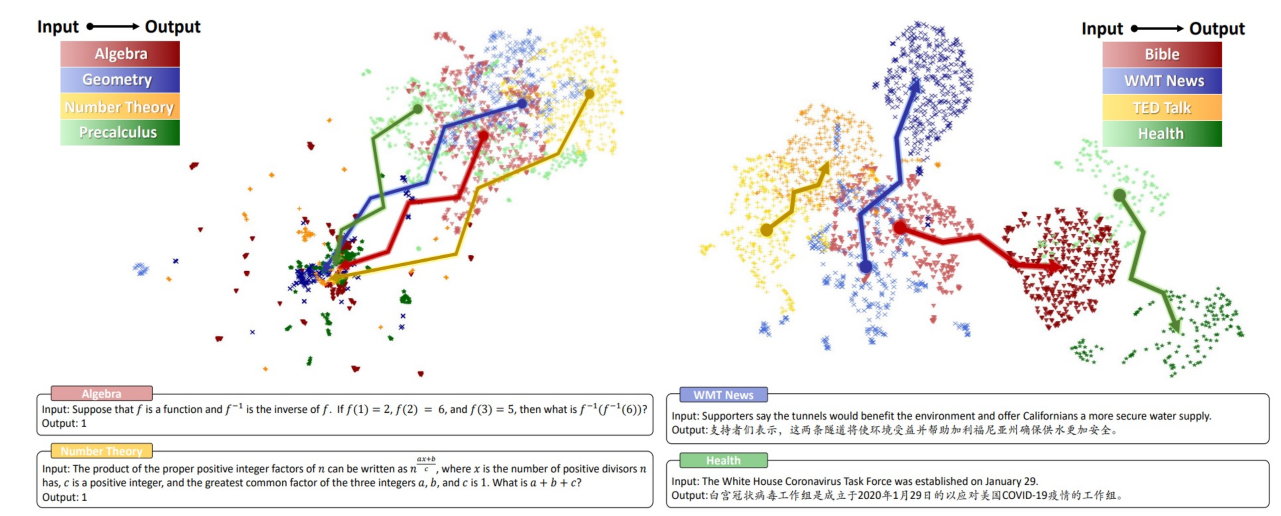
Email: [α]{yiming.wang, wangrui12}@sjtu.edu.cn
[β]yangbaosong.ybs@alibaba-inc.com

NEURAL INFORMATION PROCESSING SYSTEMS

◆ *Abbreviations: In-Distribution -> ID; Out-of-Distribution -> OOD; Mahalanobis Distance -> MaDis*

Inapplicability of Static Embedding Method in Mathematical Reasoning

Static Embedding Method (Traditional):
MaDis between the *new sample embedding* and *ID embedding distribution* **in static input or output space.**



(a) Mathematical Reasoning (b) Text Generation (Translation)

Why Static Embedding Method Inapplicable:

- ***Input space*** exhibits vague clustering features across domains.
- ***Output space*** of mathematical reasoning exhibits high-density characteristics with significant overlap between different domains -> "***Pattern Collapse***"
  - ➢ **Expression-level**: Output is symbolic/scalar (1,2,x,y...) -> compress the search space.
  - ➢ **Token-level:** Sequence tokenization used in GLMs allows for substantial token sharing among mathematically distinct expressions

(2822 -> ['2', '8', '2', '2'], 8122/8 -> ['8', '1', '2', '2', '/', '8'], ...)

TV Score: Trajectory-based OOD Detection Method

- First, we fit all ID embeddings at each layer l to a Gaussian distribution

$$\mathcal{G}_l = \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$$

- Next, for a new sample with y_l be its embedding at layer l, we map it to its MaDis

$$f(\boldsymbol{y}_l) = (\boldsymbol{y}_l - \boldsymbol{\mu}_l)^\top (\boldsymbol{\Sigma}_l)^{-1} (\boldsymbol{y}_l - \boldsymbol{\mu}_l) \quad (1 \le l \le L)$$

- Finally, we average all adjacent-layer volatilities $|f(\boldsymbol{y}_l) - f(\boldsymbol{y}_{l-1})|$ as the final trajectory volatility score

$$S = \frac{1}{L} \cdot \sum_{l=1}^{L} |f(\boldsymbol{y}_l) - f(\boldsymbol{y}_{l-1})| \qquad \textbf{(TV Score)}$$

- First, We define the k-order embedding and Gaussian distribution

$$\nabla^k \boldsymbol{y}_l = \sum_{i=0}^{k} (-1)^{k+i} \mathrm{C}_k^i \boldsymbol{y}_{l+k}, \qquad \nabla^k \mathcal{G}_l = \mathcal{N}\left(\sum_{i=0}^{k} (-1)^{k+i} \mathrm{C}_k^i \boldsymbol{\mu}_{l+k}, \sum_{i=0}^{k} \mathrm{C}_k^i \boldsymbol{\Sigma}_{l+k} \right)$$

- Next, we map $\nabla^k \boldsymbol{y}_l$ to its MaDis

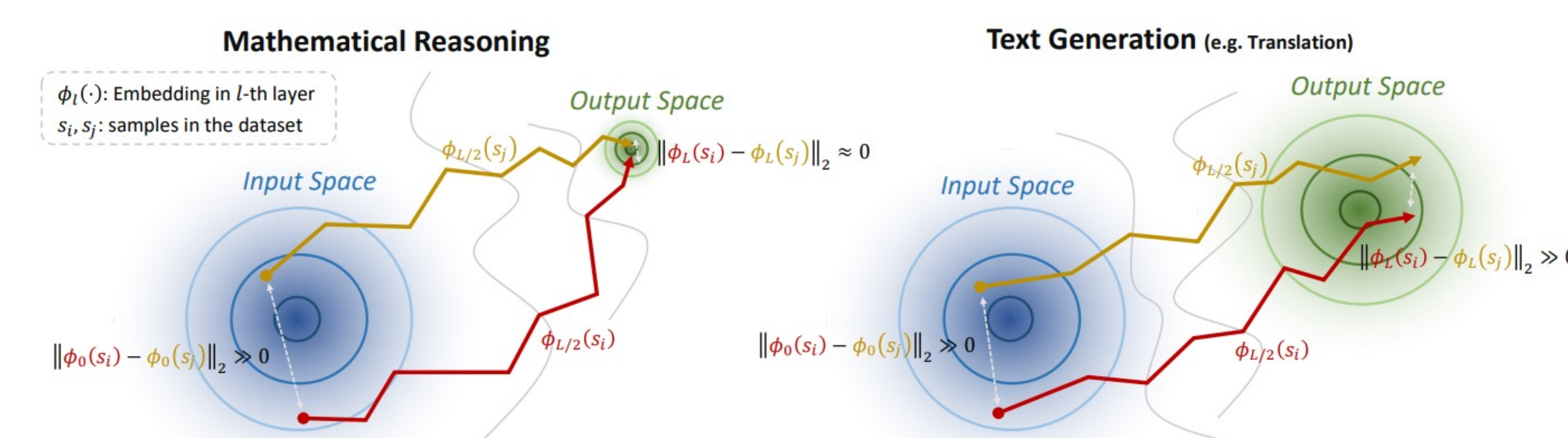
$$\nabla^k f(\boldsymbol{y}_l) = \left(\nabla^k \boldsymbol{y}_l - \sum_{i=0}^{k} (-1)^{k+i} \mathrm{C}_k^i \boldsymbol{\mu}_{l+k} \right)^\top \left(\sum_{i=0}^{k} \mathrm{C}_k^i \boldsymbol{\Sigma}_{l+k} \right)^{-1} \left(\nabla^k \boldsymbol{y}_l - \sum_{i=0}^{k} (-1)^{k+i} \mathrm{C}_k^i \boldsymbol{\mu}_{l+k} \right)$$

- Finally, we define the trajectory volatility score after *differential smoothing*

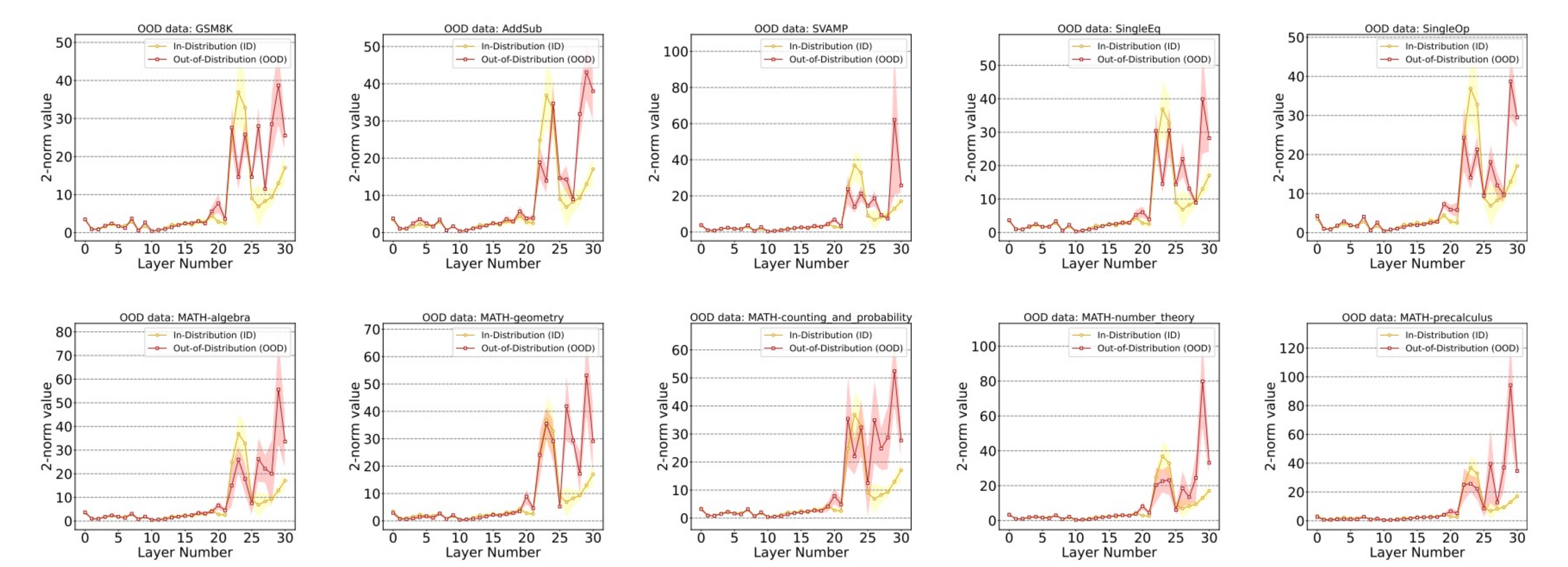
$$S = \frac{1}{L} \cdot \sum_{l=1}^{L} |\nabla^k f(\boldsymbol{y}_l) - \nabla^k f(\boldsymbol{y}_{l-1})| \qquad \textbf{(TV Score w/ DiSmo)}$$

Why Dynamic Embedding Trajectory Applicable?

➢ **What is the Embedding Trajectory?** $\boldsymbol{y}_l$ is the sentence embedding at layer l, the embedding trajectory is formed as a progressive chain of these embeddings: $\boldsymbol{y_0} \to \boldsymbol{y_1} \to \cdots \to \boldsymbol{y_l} \to \cdots \to \boldsymbol{y_L}$



- "*pattern collapse*" causes the convergence of the trajectory endpoints of different samples -> leading to **significant trajectory differences across samples.**



Change curves of embedding differences between neighboring hidden layers under ID and OOD data (Llama-2-7B)

- For ID data, GLMs **largely complete reasoning in the mid-to-late stages** (->"***Early Stablization***"), and simple adjustments are enough after that.
- For OOD data, GLMs can **still not complete accurate reasoning at a later stage**, can only randomly switch to a specific output pattern.

Experiments

Model	Llama2-7B [50]				GPT2-XL [5]			
Metric / *Method*	Far-shift OOD		Near-shift OOD		Far-shift OOD		Near-shift OOD	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Max Softmax Prob. [12]	78.66±1.38	81.44±3.56	60.14±1.54	88.91±2.41	70.54±1.42	78.29±2.02	67.12±1.20	76.27±2.66
Monte-Carlo Dropout [8]	68.63±2.21	87.04±4.88	52.33±2.21	91.92±1.89	66.18±1.87	84.69±1.65	63.54±1.72	78.08±2.50
Perplexity [3]	<u>85.64±1.46</u>	<u>53.06±4.36</u>	59.35±1.89	86.09±1.89	80.82±1.04	64.53±2.10	73.74±1.12	72.39±1.27
Input Embedding [43]	75.89±1.03	67.87±3.69	<u>60.33±1.37</u>	<u>84.65±2.53</u>	<u>86.26±0.84</u>	<u>49.33±2.10</u>	<u>83.22±0.88</u>	<u>52.90±3.16</u>
Output Embedding [43]	74.86±1.39	75.21±2.16	44.50±1.06	86.46±1.59	77.95±1.16	65.64±3.42	79.28±1.24	64.70±2.72
TV Score (Ours)	**98.76±0.11**	**5.21±0.98**	**92.64±0.39**	**28.39±1.38**	93.47±0.08	24.10±0.95	**94.86±0.23**	13.82±0.36
w/ DiSmo (Ours)	93.25±0.76	41.82±4.69	56.99±1.41	88.01±1.71	**96.54±0.11**	**9.89±0.61**	94.19±0.25	**13.66±0.69**
Δ (**bold** - <u>underline</u>)	+13.12	-47.85	+32.31	-56.26	+10.28	-39.44	+11.64	-39.24

	Far-shift OOD Setting			Near-shift OOD Setting	
Dataset	**Accuracy↑**	**Robustness↓**	*Dataset*	**Accuracy↑**	**Robustness↓**
	I-Emb. / O-Emb. / TV (ours)			I-Emb. / O-Emb. / TV (ours)	
Algebra	76.43 / 45.42 / **93.88**	5.27 / 6.94 / **0.97**	GSM8K	81.49 / 75.32 / **93.39**	10.08 / 3.36 / **2.05**
Geometry	74.32 / 54.79 / **94.47**	2.44 / 2.43 / **1.65**	SVAMP	68.66 / 63.33 / **94.88**	5.26 / 3.54 / **2.13**
Cnt.&Prob	50.31 / 27.55 / **93.74**	9.99 / **2.34** / 2.36	AddSub	**79.16** / 78.09 / 74.11	3.21 / 6.98 / **2.77**
Num.Theory	85.80 / 54.38 / **92.08**	3.31 / 11.45 / **2.34**	SingleEq	59.83 / 72.56 / **93.15**	11.57 / **3.14** / 3.17
Precalculus	80.33 / 88.50 / **99.28**	6.13 / 1.38 / **0.67**	SingleOp	69.38 / 62.20 / **95.75**	4.00 / **2.37** / 2.45
Average	73.44 / 54.13 / **94.69**	5.43 / 4.91 / **1.60**	*Average*	71.70 / 70.30 / **90.26**	6.82 / 3.88 / **2.51**

Paper **Github**