Unveiling 3D Geometries in LLMs: The Representation and Recall of Periodic Elements

Ge Lei[1], Samuel J. Cooper[1]
[1]Dyson School of Design Engineering, Imperial College London, South Kensington, London SW7 2AZ, United Kingdom

g.lei23@imperial.ac.uk

Abstract

Large language models (LLMs) excel in tasks such as translation and text generation, yet their mechanisms for storing and processing knowledge remain poorly understood. Humans tend to organize interrelated knowledge into adaptive networks, but it is unclear whether LLMs store multiple attributes—such as material and chemical properties—as independent pieces of information or as interconnected features. Understanding these processes is key to exploring LLMs' potential to represent complex properties, identify attribute relationships, and contribute to advancements in materials science.

Mechanistic interpretability aims to reverse-engineer LLMs into human-understandable algorithms. While the linear representation hypothesis suggests that LLMs store knowledge in sparse, linear forms [1], recent studies reveal more intricate structures, such as circular encodings for periodic patterns like days or months [2]. While previous work has primarily examined single-attribute representations, we systematically investigate how LLMs encode and recall complex, multi-attribute knowledge using the periodic table of elements as a case study. Our research explores the interplay and independence of linguistic and factual representations, as well as the geometric forms in which these attributes are encoded. The key findings of our study are:

1. From factual knowledge to language pattern

To investigate how LLMs encode and recall multi-attribute knowledge, we generated prompts using chemical elements and their attributes, such as atomic number and group. Multiple prompt templates were used to create paraphrased variations. Last-token activations from all transformer layers were collected and analyzed through linear probing.

The results demonstrate that the middle layers of LLMs are particularly effective at encoding factual attributes, such as atomic numbers. In contrast, the later layers prioritize context integration and refining outputs into coherent text. This finding underscores a functional transition within the model: the middle layers focus on factual knowledge, while the later layers specialize in transforming this knowledge into linguistically meaningful outputs.

2. Recall ability peaks in middle layers

To investigate recall ability—the model's capacity to retrieve related attribute information regardless of the prompt's focus—we conducted an experiment using misaligned linear probing. Activation datasets were generated from prompts targeting attributes such as atomic number and group, while the linear probe was trained exclusively on group values.

The results show that in the early and middle layers, group information can be effectively recalled even from atomic number-focused prompts, suggesting that different attributes in LLM representations are interrelated. In contrast, recall performance declines significantly in the later layers for misaligned prompts, indicating a shift toward task-specific specialization.

3. Diverse geometric relationships among attributes

We hypothesize that LLMs encode attributes in high-dimensional spaces, potentially forming patterns such as linear, circular, or spiral structures. To validate this, we project activations into a low-dimensional space (assuming a specific geometric shape), adjust them toward a target label, compute the delta, and map it back to the high-dimensional space using the pseudo-inverse weights of a linear probe to predict activations. Through activation patching, we replace specific layer activations with these predicted values during inference, assessing how it influence the model's behavior.

The results show that activation patching can be applied across multiple low-dimensional spaces, indicating that the model's internal representations for related attributes are interconnected. For instance, atomic number information is not confined to a single dimension but is also embedded within features such as group and period. Notably, two 3D spiral structures emerged as particularly effective, marking the first discovery of such geometric patterns in LLMs as shown in Figure 1. This finding aligns with the periodic nature of the periodic table, reinforcing our understanding of its inherent cyclic properties.

4. From superposition to separation

To analyze how attributes are represented and related across layers in LLMs, we conducted two experiments. First, we trained linear models for individual attributes at each layer and calculated cosine similarity between their weight vectors to evaluate the overlap and separation of attribute representations. Second, we trained linear models to map one attribute's representation to another across layers, using R^2 scores to assess the linear relationship between attributes.

The results demonstrate a clear progression in attribute representation across layers. In the middle layers, attributes exhibit significant overlap, reflecting shared and integrated representations. This superposition allows the model to encode multiple

related features simultaneously. However, in the later layers, attributes become increasingly distinct as the model refines their representations for task-specific outputs.

In conclusion, this study investigates how LLM represent and recall multi-associated attributes across transformer layers. We show that middle layers encode factual knowledge by superimposing related attributes in overlapping spaces, facilitating effective recall even when attributes are not explicitly prompted. In contrast, later layers refine linguistic patterns and progressively separate attribute representations, optimizing task-specific outputs at the cost of attribute recall. These findings are consistent with the t-SNE distributions of activations shown in Figure 2. We identify diverse encoding patterns including, for the first time, the observation of 3D spiral structures when exploring information related to the periodic table of elements.

Our findings reveal a transition in elements' attribute representations across layers, providing insights into the mechanistic interpretability of LLMs and their ability to process complex, interrelated knowledge.

References

[1] Gurnee, W. and Tegmark, M. Language models represent space and time. arXiv preprint arXiv:2310.02207, 2023.

[2] Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. Not all language model features are linear. arXiv preprint arXiv:2405.14860, 2024.

Figures

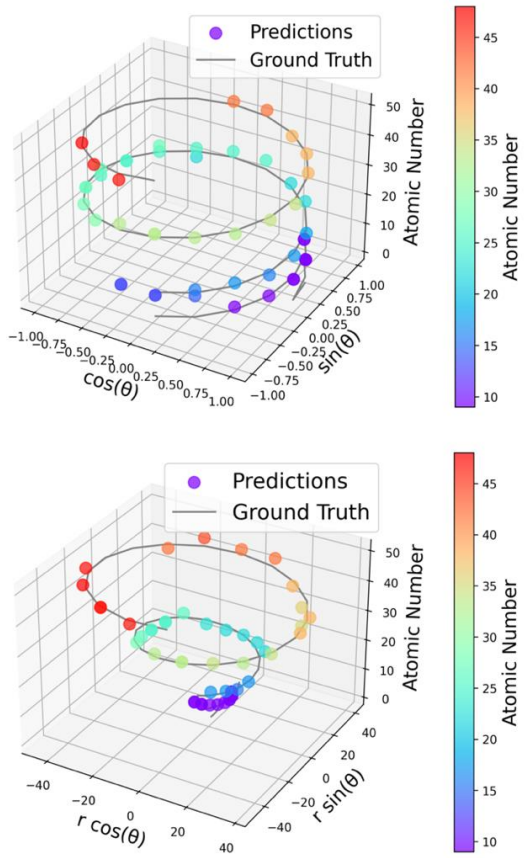

Figure 1. Visualization of predicted atomic numbers in 3D spiral shapes, generated after activation patching.

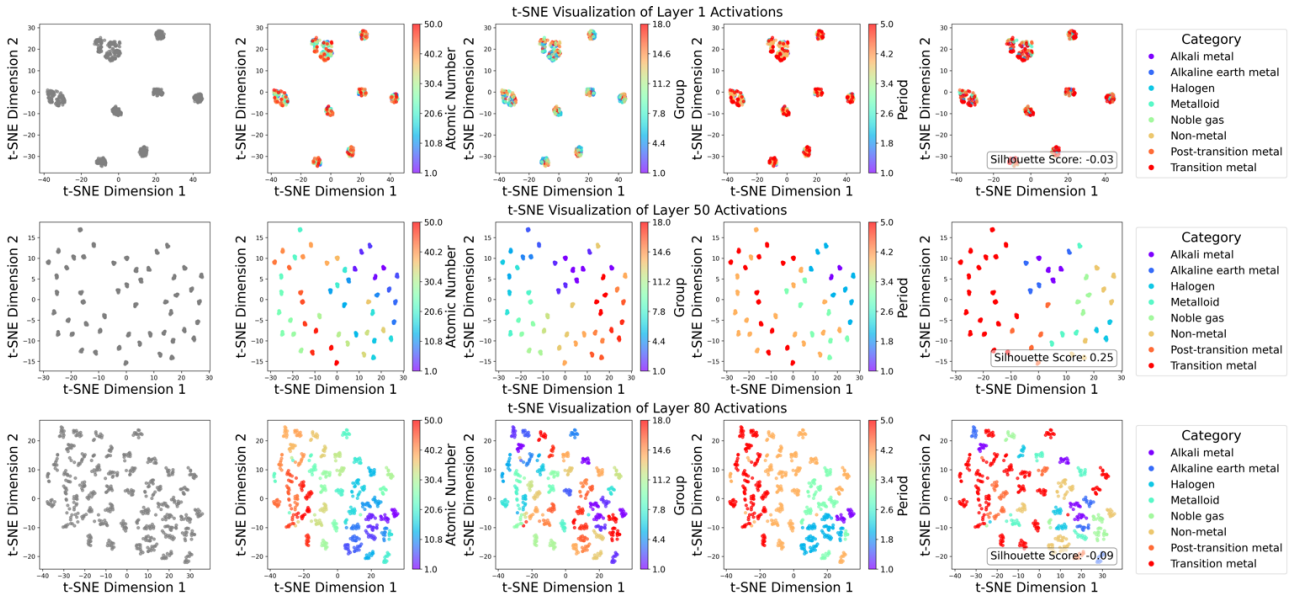

Figure 2. t-SNE visualization of Meta-Llama-3.1-70B last-token activations from the 1st, 50th, and 80th layers, generated using 11 continuation-style templates across 50 elements, focusing on the atomic number attribute. Points are colored by different attributes, revealing how the model encodes and distinguishes unmentioned related properties across layers.