arxiv.org/abs/2505.15778

Soft Thinking: Unlocking the Reasoning Potential of LLMs in Continuous Concept Space

开源代码：github.com/eric-ai-lab/Soft-Thinking

Soft Thinking：一种引入连续空间概念的解码策略

Zhen Zhang[1]* **Xuehai He**[2]* **Weixiang Yan**[1] **Ao Shen**[4] **Chenyang Zhao**[3,5]
Shuohang Wang[6] **Yelong Shen**[6] **Xin Eric Wang**[1,2]
[1]University of California, Santa Barbara, [2]University of California, Santa Cruz
[3]University of California, Los Angeles, [4]Purdue University, [5]LMSYS Org, [6]Microsoft
zhen_zhang@ucsb.edu, ericxwang@ucsb.edu

简介

本文提出Soft Thinking：一种无需训练llm即可在连续概念空间(Continuous Concept Space)中进行解码的策略。

按照论文的说法，soft thinking是对CoT reasoning的改进，我个人观点，本质上soft thinking是一种解码(decoding)策略，既然是解码策略，那么一般就具有通用性，所以不局限在CoT reasoning，或许这只是一个应用场景吧，也方便做实验。

ok下面说下什么是soft thinking，首先我们回顾下自回归式llm如何做解码的，在某个step，softmax输出整个词表的概率分布，然后用greedy/各种Sampling得到一个token，再把它作为输入去预测下一个token。本文将softmax得到的概率分布称为概念token(concept token)，不greedy/sampling了，直接根据概率分布将整个词表向量线性加权，作者说这个线性加权得到的向量空间就是连续概念空间(continuous concept space)，对应上题目了，有了**连续**的概念了，然后把线性加权后的向量作为输入去预测下一个“token”。

再说下第二个创新点cold stop，可以联想下训练模型常用的技巧early stop，既然softmax得到概率分布，那么就可以计算熵(entropy)，如果熵小，说明分布很不均匀，表示模型信心很足，如果连续多个step的熵都很小，就强行停止thinking，让llm去预测答案。

示意图

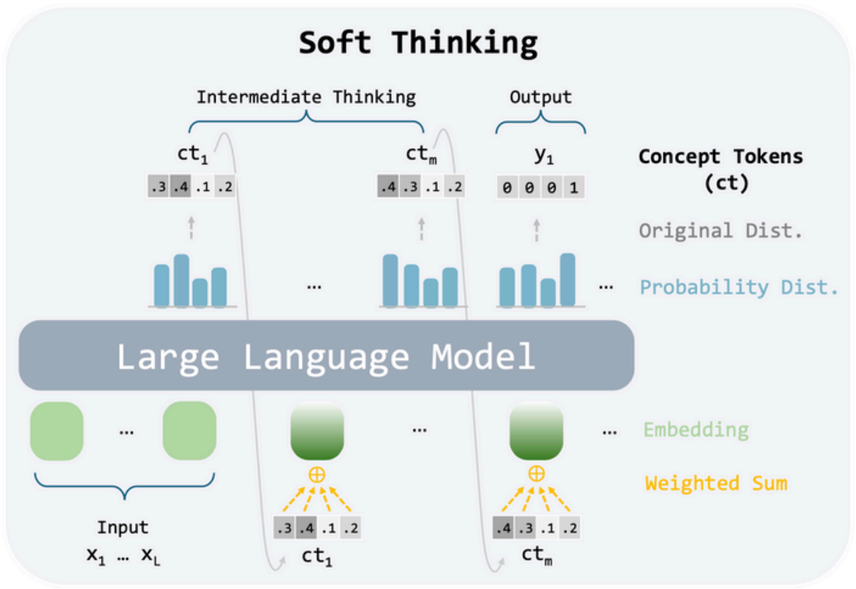

Figure 2: *Soft Thinking* replaces discrete tokens with soft, abstract *concept tokens*, enabling reasoning in continuous concept space.

部分实验结果

	Accuracy ↑					Generation Length ↓				
	MATH 500	AIME 2024	GSM8K	GPQA Diamond	Avg.	MATH 500	AIME 2024	GSM8K	GPQA Diamond	Avg.
	QwQ-32B [13]									
CoT Thinking	97.66	76.88	96.67	64.17	83.84	4156	12080	1556	8095	6472
CoT Thinking (Greedy)	97.00	80.00	96.57	65.15	84.68 (↑ 0.84)	3827	11086	1536	7417	5967 (↓ 7.8%)
Soft Thinking	**98.00**	**83.33**	**96.81**	**67.17**	**86.32** (↑ 2.48)	**3644**	**10627**	**1391**	**7213**	**5719** (↓ 11.6%)
	DeepSeek-R1-Distill-Qwen-32B [38]									
CoT Thinking	94.50	72.08	95.61	63.10	81.32	3543	9347	875	6218	4995
CoT Thinking (Greedy)	93.00	63.33	95.30	59.09	77.68 (↓ 3.64)	3651	8050	1048	8395	5286 (↑ 5.8%)
Soft Thinking	**95.00**	**76.66**	**95.83**	**64.64**	**83.03** (↑ 1.71)	**3373**	**6620**	**785**	**4722**	**3875** (↓ 22.4%)
	DeepSeek-R1-Distill-Llama-70B [38]									
CoT Thinking	94.70	70.40	94.82	65.34	81.31	3141	8684	620	5500	4486
CoT Thinking (Greedy)	94.61	**73.33**	93.60	66.16	81.92 (↑ 0.61)	**2877**	9457	606	**4443**	4345 (↓ 3.1%)
Soft Thinking	**94.80**	**73.33**	**94.90**	**66.66**	**82.42** (↑ 1.11)	3021	**6644**	**597**	4470	3683 (↓ 17.9%)

思考

自回归式LLM或者说NLP，是建立在离散语言结构基础上的，相比于连续，似乎离散型就是存在不足，作者能够去探索如何引入连续性是很值得鼓励的。只不过，我个人认为Soft Thinking 所做的“用 softmax 分布加权 embedding”这一做法稍微有那么一点点浅显了，现在llm的词表都很大，十几万的embedding加起来，哪怕再结合下top-p sampling呢？而且连续解码还有一个问题，由于不再输出具体的token，而是概率分布，那么就没有llm的思考过程了，只能看到llm输出的答案。

以上仅是我个人观点，不是批评，相比起很多easy task，作者能去啃连续这个硬骨头，已经很值得鼓励了。