AN ANALYSIS OF SUPERPOSITION IN CONTINUOUS CHAIN-OF-THOUGHT

Benjamin Roderick, Edmand Yu & Yuchen Wu
Students, Department of Computer Science
McGill University
Montreal, QC, Canada
{benjamin.roderick,edmand.yu,yuchen.wu2}@mail.mcgill.ca

Michael Rizvi-Martel
Mentor, Mila
Montreal, QC, Canada
michael.rizvi-martel@mila.quebec

1 MOTIVATION

Recently, **chain-of-thought (CoT)** has become the standard for reasoning in LLMs. Initially, CoT simply involved giving the model more "thinking time" by prompting it to produce intermediate reasoning steps during inference (Wei et al., 2022). Other methods, such as supervised fine-tuning (Yu et al., 2023) and reinforcement learning optimization (Shao et al., 2024), have also been explored. However, this form of reasoning remains constrained to the space of natural language. To overcome this limitation, researchers began exploring models capable of reasoning without the uses of explicit tokens. For instance, *Let's Think Dot by Dot* (Pfau et al., 2024) use special "thinking" tokens to facilitate internal reasoning. Another avenue is **continuous chain-of-thought** (Hao et al., 2024), which generates reasoning steps directly in the embedding space rather than in token space. Additionally, it has been shown that continuous CoT can also represent multiple potential reasoning paths simultaneously, enabling more efficient and expressive inference than discrete CoT through "superposition" of concepts (Zhu et al., 2025). A related, training-free method called **Soft Thinking** (Zhang et al., 2025) extends this idea by introducing *concept tokens*, which are continuous representations that encode multiple related meanings from discrete tokens. The authors report that **Soft Thinking** improves standard CoT while keeping outputs interpretable and readable. However, we noted that the effectiveness of **Soft Thinking** hinges on the assumption that convex combinations in embedding space decode to a mixture of distributions in probability space. As noted by the authors (Zhang et al., 2025), using *concept tokens* at inference, places the model in an out-of-distribution regime, since the model was never exposed to such continuous representations during training. This limitation motivates our primary research question: whether superpositions in embedding space correspond to superpositions in probability space. To the best of our knowledge, no prior work has systematically evaluated this assumption, and in this project, we aim to do so.

2 RESEARCH QUESTION + METHOD

2.1 RQ1

The first objective of this project is to determine whether *concept tokens* can be decoded by deconstructing the superposition or convex combination of token embeddings. In practice, this involves developing a method to reverse the process of constructing a *concept token* at any point during a series of CoT reasoning steps. This approach differs from previous work such as **Coconut** (Hao et al., 2024) because only the final hidden state is returned to the LLM as input.

2.2 RQ2

If the primary objective is attained, the secondary objective is to analyze if these decoded *concept tokens* are interpretable. Additionally, if they are indeed interpretable, we wish to confirm to what

degree these decoded hidden states are similar to those found in traditional natural language CoT models.

3 HYPOTHESIS/EXPECTED RESULTS

Considering the fact that previous work on *Soft Thinking* was able to decode the final hidden state of the CoT process, we will likely be able to decode intermediate states. However, for the secondary objective, a likely scenario is that discrete and continuous CoT hidden states have only slight similarities.

4 EXPERIMENTAL DESIGN

The project is in the experimental and analysis phase. This experiment will investigate how word embeddings behave when combined under different grouping strategies, such as semantic similarity, word type, and random baseline. Our methodology is to start with a small number of embedding vectors (e.g. $k = 2$); we will select a bin based on the chosen grouping type (for example, adjectives in the case of word type), randomly sample k embeddings from that bin, and combine them using a uniform convex combination (e.g. $0.5e_1 + 0.5e_2$). The resulting vector will then be passed through the model's unembedding matrix, and the resulting logits will be analyzed to determine whether the top-ranked tokens correspond to the expected "superposition" of the original words. If this occurs, we will further investigate whether there is a critical point at which this "superposition" emerges or disappears as k increases, and we will also examine the effects of non-uniform convex combinations.

5 PROJECT TIMELINE AND ROLES

The project will proceed in three main phases. Phase 1 will conduct preliminary convex-combination tests using pretrained embedding and unembedding matrices to verify the superposition hypothesis. Phase 2 will scale up the analysis to larger token sets and assess statistical consistency across different semantic groups. The students will handle the technical aspects of the project, including writing code and training models as needed. The mentor has provided and will continue to provide guidance on experimental design and methodology, meeting with the students approximately once every one to two weeks to ensure the project is progressing smoothly.

REFERENCES

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *arXiv preprint arXiv:2505.12514*, 2025.