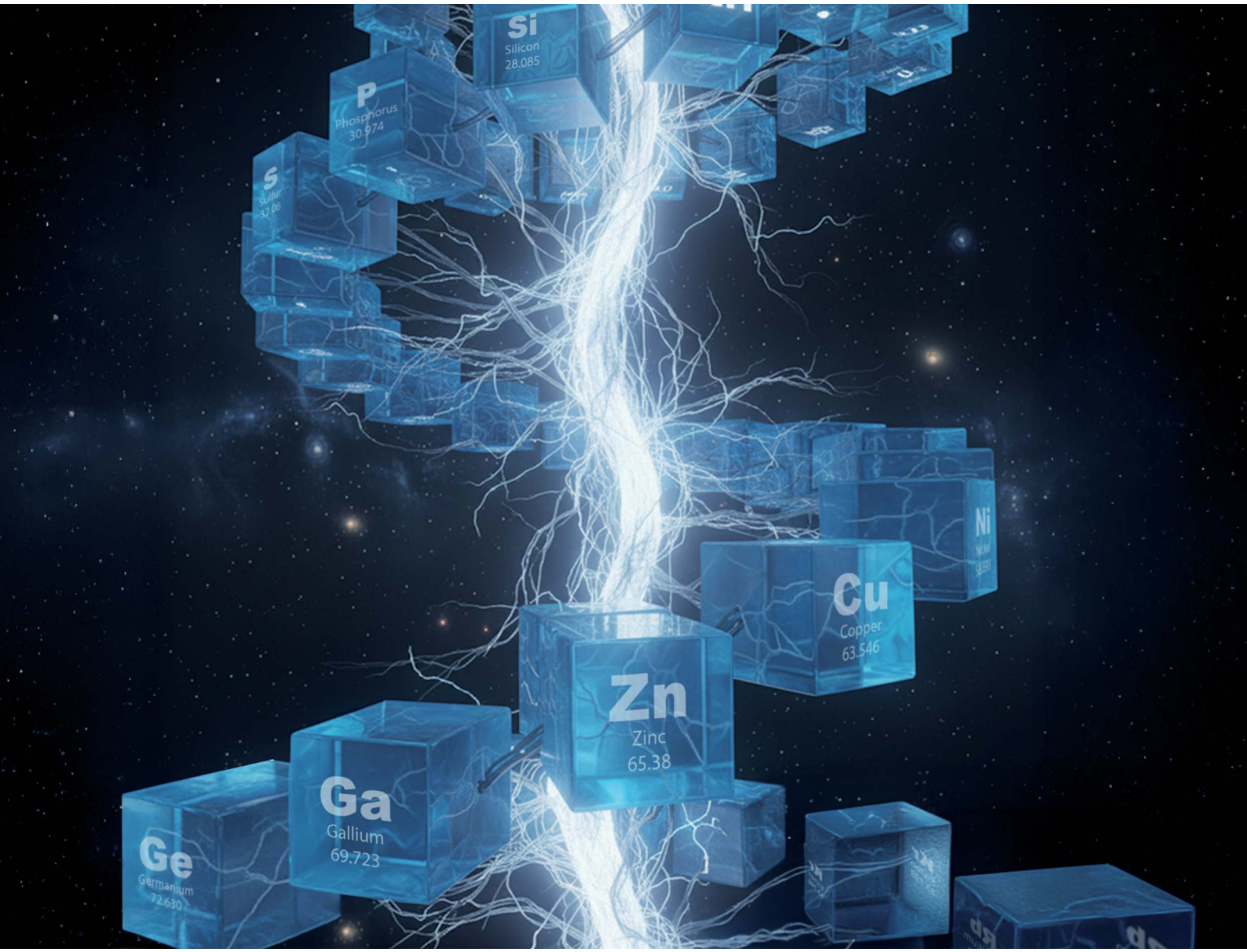
Volume 4
Number 12
December 2025
Pages 3415-3830

Digital Discovery

rsc.li/digitaldiscovery



ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PAPER
Ge Lei and Samuel J. Cooper
Do Llamas understand the periodic table?

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3455

Do Llamas understand the periodic table?

Ge Lei * and Samuel J. Cooper

Large Language Models (LLMs) demonstrate remarkable abilities in synthesizing scientific knowledge, yet their limitations, particularly with basic arithmetic, raise questions about their reliability. As materials science increasingly employs LLMs for tasks like hypothesis generation, understanding how these models encode specialized knowledge becomes crucial. Here, we investigate how the open-source Llama series of LLMs represent the periodic table of elements. We observe a 3D spiral structure in the hidden states of LLMs that aligns with the conceptual structure of the periodic table, suggesting that LLMs can reflect the geometric organization of scientific concepts learned from text. Linear probing reveals that middle layers encode continuous, overlapping attributes that enable indirect recall, while deeper layers sharpen categorical distinctions and incorporate linguistic context. These findings suggest that LLMs represent symbolic knowledge not as isolated facts, but as structured geometric manifolds that intertwine semantic information across layers. We hope this inspires further exploration into the interpretability mechanisms of LLMs within chemistry and materials science, enhancing trust of model reliability, guiding model optimization and tool design, and promoting mutual innovation between science and AI.

Received 20th August 2025
Accepted 9th October 2025

DOI: 10.1039/d5dd00374a

rsc.li/digitaldiscovery

1 Introduction

Large Language Models (LLMs) have demonstrated a notable capacity to synthesize and generate insights from vast amounts of expert knowledge, drawing attention across multiple scientific domains.[1,2] Yet, despite their impressive capabilities, researchers have observed their surprising inability to reliably perform seemingly straightforward tasks, such as basic arithmetic operations.[3–5] This phenomenon highlights an important aspect of LLMs: their fundamental reliance on learned patterns and probabilistic predictions based on token embeddings, rather than explicit arithmetic operations. Consequently, simple numerical tasks, effortlessly handled by even the most rudimentary calculators with orders of magnitude less computation, remain challenging and error-prone for these sophisticated LLMs.

In parallel, interest is rapidly growing in leveraging LLMs within the materials sciences community. Recent research has proposed intriguing applications such as laboratory orchestration,[6–8] hypothesis generation,[9–12] and complex materials property prediction.[13,14] However, despite convincing demonstrations reported in numerous studies, there remains skepticism regarding the reliability and trustworthiness of these systems in rigorous scientific research. One particular risk is that large language models are designed during training to generate responses in ways that align with the user's expectations. This can make the answers appear more authoritative than they actually are, potentially giving users a false sense of confidence in the output, even when the underlying information may be incorrect or fabricated.[15]

This raises a critical question: Can LLMs be trusted to accurately represent chemical information and serve as scientific foundation models? To approach these questions, understanding internal representations is critical. If LLMs simply memorize isolated facts, they would need constant task-specific fine-tuning, effectively functioning as many small models. By contrast, if they organize complex knowledge into structured forms, this would suggest an ability to extract and generalize abstract regularities. More ambitiously, if these representations are geometry-aware and aligned with physical laws, this may suggest that LLMs can capture aspects of universal scientific regularities. Such representations would support compression and generalization, enabling inference of unmentioned properties and transfer across tasks, thereby increasing their potential to contribute meaningfully to scientific research.

In this work, we investigate whether the prominent open-source LLMs, Llamas,[16] store chemical knowledge in a structured and rational manner, whether fragmented into isolated clusters of disconnected facts, or interconnected through rational webs of structured knowledge. We delve into how LLMs encode and recall such knowledge through layer-wise, geometry-aware representations. The contributions of our study are:

(1) We report the first observation of a 3D spiral structure in LLM hidden states that organizes chemical elements in alignment with the structure of the periodic table (Section 4).

(2) We are the first to compare regression and classification probing, showing that middle layers encode continuous

Dyson School of Design Engineering, Imperial College London, SW7 2AZ, UK. E-mail: g.lei23@imperial.ac.uk; samuel.cooper@imperial.ac.uk

attribute structure, while later layers sharpen boundaries for fine-grained decisions (Section 5.1).

(3) We show that linguistic structure increasingly shapes knowledge representations in later layers (Section 5.2).

(4) We find that LLMs recall related attributes through strong linear associations in middle layers, which weaken in deeper layers (Section 6).

2 Related work

LLMs have demonstrated strong factual recall across a wide range of domains, from history and geography to science and mathematics.[17–19] Much still remains unknown about how interconnected knowledge is internally organized in LLMs. Clarifying these mechanisms is vital for aligning LLMs with human values,[20] enhancing their design, and broadening their applications. Mechanistic interpretability offers a pathway to answer these questions.[21–23]

2.1 Superposition

The superposition hypothesis suggests that neural networks can encode far more features than neurons they have by compressing high-dimensional concepts into overlapping, nearly orthogonal directions.[24–26] Instead of assigning features to individual neurons, features are represented as sparse linear combinations across neurons, improving encoding efficiency and reducing interference. Toy models demonstrate that sparsity enhances feature disentanglement, balancing compression and accuracy.[27] Early layers encode numerous features with sparse combinations, while intermediate layers focus on higher-level contextual features.[28]

2.2 Linear representation hypothesis

The linear representation hypothesis suggests that neural networks encode high-level features as linear directions in activation space, enabling easier interpretation and manipulation.[29] Probing, introduced by,[30] assesses feature encoding in models and builds on findings in word embeddings like GloVe and Word2Vec, which capture semantic relationships through linear structures.[31,32] Empirical support spans various contexts, including spatial and temporal representations,[33] sentiment analysis,[34] task-specific features,[35] and broader relational structures.[36]

2.3 Non-linear representations

Although the linear representation hypothesis offers insights into neural network representations, studies have highlighted its limitations and emphasized the significance of non-linear structures. Non-linear structures, such as the 'pizza' and 'clock' patterns,[37,38] and circular representations observed in tasks like predicting days or months using modular arithmetic prompts (considering that every day of the week has one before and after it, so they should be represented as a loop, rather than a line),[39] reveal the complexity of these representations. These observations raise a deeper question: can LLMs represent interwoven knowledge in forms that mirror more complex conceptual geometry in the real world? Following this, we observe a 3D spiral in element representations aligned with periodic trends, suggesting geometric organization of knowledge in LLMs.

2.4 Intermediate layers matter

Recent studies underscore the importance of intermediate layers in LLMs, emphasizing their role in producing more informative representations for downstream tasks compared to final layers.[40–43] These layers are crucial for encoding abstract knowledge, enabling advanced capabilities like in-context learning and transfer learning, which are vital for understanding and optimizing LLMs.[44] Additionally, intermediate layers exhibit distinct patterns of information compression and abstraction, such as reduced entropy, allowing them to efficiently represent complex inputs.[45,46] Building on these findings, our probing results show that intermediate layers encode knowledge in a continuous form, while sharper categorical boundaries emerge in later layers.

2.5 Factual recall

Ref. 47 showed that early MLP layers at the entity token are key to recalling factual associations, while later attention layers propagate this information to the output.[48] expanded this into a three-stage process—enrichment, transfer, and extraction, revealing subject tokens carry multiple implicit attributes.[49] further validated this through detailed circuit analysis, demonstrating entity token representations linearly encode categorical attributes. Most prior studies have focused predominantly on individual attributes, leaving unclear how large language models jointly encode and retrieve multiple, interrelated pieces of information. However, chemical knowledge typically exhibits structured complexity characterized by interconnected relationships, demanding integrated and holistic representations. Motivated by this, we delve into how LLMs encode and recall complex, interwoven chemical knowledge through global, structured, and geometry-aware representations that evolve across the model's depth.

3 Preliminaries

Our study only focuses on how reliably acquired knowledge (*i.e.* things we're confident the model knows) is represented within LLMs, and excludes hallucinations or information not in the training set. We use the properties of chemical elements in the periodic table as a case study due to their frequent occurrence in training data, well-defined attributes, quantifiable properties, and making them an ideal subject for this investigation. We adopt Llama series models[16,50] in this study.

3.1 Residual stream collection

To study how LLMs represent attributes across layers, we construct a prompt dataset based on a set of attributes ($A = \{A_j\}_{j=1}^{M}$, such as 'atomic number' or 'group') and a set of elements ($X = \{X_i\}_{i=1}^{N}$, constituting the first 50 elements, such as 'Mg' or 'Al'). For linguistic diversity, we incorporate predefined template sets: $T^{\text{cont}} = \{T_k^{\text{cont}}\}_{k=1}^{11}$ for continuation-style prompts and $T^{\text{ques}} = \{T_k^{\text{ques}}\}_{k=1}^{11}$ for question-style prompts,

with 11 templates in each. The complete list of templates are provided in Appendix A.

In the continuation-style templates, the next output token would be the factual knowledge directly such as:

$$T_1^{\text{cont}}(A_j, X_i) = \text{'The } A_j \text{ of } X_i \text{ is' (}e.g.\text{, 'The atomic number of Al is')}$$

$$T_2^{\text{cont}}(A_j, X_i) = \text{'}X_i\text{'s } A_j \text{ is' (}e.g.\text{, 'Al's atomic number is')}$$

In question-style templates, the next output token is typically a syntactic word like 'The', which ensures the grammatical structure is correct, such as:

$$T_1^{\text{ques}}(A_j, X_i) = \text{'What is the } A_j \text{ of } X_i\text{?' (}e.g.\text{, 'What is the atomic number of Al?')}$$

$$T_2^{\text{ques}}(A_j, X_i) = \text{'Which value represents } X_i\text{'s } A_j\text{?' (}e.g.\text{, 'Which value represents Al's atomic number?')}$$

By substituting each element and attribute (X_i, A_j) into these templates, we generate prompts:

$$p_{i,j,k} = T_k(X_i, A_j)$$

Each prompt $p_{i,j,k}$ can then be fed into LLMs to study the corresponding residual streams at different layers. Last-token residual streams capture the full prompt context in decoder-only models with masked attention, as they integrate information from all preceding tokens. For each layer l, we collect last-token residual streams $\mathbf{h}_{i,j,k}^{(l)}$ from prompts $p_{i,j,k}$ across all elements and templates:

$$\mathbf{h}_{i,j,k}^{(l)} = f^{(l)}(p_{i,j,k}) \in \mathbb{R}^{T\times d},$$

where $f^{(l)}$ denotes the layer-l transformation, T is the token length of the prompt, and d is the hidden dimension. The initial residual stream $\mathbf{h}_{i,j,k}^{(0)}$ is obtained by embedding the prompt through an embedding layer E_0, followed by processing through L Transformer layers. Each layer applies multi-head attention and a feedforward network with residual connections and layer normalization:

$$\mathbf{h}_{i,j,k}^{(l)'} = \mathbf{h}_{i,j,k}^{(l-1)} + \text{Attention}\left(\text{Norm}\left(\mathbf{h}_{i,j,k}^{(l-1)}\right)\right)$$

$$\mathbf{h}_{i,j,k}^{(l)} = \mathbf{h}_{i,j,k}^{(l)'} + \text{FFN}\left(\text{Norm}\left(\mathbf{h}_{i,j,k}^{(l)'}\right)\right)$$

Finally, $\mathbf{h}_{i,j,k}^{(L)}$ is mapped to the vocabulary space using the vocabulary head $\mathbf{W}_{\text{vocab}}$ to produce logits:

$$\text{logits}_{i,j,k} = \mathbf{h}_{i,j,k}^{(L)}\mathbf{W}_{\text{vocab}}$$

By analyzing last-token residual streams $\mathbf{h}_{i,j,k}^{(l)}$ across layers, we investigate how attributes are represented in the model's hidden states.

3.2 Residual stream distribution

We start with a preliminary visualization of the distribution of last-token residual streams for the 'atomic number' attribute. Residual streams from each transformer layer l were collected for the atomic number attribute across the first 50 elements using 11 continuation-style templates, forming the set $H_{\text{atomic number}}^{(l)}$. During our preliminary investigations we had noticed that Llama 7B would occasionally misremember facts about the heavier elements, which is presumably because they appear less frequently in the training set. As such, we choose to investigate just the first 50 elements (where all models are confident) to avoid confounding our analysis with factual errors.

To enable informative plots to be produced efficiently, PCA was applied to each member of H and then t-SNE was use to project the first 50 principal components into 2D. Fig. 1 shows the resulting distributions for Meta-Llama-3.1-70B, with points colored by atomic number as well as the other attributes (to visualize their association to atomic number).

The first column of the figure colors residual streams by true atomic number values (explicitly requested in the prompt). In early layers, prompts with similar template (*i.e.* sentence

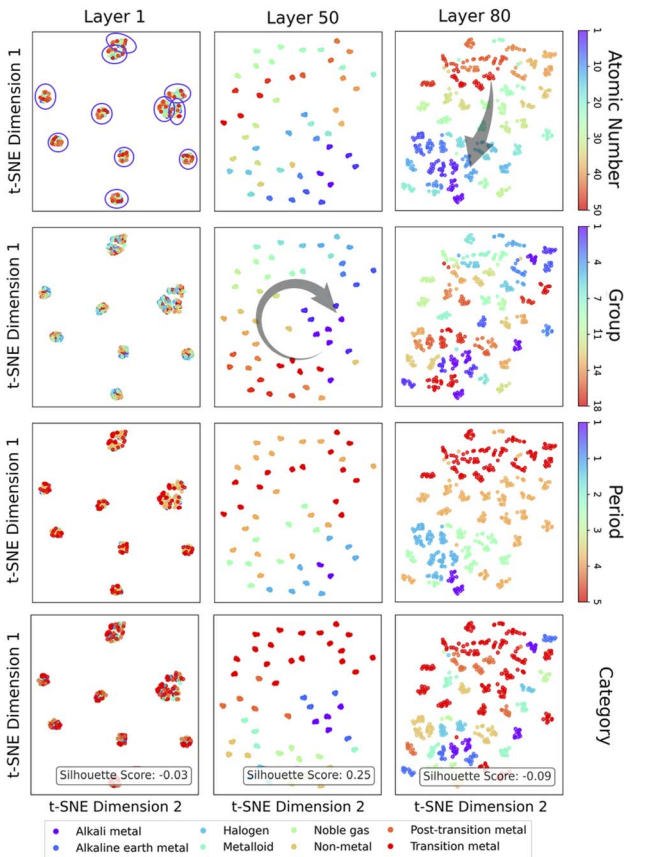


Fig. 1 t-SNE visualization of Meta-Llama-3.1-70B last-token residual streams from the 1st, 50th, and 80th layers, using 11 continuation-style templates across the first 50 elements (550 points per plot). Each column shows one layer, while rows represent different colormaps highlighting attributes: 'atomic number', 'group', 'period', and 'category'. In the top-left plot, circled clusters correspond to individual templates, each containing 50 points.

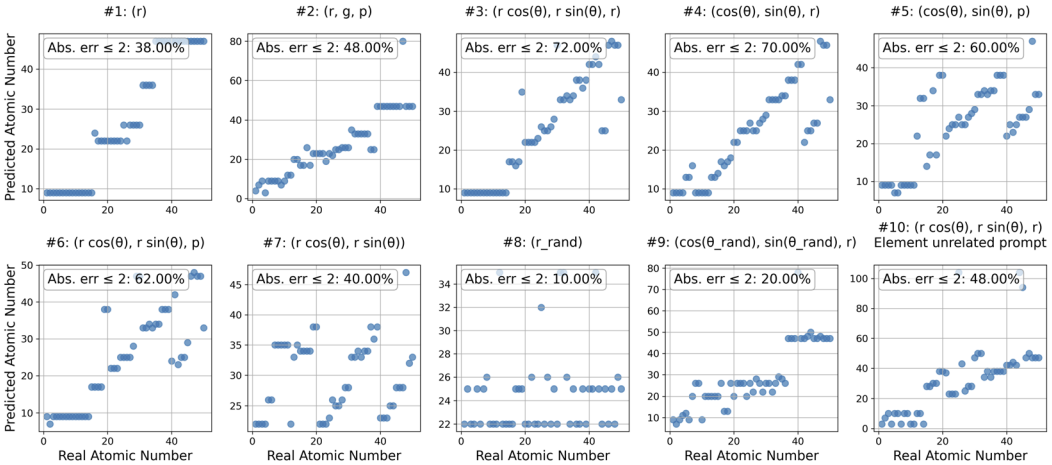


Fig. 2 Residual stream patching results for layer 20 in Meta-Llama-3.1-70B. The model's predictions are evaluated after replacing the residual stream of the 'element' token at the last token position with the predicted residual stream $\hat{\mathbf{h}}_0^{\text{pred},(20)}$.

structure and vocabulary) cluster together irrespective of atomic number, reflecting token-level similarity. In the middle layers, residual streams for the same element form tight clusters for each element, containing each of the 11 prompt templates. By the final layers, although the same elements still cluster together, individual points become distinguishable due to increased spread within the clusters.

In the next three rows, we show the same results, but the residual streams are colored by the true values of attributes unmentioned in the prompt: 'group', 'period', and 'category'. Despite not being mentioned in the prompt, in middle layers, chemically similar elements (small clusters with similar colors) cluster closely together. By the final layers, the clustering of some attributes, such as group and category, becomes less coherent, indicating a shift in representation.

Furthermore, the geometric shape of attribute distributions varies. For example, atomic numbers form a linear arrangement transitioning from red to purple (1st row, 3rd column), while the 'group' attribute activities form a cyclic pattern with sequential transitions (2nd row, 2nd column), potentially reflecting periodic relationships.

These visualizations suggest that LLM residual streams may encode attribute relationships in a structured and potentially geometric manner that reflect properties of the physical world. In particular, intermediate layers appear to capture implicit similarities even for unmentioned attributes, while later layers are tuned for task-specific outputs.

t-SNE visualizations are sensitive to hyperparameters such as perplexity and initialization. To this end, Fig. 1 is presented only as an intuitive aid for visualizing abstract representational patterns and for motivating our hypotheses. These hypotheses are examined in the following sections.

4 Geometric relationships among attributes

Previous work has argued that 2D spiral representations of the periodic table provide a more natural visualization of periodicity by arranging elements sequentially along a continuous polar axis, where recurring properties manifest as repeated turns of the spiral.[51] Unlike the conventional tabular layout that enforces line breaks, the 2D spiral highlights both the continuity of atomic number progression and the grouping of chemically similar elements along radial directions. Extending this idea, 3D spiral constructions of the periodic table have also been proposed, in which elements are mapped onto a helical structure, offering additional dimensionality to emphasize periodic trends and inter-element relationships.[52] Motivated by these connections, we investigate whether LLMs—having been exposed during training to extensive data about the properties of elements—inherently capture such physical periodicities and reflect analogous spiral structures in their learned embeddings.

We hypothesize that attributes in LLMs exist in a high-dimensional space, manifesting as linear, circular, or spiral patterns based on their structure, and then proceed to validate these geometries.

Inspired by ref. 39, we map the last-token residual streams $\mathbf{h}^{(l)} \in \mathbb{R}^k$ at layer l to a geometric space $f(r, g, p)$, which encodes atomic number r, group g, and period p. To learn this mapping, we first reduce the dimensionality of the residual streams to 30 using PCA, denoted as $\mathbf{P}(\mathbf{h}^{(l)})$, and fit a linear projection using all 50 elements except one held-out target:

$$\mathbf{W}^{(l)}, \mathbf{b}^{(l)} = \underset{\mathbf{W}', \mathbf{b}'}{\operatorname{argmin}} \sum_{i \neq 0} \left\| \mathbf{W}' \mathbf{P}\left(\mathbf{h}_i^{(l)}\right) + \mathbf{b}' - f_i \right\|_2^2$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d' \times 30}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{d'}$, and $f_i = f(r_i, g_i, p_i)$ denotes the mapping of the i-th element in the geometric space.

To perform the intervention, we compute the centroid of the PCA-reduced residual streams for the remaining $N = K - 1$ elements:

$$\bar{\mathbf{h}}^{(l)} = \frac{1}{N} \sum_{i \neq 0} \mathbf{P}\left(\mathbf{h}_i^{(l)}\right)$$

then map it to the geometric space: $\mathbf{z} = \mathbf{W}^{(l)} \bar{\mathbf{h}}^{(l)} + \mathbf{b}^{(l)}$. Let $f_0 = f(r_0, g_0, p_0)$ denote the target element's embedding in the geometric

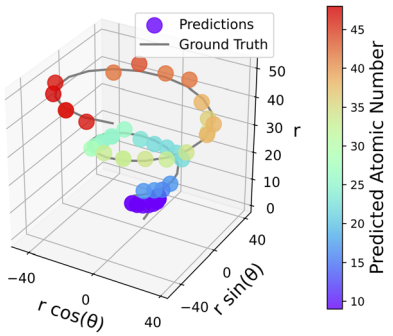


Fig. 3 Predicted atomic numbers after intervention in 3D spiral space ($r\cos\theta$, $r\sin\theta$, r). Colored points indicate the tokens with highest logits.

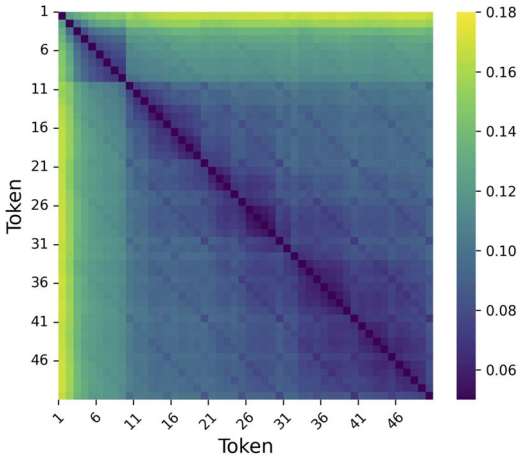


Fig. 4 Euclidean distance heatmap of approximated vector representations for numeric tokens (1–50) in the hidden space of the last layer in Meta-Llama-3.1-70B.

space. The deviation $f_0 - \mathbf{z}$ is projected back to the residual stream space using the pseudo-inverse of $\mathbf{W}^{(l)}$, giving the predicted (intervened) residual stream:

$$\hat{\mathbf{h}}_0^{\text{pred},(l)} = \mathbf{P}^{-1}(\bar{\mathbf{h}}^{(l)} + (\mathbf{W}^{(l)})^{+}(f_0 - \mathbf{z}))$$

Importantly, the model never accesses the original residual stream of the target element; the predicted residual stream is computed solely from its geometric representation and the residual streams of other elements. During inference, we replace the residual stream of 'element' (last token position) in the 20th layer† with $\hat{\mathbf{h}}_0^{\text{pred},(20)}$, using the prompt 'In the periodic table, the atomic number of element'. We then evaluate whether the model can correctly output the target token without ever seeing its original residual stream.

We evaluate the effectiveness of different geometric spaces for interventions, including linear, 2D spiral, and 3D spiral (*i.e.* conical helix) geometries, as shown in Fig. 3. Angular variables $\theta = \frac{2\pi g}{18}$ are used to capture periodic relationships. To test the impact of disrupted geometry, two random spaces are introduced: in Space 8, atomic numbers r are shuffled; in Space 9, θ is randomly permuted. Additionally, in Space 10, the prompt 'In numbers, the Arabic numeral for number' generates numbers 1–50, testing whether periodic patterns emerge without explicit element references. We designed this control to examine whether the observed geometric shapes arise from element-related knowledge or simply from numerical sequences.

Effective residual stream patching suggests that the target space $f(r, g, p)$: (1) retains sufficient information for accurate reconstruction during transformations with the residual stream space, and (2) preserves geometric structures similar to those in the residual stream space to ensure valid adjustments in the high-dimensional space.

Patching results for Meta-Llama-3.1-70B are shown in Fig. 2, with detailed values in Table B.1 (Appendix). Results show that intervention can be applied in various geometric spaces, with some performing significantly better. Spaces such as ($\cos\theta$, $\sin\theta$, r) and ($r\cos\theta$, $r\sin\theta$, r) over 70% predictions of the atomic number have an absolute error within 2, suggesting the potential existence of latent 3D structures in LLMs resembling 3D spirals. Fig. 3 illustrates the LLM's output post-intervention in 3D spiral geometry. Additional geometric analyses are in Appendix B.2. Randomly generated prompts perform very poorly, which is expected given their lack of coherent semantic structure and context, but even element unrelated prompts with clear linguistic form also yield poor performance. This suggests that the geometry of the embedding space is not merely tied to numerical correlations or surface-level semantics, but is inherently aligned with the background knowledge invoked by the prompt, reflecting real-world knowledge structures.

In a concurrent study,[53] observed spiral-like structures in number space with periods of 2, 5, 10, and 100, likely reflecting common human conventions in numerical representation. In contrast, our model exhibits a distinct 18-period spiral‡ aligned with the periodic structure of chemical elements. This representation performs notably worse for ordinary numbers without elemental context (which aligns with their observation that the 18-period does not prominently emerge), indicating that such geometric patterns emerge from underlying physical or semantic regularities rather than arbitrary structures.

In the intervention experiments, it is actually not obvious whether a smaller numerical difference between the output token and the true value always implies smaller error. To investigate this, we project token IDs for numbers 1–50 into the last hidden layer using the pseudoinverse of the vocabulary projection matrix $\mathbf{W}_{\text{vocab}}^{+}$. This operation reconstructs an approximation of the hidden representations that would produce these token IDs as logits. Fig. 4 shows that smaller numerical differences generally correspond to closer

† See Appendix B.1 for details on intervention performance. Interventions become effective from layer 20 onward.

‡ In our study, each period is arranged on an 18-sector scale corresponding to the 18 groups; in the first three periods, only 2, 8, 8 sectors are filled, with the remaining sectors left empty. Thus, the finer 2/8/8/18 period is preserved within the 18-sector framework.

representations, while larger differences often result in inconsistent distances, reflecting the model's difficulty with numerical consistency over larger gaps. For instance, the vector for '1' is closer to '2' than to '5', while the distances between '10' and '40' is closer than between '10' and '21'. In the intervention, when the predicted value is close to the true value, hidden logits align well with true logits, suggesting higher accuracy. However, large numerical deviations cannot fully capture prediction errors, so we evaluate results using an absolute error threshold (≤ 2) in Fig. 2, representing a small distance.

5 Direct attribute recall

In the previous section, we observed that elemental knowledge in LLMs forms a 3D spiral structure. Interestingly, although prompts mentioned only atomic numbers, the embeddings also reflected elemental groups, suggesting that LLMs retrieve both explicitly requested and implicitly related attributes. To better understand these mechanisms, this section investigates direct attribute knowledge recall and Section 6 will explores how LLMs access related but unprompted knowledge.

5.1 From continuity to boundary sharpening

Some elemental attributes, such as group and period, naturally exist in both categorical and numerical forms. This duality enables both classification and regression probing, allowing for direct comparisons that have been underexplored in prior work, which often focused exclusively on a single type.

To examine how LLMs access explicitly mentioned knowledge, we use the last-token residual stream from the continuation style prompt $\mathbf{h}_j^{(l)} \in \mathbb{R}^k$ as the representation of attribute A_j, and fit a linear probe to predict its corresponding values *via*:

$$f_j^{(l)}(\mathbf{h}_j) = \mathbf{W}_j^{(l)}\mathbf{h}_j^{(l)} + \mathbf{b}_j^{(l)}$$

For categorical attribute forms (*e.g.*, category, group, period), $\mathbf{W}_j^{(l)} \in \mathbb{R}^{|C_j| \times k}$, $\mathbf{b}_j^{(l)} \in \mathbb{R}^{|C_j|}$. Predictions are made by:

$$\hat{y}^{(l)} = \underset{c \in C_j}{\operatorname{argmax}} \left[f_j^{(l)}\left(\mathbf{h}_j^{(l)}\right) \right]_c .$$

For continuous attributes, we perform scalar regression by setting $\mathbf{W}_j^{(l)} = \mathbf{w}_j^{(l)\top}$, $\mathbf{w}_j^{(l)} \in \mathbb{R}^k$, $\mathbf{b}_j^{(l)} \in \mathbb{R}$, yielding:

$$\hat{y}^{(l)} = \mathbf{w}_j^{(l)\top}\mathbf{h}_j^{(l)} + b_j^{(l)}$$

Probes are trained using 5-fold cross-validation on last-token residual streams. We use a linear Support Vector Machine (SVM) for categorical tasks and Support Vector Regression (SVR) with a linear kernel for continuous tasks. The resulting classification accuracies and regression R^2 scores are shown in Fig. 5, with best-layer results provided in Appendix F.5.

Regression probes reveal that continuous numerical features are effectively represented in intermediate layers, as indicated by high R^2 values (while not reaching 1, see Appendix F.1). These intermediate layers sometimes even outperform the final

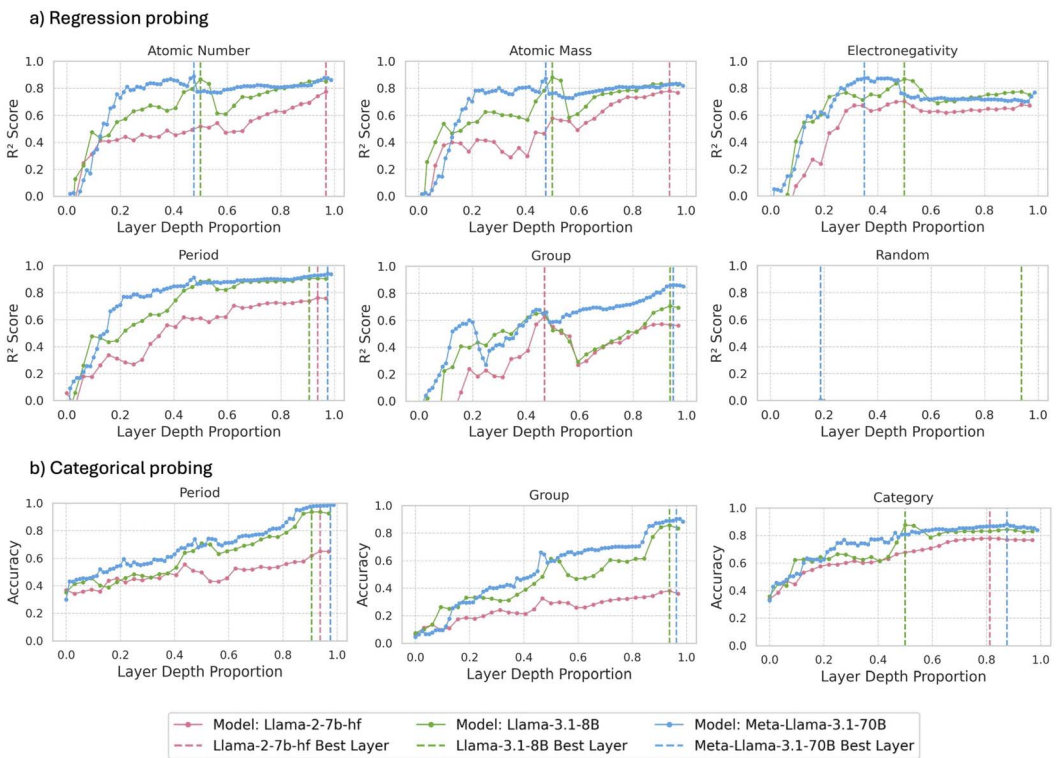


Fig. 5 Linear probing results on last token across layers. (a) Regression (R^2) for numerical attributes and a random baseline. (b) Classification (accuracy) for categorical attributes. All results use 5-fold cross-validation on last-token residual streams.

layers, suggesting that numerical knowledge is already encoded before the final output stage. This aligns with findings by ref. 47, which show that factual knowledge recall is already mediated by intermediate MLP layers.

In classification probes, intermediate layers perform similarly or even better than final layers for clearly distinct non-numerical categories (*e.g.*, metal *vs.* non-metal), aligning with prior work.[49] However, they significantly underperform in fine-grained numerical classification, *e.g.*, period accuracy drops from ∼1.0 (final) to ∼0.7, and group from ∼1.0 to ∼0.6.

This suggests that while intermediate layers already encode meaningful numerical structure, additional processing in later layers is required to sharpen boundaries and support accurate fine-grained classification. This aligns with intuition: later layers prepare for discrete token outputs, where clearer classification boundaries must emerge. As shown in Appendix Fig. F.1, the confusion matrix from Layer 40 (70B middle layer) is not perfectly accurate, but most misclassifications fall near the diagonal, further demonstrating that intermediate layers encode coherent numerical structure, albeit with blurred categorical boundaries. These observations may provide useful insights for choosing between intermediate and later-layer embeddings in downstream tasks.

Notably, Llama2 7B shows low accuracy (<0.4) on group classification compared to Llama3.1 8B (>0.8) (but similar performance in group regression probing) potentially due to its single number tokenization (splitting numbers like '12' into '1' and '2'), which may cause confusion between the representations of output tokens like '12' and '1'. In contrast, Llama 3 uses separate tokens for numbers below 1000.

5.2 Higher language sensitivity in later layers

The sharpening of numerical representations into categorical boundaries in later layers suggests that these layers might be shaped by the expected output tokens. This raises a question: does the linguistic structure influence the factual representations across layers?

We compared question-style and continuation-style prompts using linear regression probes. Continuation prompts generally lead to direct generation of fact-related tokens, whereas question-style prompts tend to introduce syntactic fillers (*e.g.*, 'The') and are more influenced by superficial language patterns.

Fig. 6 reports the average delta R^2 across five attributes, with per-attribute results shown in Fig. F.2 (Appendix). As analyses in earlier sections show stronger semantic signals and higher R^2 in mid-to-late layers, we focus on depths 0.6–1.0. ΔR^2 increases in the mid-to-late layers, indicating a growing gap between prompt types. Among the 15 attribute–model combinations (3 models × 5 attributes), 12 show a significant increasing trend (FDR-corrected $p < 0.05$), with a median Mann–Kendall τ of 0.55 (Appendix F.3).

The results indicate that, as depth increases, question prompts become progressively less effective than continuation prompts at encoding factual attributes, hinting that the prompt's linguistic structure exerts a stronger influence on representations in deeper layers. Interestingly, the larger

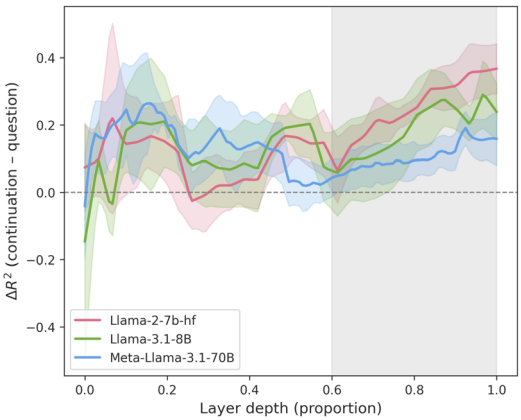


Fig. 6 Average ΔR^2 across five attributes, with 95% confidence interval shaded. $\Delta R^2 = {R_{\text{cont}}}^2 - {R_{\text{ques}}}^2$.

models show a slower increase in ΔR^2 across layers than the smaller models, suggesting they maintain more stable factual representations across prompt types and thus exhibit a smaller distinction between continuation and question prompts.

The rising ΔR^2 suggests that deeper layers increasingly blend factual content with linguistic structure to prepare the final tokens. To further test this, we applied the logit lens[54] and tuned-lens.[55] These analyses estimate the token distribution each layer would produce if decoding were halted at that depth, and show that the correct numerical token becomes highly ranked only in the later layers (Appendix D). Complementary attention statistics (Appendix C) reveal that mid-layers focus tightly on the factual token, whereas later layers spread attention over a wider context patterns consistent with increased syntactic and contextual integration.

6 Indirect attribute recall

In the previous section, we analyzed direct recall of explicitly mentioned attributes across layers. Our earlier geometric analysis showed that LLMs can also recall related attributes that are not explicitly mentioned. In this section, we explore how related but unmentioned attributes are recalled.

6.1 Middle layers excel at indirect recall

We conducted experiments using linear probing to examine the relationships between distinct attributes. Specifically, we extracted last-token residual streams from continuation prompts that mention attribute A_{j_1} (matching) or a different attribute A_{j_2} (non-matching), *i.e.* seeing if we can extract information that was not explicitly requested in the prompt. We also extracted the residual stream at the element token position, before any attribute is introduced (no mention). Separate probes were trained for each residual stream dataset, always using labels of attribute A_{j_1} as targets. To avoid confounding factors, we selected six attribute pairs without direct linear relationships for non matching probe (see Appendix F.4.1). Average R^2 curves for all attributes are shown in Fig. 7; detailed

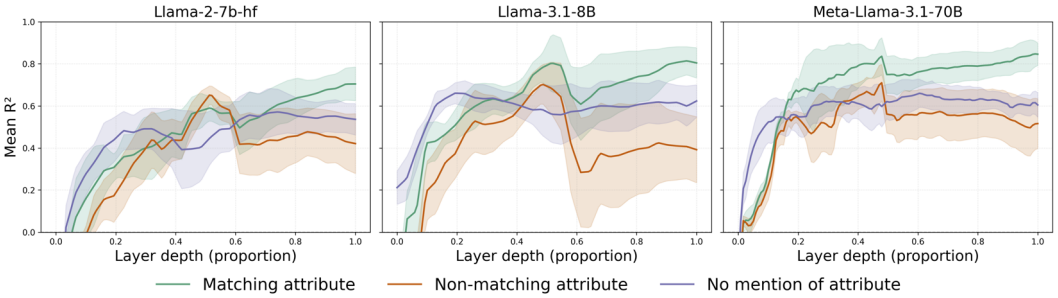


Fig. 7 Average R^2 scores from regression probing across layers for three prompt types: matching, non-matching, and no-mention. All probes predict a fixed target attribute A_{j_1}; no-mention uses element token residual streams before any attribute appears. Shaded areas show 95% confidence intervals.

case-wise linear probing results appear in Appendix Fig. F.3 and F.4.

Attribute information was detectable across all prompt styles. Intuitively, matching prompts should perform best by providing explicit cues, no-mention comes next as it relies on inference, and non-matching prompts perform worst due to misleading signals. Surprisingly, at intermediate layers (around 0.5 depth), non-matching prompts yielded higher linear R^2 scores than no-mention prompts, suggesting stronger inter-attribute interactions at these depths. This may reflect entangled representations between related attributes, which we analyze further in Section 6.2.

Beyond 60% depth, performance follows the expected trend: matching > no-mention > non-matching. The gap between matching and non-matching prompts increases steadily from 0.6 to 1.0 depth. Across 15 model–attribute tests, 14 exhibited statistically significant divergence (FDR corrected $p < 0.05$), with a median Mann–Kendall τ of 0.77 (Appendix Fig. F.5 and Table F.3). It suggests that attribute representations become more specialized and context-sensitive in deeper layers. Further analyses in Section 6.2 provide a more direct explanation, examining how structural relationships between attributes contribute to this layered specialization.

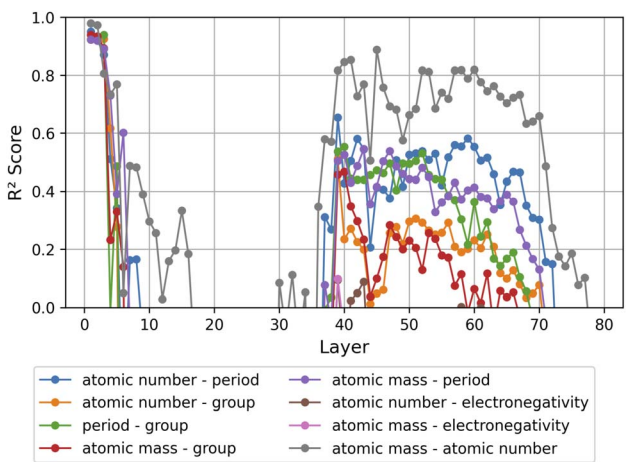


Fig. 8 R_2 scores across layers in Meta-Llama-3.1-70B for linear mappings between attribute pairs using the final residual stream from a fixed prompt.

The fact that the 'no-mention' prompts perform best in the early layers may seem counterintuitive; however, this is likely because, unlike the other two scenarios, in the 'no-mention' case, the last token is the element itself, which may aid recall. In contrast, matching prompts extract residual streams at the final token (such as 'is'), requiring holistic semantic understanding. As layer depth increases, semantic clarity improves, enhancing explicitly mentioned attributes and reversing this initial trend.

To explicitly capture relationships between attribute representation, we train a linear mapping from the representation of attribute A_{j_1} to attribute A_{j_2} at each model layer. Specifically, we utilize the final residual streams from a fixed prompt template (after applying PCA to reduce the dimensionality to 20). The mapping performance is evaluated using R^2 scores obtained *via* 5-fold cross-validation.

Fig. 8 illustrates the variation of R^2 scores across layers for different attribute pairs. In early layers, R^2 scores are high; however, this observation alone does not necessarily indicate meaningful attribute-level relationships, as initial representations are predominantly sensitive to token-level similarity (see t-SNE analysis in Fig. 1). Due to the use of a fixed input template, the resulting inputs exhibit substantial token-level overlap.

6.2 Stronger linear correlations in middle layers

In the intermediate layers, where concept-level understanding is evident (as shown by t-SNE and linear probing), we observe a peak in R^2 scores. This indicates that even simple linear models can effectively capture relationships between different attributes, reflecting their connection in the learned representation space. This also explains why prompts with non-matching attributes outperform those with no attribute mention at these layers in the last Section 6.1. In deeper layers, R^2 scores decline, suggesting a shift toward specialized representations. Similar conclusions from the linear probing weight analysis further support this, as shown in Appendix E.1.

7 Discussion and conclusions

This study highlights that despite their exclusive reliance on textual training data, LLMs internally develop structured representations aligning closely with scientific knowledge. Specifically, we observe a 3D spiral structure within the hidden states of LLMs

that mirrors the conceptual organization of the periodic table, indicating the models' implicit grasp of domain-specific regularities without explicit supervision.

Probing experiments reveal that the encoding of chemical knowledge evolves across model depth: middle layers encode continuous, overlapping attribute subspaces suitable for coarse categorization, while deeper layers sharpen decision boundaries and integrate linguistic structure. In addition, we find that related attributes are strongly linearly associated in middle layers, enabling indirect recall.

Our results demonstrate that symbolic scientific knowledge, particularly in chemistry, is represented within LLMs as coherent, geometry-aware manifolds where conceptual information is systematically intertwined across model layers. Furthermore, this geometric structure aligns with the laws observed in the physical world, indicating that knowledge within LLMs is not arbitrary, but rather organized and reflective of inherent natural order. Moreover, it is unsurprising that these large models discover meaningful relationships between concepts and these must often represent efficient compression.

We hope this work inspires further investigation into how LLMs represent and reason about scientific knowledge, such as materials property prediction, and informs the design of downstream embedding-based tasks. We believe interpretability in LLMs is essential for AI safety, reducing unintended behaviors and building trust. Understanding how knowledge is stored and recalled across layers can inspire more interpretable, efficient models, advance knowledge editing and scientific discovery.

7.1 Limitations

Our prompts have specifically targeted chemical elements in the periodic table; future studies could expand this to include other chemical structures and properties. The hypothesis-driven validation of geometric structures may oversimplify LLMs' non-linear interactions.

Conflicts of interest

There are no conflicts to declare.

Data availability

The code required to reproduce the results presented in this paper is available at **https://github.com/tldr-group/LLM-knowledge-representation** with an MIT license agreement and archived on Zenodo with DOI: **https://doi.org/10.5281/zenodo.17280841**.

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00374a**.

Acknowledgements

This project has received funding from the European Union's research and innovation programme Horizon Europe under the grant agreement No. 101192848 and the Imperial Lee Family Scholarship. We would like to thank the members of the TLDR group for their valuable comments and insightful discussions.

References

1 O. Wysocki, M. Wysocka, D. Carvalho, A. T. Bogatu, D. M. Gusicuma, M. Delmas, *et al.*, An llm-based knowledge synthesis and scientific reasoning framework for biomedical discovery, *arXiv*, 2024, preprint, arXiv:2406.18626, DOI: **10.48550/arXiv.2406.18626**.

2 G. Lei, R. Docherty and S. J. Cooper, Materials science in the era of large language models: a perspective, *Digital Discovery*, 2024, **3**(7), 1257–1272.

3 J. Qian, H. Wang, Z. Li, S. Li and X. Yan, Limitations of language models in arithmetic and symbolic induction, *arXiv*, 2022, preprint, arXiv:2208.05051, DOI: **10.48550/arXiv.2208.05051**.

4 T. Baeumel, J. van Genabith and S. Ostermann, The lookahead limitation: Why multi-operand addition is hard for llms, *arXiv*, 2025, preprint, arXiv:2502.19981, DOI: **10.48550/arXiv.2502.19981**.

5 A. Gambardella, Y. Iwasawa and Y. Matsuo, Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks, *arXiv*, 2024, preprint, arXiv:2406.02356, DOI: **10.48550/arXiv.2406.02356**.

6 M. Sim, M. G. Vakili, F. Strieth-Kalthoff, H. Hao, R. J. Hickman, S. Miret, *et al.*, ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories, *Matter*, 2024, **7**(9), 2959–2977.

7 K. Darvish, M. Skreta, Y. Zhao, N. Yoshikawa, S. Som, M. Bogdanovic, *et al.*, ORGANA: A robotic assistant for automated chemistry experimentation and characterization, *Matter*, 2025, **8**(2), 101897.

8 E. A. Olowe and D. Chitnis, LABIIUM: AI-Enhanced Zero-configuration Measurement Automation System, in *2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, IEEE, 2025, pp. 1–6.

9 Q. Liu, M. P. Polak, S. Y. Kim, M. A. A. Shuvo, H. S. Deodhar, J. Han, *et al.*, Beyond designer's knowledge: Generating materials design hypotheses via a large language model, *Acta Mater.*, 2025, 121307.

10 S. Kumbhar, V. Mishra, K. Coutinho, D. Handa, A. Iquebal and C. Baral, Hypothesis generation for materials discovery and design using goal-driven and constraint-guided llm agents, *arXiv*, 2025, preprint, arXiv:2501.13299, DOI: **10.48550/arXiv.2501.13299**.

11 A. Bazgir, Y. Zhang, *et al.*, Agentichypothesis: A survey on hypothesis generation using llm systems, *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025.

12 A. Bazgir, Y. Zhang, *et al.*, MatAgent: A human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle, in *AI for Accelerated Materials Design-ICLR 2025*, 2025.

13 S. Liu, T. Wen, B. Ye, Z. Li, H. Liu, Y. Ren, *et al.*, Large language models for material property predictions: elastic

constant tensor prediction and materials design, *Digital Discovery*, 2025, **4**(6), 1625–1638.
14 A. N. Rubungo, K. Li, J. Hattrick-Simpers and A. B. Dieng, LLM4Mat-bench: benchmarking large language models for materials property prediction, *Mach. Learn.: Sci. Technol.*, 2025, **6**(2), 020501.
15 M. Steyvers, H. Tejeda, A. Kumar, C. Belem, S. Karny, X. Hu, *et al.*, What large language models know and what people think they know, *Nat. Mach. Intell.*, 2025, **7**(2), 221–231.
16 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, *et al.*, The llama 3 herd of models, *arXiv*, 2024, preprint, arXiv:2407.21783, DOI: **10.48550/arXiv.2407.21783**.
17 H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, *et al.*, A comprehensive overview of large language models, *arXiv*, 2023, preprint, arXiv:2307.06435, DOI: **10.48550/arXiv.2307.06435**.
18 Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, *et al.*, A survey on evaluation of large language models, *ACM Trans. Intell. Syst. Technol.*, 2024, **15**(3), 1–45.
19 J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu and R. McHardy, Challenges and applications of large language models, *arXiv*, 2023, preprint, arXiv:2307.10169, DOI: **10.48550/arXiv.2307.10169**.
20 J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, *et al.*, Ai alignment: A comprehensive survey, *arXiv*, 2023, preprint, arXiv:2310.19852, DOI: **10.48550/arXiv.2310.19852**.
21 L. Bereska and E. Gavves, Mechanistic Interpretability for AI Safety–A Review, *arXiv*, 2024, preprint, arXiv:2404.14082, DOI: **10.48550/arXiv.2404.14082**.
22 C. Singh, J. P. Inala, M. Galley, R. Caruana and J. Gao, Rethinking interpretability in the era of large language models, *arXiv*, 2024, preprint, arXiv:2402.01761, DOI: **10.48550/arXiv.2402.01761**.
23 G. Dar, M. Geva, A. Gupta and J. Berant, Analyzing transformers in embedding space, *arXiv*, 2022, preprint, arXiv:2209.02535, DOI: **10.48550/arXiv.2209.02535**.
24 S. Arora, Y. Li, Y. Liang, T. Ma and A. Risteski, Linear algebraic structure of word senses, with applications to polysemy, *Trans. Assoc. Comput. Linguist.*, 2018, **6**, 483–495.
25 A. Scherlis, K. Sachan, A. S. Jermyn, J. Benton and B. Shlegeris, Polysemanticity and capacity in neural networks, *arXiv*, 2022, preprint, arXiv:2210.01892, DOI: **10.48550/arXiv.2210.01892**.
26 A. S. Jermyn, N. Schiefer and E. Hubinger, Engineering monosemanticity in toy models, *arXiv*, 2022, preprint, arXiv:2211.09169, DOI: **10.48550/arXiv.2211.09169**.
27 N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, *et al.*, Toy models of superposition, *arXiv*, 2022, preprint, arXiv:2209.10652, DOI: **10.48550/arXiv.2209.10652**.
28 W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii and D. Bertsimas, Finding neurons in a haystack: Case studies with sparse probing, *arXiv*, 2023, preprint, arXiv:2305.01610, DOI: **10.48550/arXiv.2305.01610**.
29 N. Nanda, *et al.*, Transformer Circuit Faithfulness Metrics Are Not Robust, *arXiv*, 2023, preprint, arXiv:2407.08734, DOI: **10.48550/arXiv.2407.08734**.
30 G. Alain, Understanding intermediate layers using linear classifier probes, *arXiv*, 2016, preprint, arXiv:1610.01644, DOI: **10.48550/arXiv.1610.01644**.
31 T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, in *Advances in Neural Information Processing Systems*, vol. 26, 2013.
32 J. Pennington, R. Socher and C. D. Manning, GloVe: Global Vectors for Word Representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
33 W. Gurnee and M. Tegmark, Language models represent space and time, *arXiv*, 2023, preprint, arXiv:2310.02207, DOI: **10.48550/arXiv.2310.02207**.
34 C. Tigges, O. J. Hollinsworth, A. Geiger and N. Nanda, Linear representations of sentiment in large language models, *arXiv*, 2023, preprint, arXiv:2310.15154, DOI: **10.48550/arXiv.2310.15154**.
35 R. Hendel, M. Geva and A. Globerson, In-context learning creates task vectors, *arXiv*, 2023, preprint, arXiv:2310.15916, DOI: **10.48550/arXiv.2310.15916**.
36 E. Hernandez, A. S. Sharma, T. Haklay, K. Meng, M. Wattenberg, J. Andreas, *et al.*, Linearity of relation decoding in transformer language models, *arXiv*, 2023, preprint, arXiv:2308.09124, DOI: **10.48550/arXiv.2308.09124**.
37 N. Nanda, L. Chan, T. Lieberum, J. Smith and J. Steinhardt, Progress measures for grokking via mechanistic interpretability, *arXiv*, 2023, preprint, arXiv:2301.05217, DOI: **10.48550/arXiv.2301.05217**.
38 Z. Zhong, Z. Liu, M. Tegmark and J. Andreas, The clock and the pizza: Two stories in mechanistic explanation of neural networks, *Adv. Neural Inf. Process. Syst.*, 2024, **36**, 27223–27250.
39 J. Engels, E. J. Michaud, I. Liao, W. Gurnee and M. Tegmark, Not all language model features are linear, *arXiv*, 2024, preprint, arXiv:2405.14860, DOI: **10.48550/arXiv.2405.14860**.
40 O. Skean, M. R. Arefin, D. Zhao, N. Patel, J. Naghiyev, Y. LeCun, *et al.*, Layer by Layer: Uncovering Hidden Representations in Language Models, *arXiv*, 2025, preprint, arXiv:2502.02013, DOI: **10.48550/arXiv.2502.02013**.
41 P. Kavehzadeh, M. Valipour, M. Tahaei, A. Ghodsi, B. Chen and M. Rezagholizadeh, Sorted LLaMA: Unlocking the Potential of Intermediate Layers of Large Language Models for Dynamic Inference, in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 2129–2145.
42 T. Ju, W. Sun, W. Du, X. Yuan, Z. Ren and G. Liu, How large language models encode context knowledge? a layer-wise probing study, *arXiv*, 2024, preprint, arXiv:2402.16061, DOI: **10.48550/arXiv.2402.16061**.
43 Z. Liu, C. Kong, Y. Liu and M. Sun, Fantastic Semantics and Where to Find Them: Investigating Which Layers of Generative LLMs Reflect Lexical Semantics, *arXiv*, 2024, preprint, arXiv:2403.01509, DOI: **10.48550/arXiv.2403.01509**.
44 Y. Zhang, Y. Dong and K. Kawaguchi, Investigating Layer Importance in Large Language Models, in *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, ed. Y. Belinkov, N. Kim, J.

Jumelet, H. Mohebbi, A. Mueller and H. Chen, Association for Computational Linguistics, Miami, Florida, US, 2024, pp. 469–479, **https://aclanthology.org/2024.blackboxnlp-1.29/**.

45 D. Doimo, A. Serra, A. Ansuini and A. Cazzaniga, The representation landscape of few-shot learning and fine-tuning in large language models, *arXiv*, 2024, preprint, arXiv:2409.03662, DOI: **10.48550/arXiv.2409.03662**.

46 M. Yin, C. Wu, Y. Wang, H. Wang, W. Guo, Y. Wang, *et al.*, Entropy law: The story behind data compression and llm performance, *arXiv*, 2024, preprint, arXiv:2407.06645, DOI: **10.48550/arXiv.2407.06645**.

47 K. Meng, D. Bau, A. Andonian and Y. Belinkov, Locating and editing factual associations in gpt, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 17359–17372.

48 M. Geva, J. Bastings, K. Filippova and A. Globerson, Dissecting recall of factual associations in auto-regressive language models, *arXiv*, 2023, preprint, arXiv:2304.14767, DOI: **10.48550/arXiv.2304.14767**.

49 N. Nanda, S. Rajamanoharan, J. Kramár and R. Shah, Fact Finding: Attempting to Reverse-Engineer Factual Recall on the Neuron Level, *Alignment Forum post*, 2023, accessed April 29, 2025, **https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall**.

50 H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, *et al.*, Llama 2: Open foundation and fine-tuned chat models, *arXiv*, 2023, preprint, arXiv:2307.09288, DOI: **10.48550/arXiv.2307.09288**.

51 M. Rodríguez Peña and J. Á. García Guerra, The periodic spiral of elements, *Found. Chem.*, 2024, **26**(2), 315–321.

52 G. B. Kauffman, ElemenTree: A 3-D Periodic Table. By Fernando Dufour, *Chem. Educat.*, 1999, **4**, 121–122.

53 S. Kantamneni and M. Tegmark, Language Models Use Trigonometry to Do Addition, *arXiv*, 2025, preprint, arXiv:2502.00873, DOI: **10.48550/arXiv.2502.00873**.

54 nostalgebraist. Interpreting GPT: The Logit Lens; 2020. AI Alignment Forum/LessWrong blog post. Available from: **https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens**.

55 N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, *et al.*, Eliciting latent predictions from transformers with the tuned lens, *arXiv*, 2023, preprint, arXiv:2303.08112, DOI: **10.48550/arXiv.2303.08112**.