# Multimodal chain-of-thought

**Anonymous Authors**[1]

## Abstract

Language models (LMs), particularly large-scale models (LLMs), have demonstrated remarkable capabilities across various tasks due to a reasoning process, called chain-of-thought (CoT), emerged in them. However, CoT, while improving predictive quality, comes at the cost of increased inference time. In this project, we propose a novel approach to mitigate this trade-off by introducing compressed CoT representations. Inspired by successful techniques in light-weight multimodal LLM extensions, we employ a two-step process: first, training compressed data representations using effective compressing methods i.e quantized autoencoders, and then fine-tuning pretrained LLMs with an extended vocabulary incorporating ”tokens” from different modalities. We present our methodology and discuss the potential impact on improving both efficiency and effectiveness of LLMs. Experimental results are produced on two different multi-modal datasets: (1) ScienceQA multi-modal dataset with text and images; (2) Sk3D dataset with images, depth maps and meshes. The second dataset is additionaly annotated with detailed instructions suitable for training LLMs with CoT. Please find out project on Git: `https://github.com/BobrG/MMDCOT`

## 1. Introduction

Large-scale language models (LLMs) have revolutionized natural language processing (NLP) tasks with their ability to generate coherent and contextually relevant text. In the realm of Language Modeling, strategies such as Chain-of-Thoughts have demonstrated the ability to enhance the predictive capabilities of LLMs across various tasks. However, the use of such methods comes at the cost of increased inference complexity.

The recent works on Multimodal LLMs and Chain-of-Though inspired us to investigate the capabilities of injesting qualified reasoning in LLMs via training on diverge data modalities. So, Multimodal CoT (Zhang et al., 2024) work expands on the Chain-of-Thought approach by integrating multimodal data, enabling the model to process and reason across textual, visual, and auditory inputs what allows for a more nuanced understanding and generation of responses that reflect a deeper comprehension of mixed-media content. AnyGPT framework (Zhan et al., 2024) proposes a novel approach to building generative models that are not limited to text but can seamlessly incorporate and generate across different modalities, including images and speech. Authors demonstrate that such an approach not only broadens the applicative potential of LLMs, but also improves the model's robustness and reasoning by leveraging cross-modal synergy during training.

Several scholarly works have highlighted the observation that language models possess redundant capacity which remains unused. In response, (Ashkboos et al., 2024) advocate for a methodological approach to alleviate this redundancy by employing orthogonal transformations, subsequently followed by the removal of dormant rows and columns from the weight matrices. A series of empirical investigations demonstrated that the embedding dimensionality of the resultant model is reduced by approximately 10-30%. Furthermore, the evaluation metrics, particularly perplexity, exhibit minimal deterioration, notably in the case of large-scale language models such as LAMA-70B and OPT-66B.

Drawing inspiration from these concepts in our project, we propose the idea to develop compressed representations of chains of thought facilitating the substitution of the LLM reasoning process with a reduced set of distinctive tokens. We propose a two-stage framework by training compressed data representations using techniques such as quantized autoencoders to reduce inferential complexities, then fine-tune pre-trained LLMs with an expanded vocabulary to incorporate multimodal ”tokens”. We introduce Sk3D dataset with additional annotations for 3D reconstruction from multiple modalities. Thus, we prepare a dataset of mixed-modality data and fine-tune LLM Incorporate point clouds as an additional modality, leveraging their spatial data characteristics

---

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

to enrich the model's understanding and predictive capabilities.

Expected advantages of our method: (1) improved efficiency due to compressed chain-of-thought reasoning; (2) broadened modalities landscape by introducing 3D data input along with 2D visual and text data on novel dataset.

## 2. Related Work

Recent advancements in language model (LM) prompting techniques have unveiled the potential for improved performance by encouraging models to engage in a form of deliberative thinking. While conventional zero-shot prompts aim for immediate responses, recent articles suggest that prompting LMs to ”think aloud” about the problem space can lead to more accurate outcomes. (Wei et al., 2022) (Kojima et al., 2022) This concept, often referred to as ”chain-of-thoughts” (CoT) entails guiding the LM through a step-by-step reasoning process similar to human deliberation. The addition of a simple ”Let's think step-by-step” prompt enabled the PaLM language model to outperform humans on 10 out of 23 tasks in the Big-Bench benchmark. (Suzgun et al., 2022)

An alternative method, termed Quiet-STaR (Zelikman et al., 2024) , enables a model to engage in silent contemplation at each token, with a trained distribution optimized for utility. The authors proposed an idea to generalize CoT prompting, approximating it to the semblance of human thought process (not expressing everything one thinks). To achieve this, the authors suggested allowing LLM to sometimes ”think” — the start and end of such thought are defined by trainable reinforcement learning (RL) special tokens, while the thought itself often consists of gibberish, reducing the perplexity of the subsequent text. Unlike CoT, there is no requirement to write coherently here, so what is generated inside the thought is not always interpretable. Nevertheless, this significantly boosts metrics and practically does not slow down inference, as generation occurs in parallel.

Data tokenizers play a key role in processing and understanding different data modalities in LLMs. For common modalities such as text, images, and speech, well-established tokenizers have been developed, enabling the efficient training of models on diverse datasets i.e AnyGPT. (Zhan et al., 2024) However, tokenizing more complex data types that are crucial for deep understanding of a 3D environment i.e 3D point clouds or meshes, presents unique challenges due to their intricate structures and the spatial relationships. Recent innovations in this area include PointBERT (Yu et al., 2022) and MeshGPT (Siddiqui et al., 2023), which have presented innovative approaches for tokenization of 3D data forms. PointBERT proposes a tokenizer for point clouds – a 3D data representation in a form of a discrete set of data points in space. This tokenizer is based on a discrete Variational AutoEncoder which generates discrete point tokens containing meaningful local information and is pre-trained on ShapeNet dataset. MeshGPT is aimed to process meshes – a different 3D representation, which consists of vertices and faces. MeshGPT learns a vocabulary of latent quantized embeddings, using graph convolutions. This approach not only preserves the geometric and topological properties of the meshes, but also facilitates the generation and manipulation of 3D shapes using language model-like techniques. In this project, apart from working with well-known modalities, we aim to dive deeper into the 3D domain and utilize tokenizers proposed for 3D data modalities.

## 3. Methods

### 3.1. Problem Formulation

The primary objective of this research is to devise a compressed representation methodology for the multi-modal reasoning process within LLMs. This approach aims to facilitate the substitution of the conventional multi-modal chain-of-thought reasoning generation mechanism with a reduced set of specialized tokens. Compressing the intricate reasoning process into a more concise form, the proposed methodology seeks to enhance the efficiency and scalability of LLM-based reasoning tasks while maintaining or even augmenting their predictive capabilities.

The overall procedure is composed of three main components:

- **Train Compressed Data Representation**. Initially, a compressed data representation is trained. One possible way that we propose in this project is to use a quantized autoencoder to fuse and compress multi-modal tokens, leading to creation of compact multi-modal reasoning generation.

- **Multi-modal data preparation and processing**. A dataset that will force LLM model to reason in 3D should be created from the existing multi-modal Skoltech3D dataset, which lack text descriptions, but has rich visual and 3D data annotations. 3D tokenizers should be selected and incorporated in the pipeline, as well.

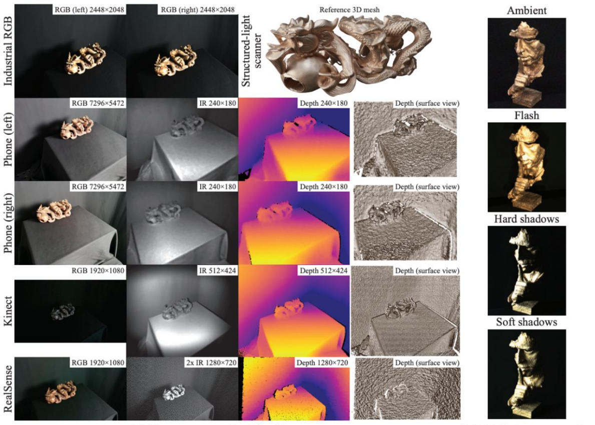
- **Fine-tune pre-trained LLM**. Subsequently, the LLM is fine-tuned using the compressed data representation. This fine-tuning process involves extending the vocabulary of the LLM to include ”tokens” representing different modalities. These modalities could include various types of data such as text, images, audio, as well as 3d data i.e. point clouds. By incorporating tokens representing different modalities, the LLM becomes capable of processing and generating outputs based on a diverse range of inputs, thereby enhancing its versatility and applicability.

# 4. Experiments and Results

## 4.1. Dataset

The aim to build a deep multi-modal reasoning in LLMs requires enough diverse data with text and other data modalities. We assess the effectiveness of our approach using the **ScienceQA** benchmark (Lu et al., 2022), renowned as the pioneering large-scale multimodal science question dataset, which meticulously annotates answers with extensive lectures and explanations. Comprising 21,000 multimodal multiple-choice questions, the dataset exhibits significant domain breadth, spanning three subjects, 26 topics, 127 categories, and 379 skills.

Another dataset, **Sk3D** (Voynov et al., 2023) is one of the recently proposed benchmarks for 3D reconstruction from multiple modalities. It provides more than 1.4M data samples of different modalities, such as images, depth maps from different types of scanners and meshes for 110 real-life objects. Yet, this dataset lacks text instructions to be used as annotations for LLM finetunning.



(a) Data captured with all sensors (b) Lighting examples

*Table 1.* Sk3D dataset data examples.

### Dataset preparation

In order to prepare Sk3d dataset for LLM, an additional automated approach was developed, which uses a GPT model to generate questions and conversations with detailed instructions in semi-automatized manner for multi-modal data samples. This script is going to be massively applied to all 110 scenes and 100 viewpoints. To answer the generated questions LLM will have to utilize visual and 3D data modalities2.

## 4.2. Tokenizers Preparation

### Quantized tokenizer

To develop compressed representations of a chain-of-thoughts, the initial step involves constructing a tokenizer. This is achieved by training an autoencoder that first reduces the number of elements within the chain and subsequently quantizes the values in the resulting shortened sequence. By performing these two steps, the autoencoder transforms an extended chain of text tokens into a concise chain of thought tokens.

We decided to start our experiments with the VQ-VAE architecture (Van Den Oord et al., 2017), adapting the model to handle multimodal input in the form of embeddings of conversation paired with an image. To obtain visual embeddings for images, we employed the CLIP-ViT-L visual encoder. The projection layer transforms the image embeddings to match the dimensionality of the text embeddings, enabling their seamless combination. Its also helps align the semantics of the image and text embeddings. And their combination then fed into the input of the autoencoder.

### 3D tokenizer

Point-MAE was chosen for encoding of 3D point cloud data that will be used in the experiments with Sk3D dataset. Before applying this tokenizer to a new data, it was debugged according to official guidelines. In finetune setup on ShapeNet dataset it was checked that pre-trained model is able to produce adequate reconstructions, when used as autoencoder. Some reconstruction examples are visualised in Table 2.

## 4.3. Experimental Setup

## 4.4. Baselines

To compare with existing solutions , we ran the Quiet-STaR method (Zelikman et al., 2024) together with a zero-shot prompt (“Let’s think step by step.”) without in-context examples. Validation on a sample of 128 GSM8K test items gives us the majority vote accuracy over 8 samples equal to 43.54 %. Examples of results on different iterations could be find in Appendix A **??**

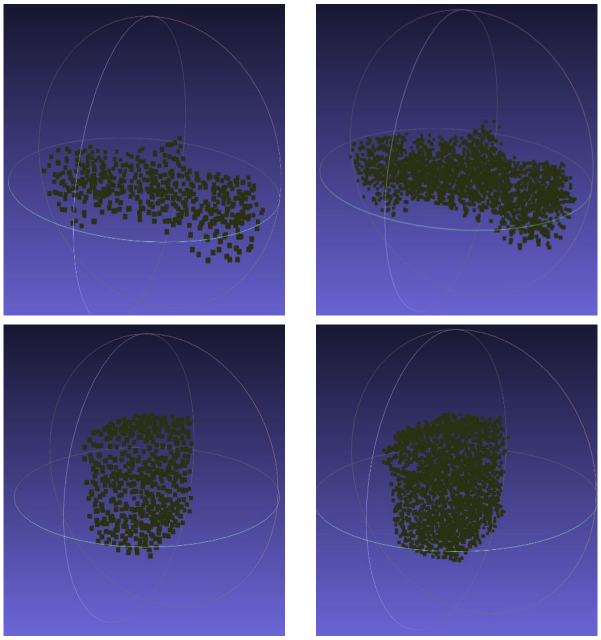
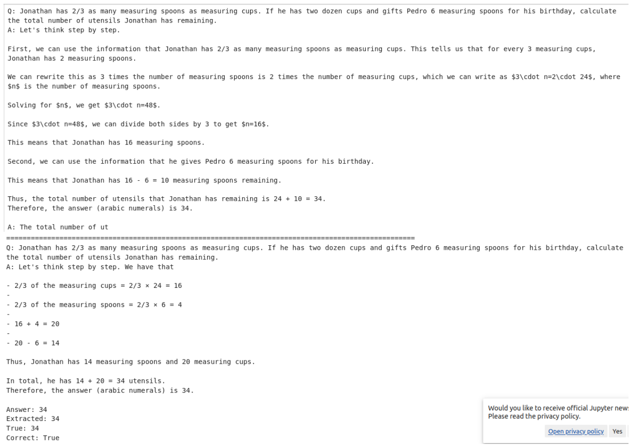


*Table 2.* Point MAE tokenizer reconstruction results. *left* images show original ShapeNet point clouds, *right* images show reconstructed results.



```
Q: Jonathan has 2/3 as many measuring spoons as measuring cups. If he has two dozen cups and gifts Pedro 6 measuring spoons for his birthday, calculate
the total number of utensils Jonathan has remaining.
A: Let's think step by step.

First, we can use the information that Jonathan has 2/3 as many measuring spoons as measuring cups. This tells us that for every 3 measuring cups,
Jonathan has 2 measuring spoons.

We can rewrite this as 3 times the number of measuring spoons is 2 times the number of measuring cups, which we can write as $3\cdot n=2\cdot 24$, where
$n$ is the number of measuring spoons.

Solving for $n$, we get $3\cdot n=48$.

Since $3\cdot n=48$, we can divide both sides by 3 to get $n=16$.

This means that Jonathan has 16 measuring spoons.

Second, we can use the information that he gives Pedro 6 measuring spoons for his birthday.

This means that Jonathan has 16 - 6 = 10 measuring spoons remaining.

Thus, the total number of utensils that Jonathan has remaining is 24 + 10 = 34.
Therefore, the answer (arabic numerals) is 34.

A: The total number of ut
====================================================================================================
Q: Jonathan has 2/3 as many measuring spoons as measuring cups. If he has two dozen cups and gifts Pedro 6 measuring spoons for his birthday, calculate
the total number of utensils Jonathan has remaining.
A: Let's think step by step. We have that

- 2/3 of the measuring cups = 2/3 × 24 = 16
-
- 2/3 of the measuring spoons = 2/3 × 6 = 4
-
- 16 + 4 = 20
-
- 20 - 6 = 14

Thus, Jonathan has 14 measuring spoons and 20 measuring cups.

In total, he has 14 + 20 = 34 utensils.
Therefore, the answer (arabic numerals) is 34.

Answer: 34
Extracted: 34
True: 34
Correct: True
```

Would you like to receive official Jupyter new
Please read the privacy policy.

*Figure 1.* Generations examples of Quiet-STaR method

# A. Quiet-STaR baseline

## A.1. Citations and References

# Software and Data

# Acknowledgements

# References

Ashkboos, S., Croci, M. L., Nascimento, M. G. d., Hoefler, T., and Hensman, J. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., and Nießner, M. Meshgpt: Generating triangle meshes with decoder-only transformers. *arXiv preprint arXiv:2311.15475*, 2023.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Voynov, O., Bobrovskikh, G., Karpyshev, P., Galochkin, S., Ardelean, A.-T., Bozhenko, A., Karmanova, E., Kopanev, P., Labutin-Rymsho, Y., Rakhimov, R., Safin, A., Serpiva, V., Artemov, A., Burnaev, E., Tsetserukou, D., and Zorin, D. Multi-sensor large-scale dataset for multi-view 3d reconstruction. *CVPR*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. *CVPR*, 2022.

Zelikman, E., Harik, G., Shao, Y., Jayasiri, V., Haber, N., and Goodman, N. D. Quiet-star: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Zhang, Z., Zhang, A., Li, M., hai zhao, Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *ICLR*, 2024.