

Uncovering Ontology Knowledge in Large Language Models

Xiaotong Qin^{1,2}, Tong Zhou^{1,2}, Yubo Chen^{1,2*}, Kang Liu^{1,2}, Jun Zhao^{1,2}

¹The Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China

Keywords: Ontological knowledge; Large language models; Sparse autoencoders; Model interpretability; Knowledge features

Citation: Qin X.T., Zhou T., Chen Y.B., et al.: Uncovering Ontology Knowledge in Large Language Models. Data Intelligence, 2026, Vol. 8, Art. No.: 2026r08. DOI: <https://doi.org/10.3724/2096-7004.di.2025.0079>

ABSTRACT

Ontological knowledge, which structurally represents concepts and their interrelations, is crucial for artificial intelligence to effectively mimic human cognition and reasoning. Despite prior work indicating that Large Language Models (LLMs) encode substantial ontological knowledge, the mechanisms through which these models learn and utilize such knowledge remain poorly understood. This gap in knowledge representation and interpretability is critical, as it hampers our ability to enhance and control these models. Sparse Autoencoders (SAEs) have emerged as a promising approach to demystify the internal representations within LLMs, offering new avenues for model interpretability. In this work, we harness SAEs to decompose and interpret the complex activation patterns within LLMs, focusing specifically on the encoding of ontological knowledge. Through a series of systematic experiments, we identify and manipulate key features related to ontological relationships, demonstrating the models' sensitivity to interventions on these features. Our findings not only shed light on the black-box nature of LLMs but also pave the way for developing models with enhanced reasoning capabilities and human-like interactions. This research contributes to the broader endeavor of creating AI systems that can effectively process and reason with structured knowledge, thus expanding their applicability in diverse cognitive tasks.

*Corresponding author: Yubo Chen (E-mail: yubo.chen@nlpria.ac.cn).

1. INTRODUCTION

Ontology is an abstract conceptual framework that serves as a foundation for organizing and representing knowledge to reflect the complexities of the real world. It defines the fundamental classes, properties, and their relationships, providing a structure for understanding and categorizing the various aspects of human experience [1]. Knowledge of ontology is not only vital for human cognition but also underpins effective communication and information sharing [2-5].

Unlike factual knowledge, ontological knowledge is more abstract and structured. Editing ontological knowledge [6-7] is more challenging than editing factual knowledge, as modifying concepts and properties proves to be difficult. This suggests that the learning and modeling approaches for ontological and factual knowledge are different. The abstract nature of ontological knowledge may require distinct computational frameworks for effective learning and representation.

Through extensive pre-training on diverse datasets, LLMs have demonstrated impressive abilities to encode and utilize vast amounts of knowledge, including ontological knowledge [8-10]. It is essential for LLMs to master ontological knowledge because it allows these models not only to store information but also to construct relationships and structures related to various concepts in the world. By mastering ontological knowledge, models can engage in thoughtful and well-reasoned decision-making.

However, the mechanisms that enable models to learn and internalize knowledge, particularly ontological knowledge, are not yet fully understood. A major challenge in this field is how to interpret the internal representations of models in a way that aligns with specific semantics or conceptual terms that are understandable to humans.

Recent advances in SAEs have offered new avenues for decomposing model activations into interpretable features. Building upon the linear representation hypothesis [11], SAEs enable the extraction of sparse feature dictionaries that capture semantic patterns in LLMs [12]. However, existing studies primarily focus on interpreting individual features or linguistic/symbolic representations, leaving the encoding of ontological relationships largely unexplored. Recent studies have shown that SAEs enable the extraction of sparse feature dictionaries that capture semantic patterns in LLMs.

In this work, we aim to interpret and manipulate the features in high-dimensional spaces from an ontological perspective. While previous studies have often interpreted each dimension of the learned dictionary as indicative of a specific semantic or linguistic feature, our focus is different. We will identify a set of features that are specifically related to relationships in ontological knowledge, observe their characteristics, and manipulate them to determine their impact on the model's specific expressions.

In this paper, we propose a framework for interpreting ontological knowledge in large models based on the feature dictionary theory. We utilize a sparse autoencoder to decompose features from model activations, enabling the capture and manipulation of ontological features. We have discovered that ontological knowledge can be represented by the multi-dimensional combinations of specific sets of features. In detail, (1) we effectively capture multi-dimensional ontological knowledge features through

a data-driven approach within the SAEs which has a sparse feature dictionary, identifying feature combinations related to specific ontological knowledge; **(2)** by leveraging the captured multi-dimensional ontological features, we conduct interventions such as feature set replacement, which allow us to manipulate the ontological concepts within the model and verify the existence of ontological features; **(3)** we perform a series of experimental analyses, demonstrating that a model's mastery of ontological knowledge depends on the concentration of multi-dimensional features (or the sharpness of the feature vector), and successful ontological knowledge intervention relies on the quality of the ontological knowledge modeling.

Our main contributions can be summarized as follows:

1. Interpretation of Ontological Knowledge in LLMs: We introduce a novel framework that leverages SAEs to decompose and analyze the internal representations of ontological knowledge in LLMs. This framework provides new insights into how structured and abstract knowledge is encoded and utilized by these models.

2. Feature-Based Ontological Interventions: We demonstrate the feasibility of manipulating specific ontological features within LLMs through feature interventions, such as feature set replacement and superposition. This enables us to alter the model's predictions by targeting key features that represent ontological relationships.

3. Multidimensional Representation of Ontological Knowledge: We identify that ontological knowledge is encoded in a multi-dimensional manner within LLMs. Our experiments reveal that a small set of features can accurately represent complex ontological concepts, allowing for precise control over the model's output.

4. Comprehensive Experimental Validation: Through extensive experiments, we validate the effectiveness of our framework in capturing and intervening in the ontological representations of LLMs. Our results underscore the sensitivity of LLMs to structured knowledge and their capacity to reflect the hierarchical nature of ontological relationships.

Our research is driven by the urgent need to understand how LLMs handle ontological knowledge. The opacity of LLMs hinders model interpretability and control. Our study aims to decode these mechanisms, paving the way for more interpretable, reliable, and controllable AI systems that better align with human understanding of structured knowledge. By revealing how ontological knowledge is encoded and can be manipulated through multi-dimensional feature superposition, our findings provide a new perspective to understand the internal mechanisms of LLMs better, enhancing the interpretability of LLMs.

2. RELATED WORKS

Ontological Knowledge in Large Models. Previous works have largely concentrated on the knowledge encoded within large language models. Studies [8, 13] have demonstrated that LLMs can effectively retain and utilize information from their extensive pre-training on diverse textual corpora. Their findings suggest

that language models can store and retrieve a vast amount of factual and commonsense knowledge without fine-tuning. However, these approaches treat knowledge as isolated facts rather than structured relationships. Recent research has begun to explore how LLMs encode ontological knowledge. [14-16] believe that LLMs are capable of memorizing some ontological knowledge. However, there are disparities in the extent to which they master knowledge from different domains, and the models encounter difficulties in applying ontological knowledge.

Despite these advancements, there is still a significant gap in understanding how these models process and thus comprehend various types of knowledge, particularly ontological knowledge [17, 18]. This line of research is essential, as it lays the groundwork for developing models that can reason with structured knowledge, similar to human cognitive processes.

Interpretability of LLMs and the SAEs. The quest for interpretability in neural networks, especially within the context of language models, has been a challenging endeavor. While these models have shown remarkable performance across various tasks, their internal mechanisms often remain opaque [19]. [20-21] attempt to explain the behavior of individual neurons or small groups of neurons in the model. Although these efforts have had some success, a significant challenge is that neurons often exhibit polysemy [11], activating multiple unrelated types of features. To tackle this issue, researchers have turned to unsupervised learning techniques, such as training a sparse autoencoder, to decompose the complex activation patterns within language models into more interpretable features [22-23]. One crucial hypothesis in this field is that each activation in neuron networks can be represented as a sparse linear combination of feature vectors, known as the linear representation hypothesis [24-25]. Recent studies have shown that large language models represent linguistic, symbolic, and semantic concepts in higher-dimensional spaces. This includes representations of the Japanese language, special symbols, and even the Golden Gate Bridge [12]. In contrast, [26] argues for multi-dimensional representations, suggesting that the model's encoding is not composed of linearly separable features but rather a complex interplay of various dimensions. Anyway, by enforcing sparsity, SAEs have been successful in identifying a dictionary of features that are not only mono semantics but also have predictable effects on the model outputs demonstrating that LLMs encode a rich structure of abstract concepts that can be extracted and utilized to enhance our understanding and manipulation of the model's behavior.

Previous research has primarily concentrated on summarizing the semantics expressed by features individually, taking into account the context in which each feature is activated. In contrast, our work adopts a reverse-thinking approach. We start with a specific concept and investigate the patterns in which its features are activated.

3. PRELIMINARY

3.1 Ontological Knowledge

Ontological knowledge refers to abstractly understanding and representing the essential nature, existence, and relationships of entities. In computer science and artificial intelligence, ontological

knowledge systematically represents concepts, objects, and their relationships, enabling machines to understand and reason about the real world.

The core idea of ontological knowledge is to structure a domain of knowledge using a set of well-defined concepts and the relationships between them. In an ontology, concepts often represent classes (such as “Human” or “Animal”), while relationships define how these concepts are related to one another (e.g., “A human is a type of animal” or “A cat is a subclass of animals”).

As shown in Figure 1, there are four key elements in Ontological Knowledge:

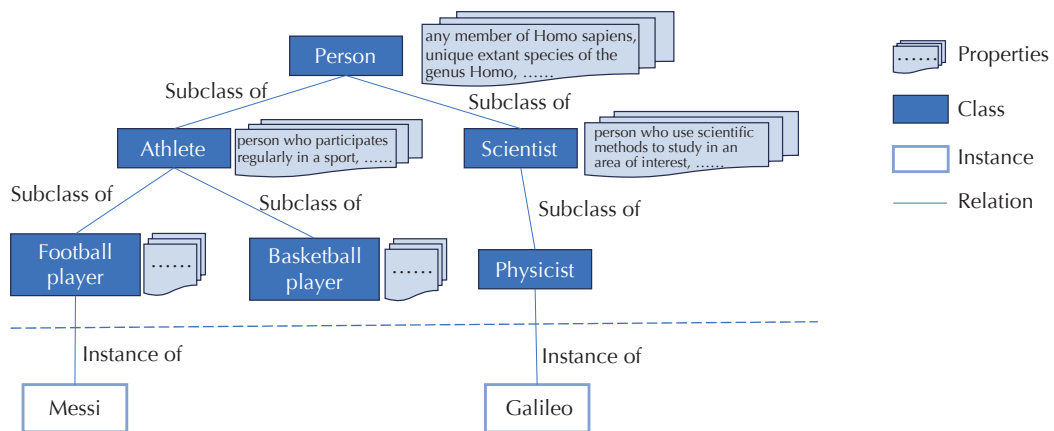


Figure 1. Ontology schema.

Classes: Basic units in the knowledge domain, representing categories of entities. For example, “person” is a class, and “athlete” is a subclass of “person”.

Instances: Specific objects belonging to a particular class. For instance, “Messi” is an instance of the concept “football player”.

Properties: Characteristics of each concept or instance. For example, an “athlete” may have the properties “person who participates regularly in a sport”.

Relationships: Links between different concepts or instances. For example, “A football player is a type of athlete,” or “Athletes and scientists have different careers”.

3.2 Sparse Autoencoder

To figure out how LLMs store and utilize knowledge, we expect to decompose the activations of the models into smaller units. There is a superposition phenomenon in neural networks [11]. In detail, neural networks often use limited quantities of neurons to represent semantic features that far exceed the number

of neurons. This is achieved by assigning features to an overcomplete set of directions in the activation space, rather than to individual neurons. In neural networks, superposition allows efficient encoding of a large number of sparse and distinct features within a limited number of neurons, which is particularly useful for handling polysemous input with many potential interpretable features.

SAEs are used to address the superposition phenomenon, which are designed to learn a sparse representation of the input data. SAEs are unsupervised neural networks that reconstruct activations as weighted sums of m features with non-negative weights, decomposing model activations into sparse linear combinations of features from a large, shared feature dictionary. Specifically, given a location in the model (e.g., an attention head output), suppose that there exists a set of vectors $\{f_i\}_{i=1}^m$ that can sparsely represent different semantics, called dictionary features. Each activation $\mathbf{a} \in \mathbb{R}^n$ at this location can be approximated as a linear combination of f_i with non-negative coefficients. Note that $m \gg n$. SAEs are trained to learn dictionary features. An SAE typically consists of an encoder to decompose activation $\mathbf{a}$ into sparse features, and a decoder to compose the features into reconstructed activation as

$$\mathbf{f} = \text{ReLU}(\mathbf{W}_{\text{enc}}\mathbf{a} + \mathbf{b}_{\text{enc}}) \quad (1)$$

$$\hat{\mathbf{a}} = \mathbf{W}_{\text{dec}}\mathbf{f} + \mathbf{b}_{\text{dec}} \quad (2)$$

Where $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{m \times n}$, $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{n \times m}$, $\mathbf{b}_{\text{enc}} \in \mathbb{R}^m$, and $\mathbf{b}_{\text{dec}} \in \mathbb{R}^n$ are parameters to learn. A key property of SAEs is the sparsity of dictionary features. The sparsity is enforced by adding a penalty term to the loss function that encourages the hidden layer activations to be as inactive (i.e., zero) as possible, except for a few that are necessary to capture the most important aspects of the input. Thus, the loss function is defined as

$$L(\mathbf{a}) = \|\mathbf{a} - \hat{\mathbf{a}}\| + \alpha \|\mathbf{f}\|_1 \quad (3)$$

where α is a hyperparameter controlling the sparsity of the reconstruction.

A well-trained SAE can be used to discover a compact representation of the data by learning sparse and distinct features, enhancing the interpretability of the network's internal representations. This aids in understanding the patterns the network extracts from the data and how the models deduce the output.

4. METHOD

Given an input that describes a piece of ontological knowledge, we will obtain a set of feature vectors using the sparse autoencoder. We view the positive features as the feature set that represents that piece of knowledge. In this section, we explain how to capture ontological feature sets and intervene in these features to manipulate the models. The overall process is as shown in Figure 2.

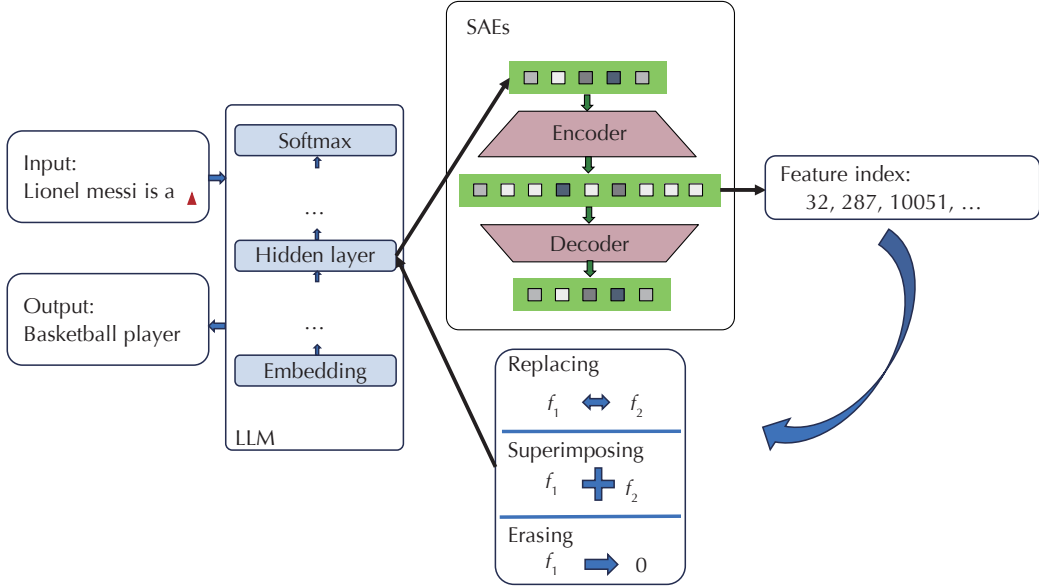


Figure 2. The overall process of our method.

4.1 Ontology Features Capturing

We provide models with an input describing ontological knowledge, requiring the models to predict the tail entity. For example, the sentence “a cat is a subclass of an animal” expresses the relationship between “cat” and “animal.” Thus, the input is formatted as “a cat is a subclass of a(an)”. The models are tasked with predicting the tail entity, while the activation at the residual module before the predicted tokens will be input to the sparse autoencoder, activating specific features in the dictionary. The feature vector is expressed as

$$\mathbf{f} = (f_1, f_2, \dots, f_m) \quad (4)$$

From this process, we obtain the feature vector and consider the positive features as the feature set representing the relationship. The obtained feature set is

$$S = \{i \in \{1, 2, \dots, m\} \mid f_i > 0\} \quad (5)$$

For each class relationship in the schema, there are multiple related inputs, resulting in different sets of features. We take the intersection of these feature sets, assuming it reflects the common relationship shared among the various inputs.

4.2 Ontology Features Intervention

To validate the identified feature sets, we conduct manipulation experiments by modifying the activation values of these feature sets to influence the model's output. The specific manipulation methods include:

(1) Replacing: Swap the feature sets corresponding to different classes (e.g., “athlete” vs. “scientist”) to force the model to replace the activation pattern of the source class with that of the target class, and observe whether the output shifts toward the target class.

(2) Superimposing: Superimpose the activation values of feature sets from different classes with weighted coefficients to observe whether the model can respond to multiple ontological relationships simultaneously or produce semantically mixed outputs.

(3) Erasing: Delete the feature set corresponding to a specific ontological relationship (i.e., set its activation values to zero) to test whether the model fails to reason correctly due to the loss of key features.

5. EXPERIMENT

5.1 Dataset

We construct a dataset for detecting the features that define the semantics of a class in an ontology schema. Considering classes, instances, and the hierarchical relationships in ontology, we have developed a comprehensive ontology schema that includes general classes.

We begin our project by leveraging DBpedia [27] to establish an initial concept schema, identifying a total of 783 classes. To enhance clarity and organization, we manually curate the schema to ensure well-defined and logical hierarchical relationships, ultimately refining it to 370 classes. To further enrich the schema, we gather representative instances for each class. Using SPARQL queries, we extract 20 instances for each class. We also generate inputs that describe the relationships between entities in our schema, such as *Lionel Messi is a football player*, and *Football player belongs to athlete*. By inputting these sentences, we aim to capture the features that represent the relationships between entities in our ontological knowledge. Additionally, we design tasks to generate tail entities using the relationships defined in the ontology for model response analysis.

Besides, for contrastive experiments, we also prepare a dataset portion with sentences containing entities from our ontology schema. These sentences are carefully designed to test the model's ability to distinguish between valid ontological relationships and mere co-occurrence of entities. Specifically, the positive examples are sentences that explicitly express the ontological relationships defined in our schema. For instance, a positive example is “Alfonso Reyes is a famous basketball player”, which clearly states the relationship between the entity “Alfonso Reyes” and the category “basketball player.” In contrast, the negative examples include entities from our ontology schema but do not express any valid ontological relationships. Instead, they present more general or contextual associations between entities. For example, a negative example is “The basketball game attracted many enthusiastic basketball players, and Alfonso

Reyes led the team to win the game.” Although this sentence includes entities like “basketball game,” “basketball players,” and “Alfonso Reyes,” it does not express a direct ontological relationship such as “is a” or “plays for,” which are defined in our ontology. Instead, it describes a situational context where these entities are involved. We manually construct several negative examples and prompt GPT-4 to automatically generate corresponding negative examples for all ontological relationships.

5.2 Experiment Setting

We conduct experiments on the mistral-7b and we use the SAE trained by [28]. The SAE we use is trained on the residual module of the 24th layer in mistral-7b, and it has a dimension of 65536, which is 16 times the model dimension of 4096. The sparsity coefficient used in SAE training is $\alpha = 0.012$, with each token activating an average of $l_0 = 75$ features.

5.3 Results

In this section, we analyze the ontological knowledge features captured by our experiments and the outcomes of intervening in these features from different perspectives. Overall, the model has internalized a certain level of ontological knowledge, which is represented within the SAE through feature vectors. This internalization is evident from the model’s sensitivity to feature interventions, which, in turn, influences the predictive probabilities. The effectiveness of these interventions depends on the activation patterns and feature confidence levels, suggesting that the model’s representation of ontological relationships is both nuanced and manipulable.

5.3.1 Ontology knowledge is represented by feature superposition

In this study, we implement the three intervention methods previously outlined to manipulate the model. As illustrated in Figure 3, replacing the feature activation vectors leads to a decrease in the source knowledge probability and an increase in the target knowledge probability. When the feature activation vectors are superimposed, the probabilities for the source and target knowledge become more similar. Erasing the feature activation vectors leads to a noticeable reduction in the source knowledge probability, with this decrease being more pronounced than that observed during the replacement experiments. This observation may indicate certain conceptual similarities between source knowledge and target knowledge.

These observations allow us to conclude that ontology knowledge can be represented by feature superposition. These findings support the conclusion that ontological knowledge can be effectively represented through the superposition of features. This collection of features captures the hierarchical relationships inherent in ontological knowledge. By intervening in the activation values of these features, we are able to manipulate the model’s output. **This manipulation is indicative of the model’s sensitivity to the underlying ontological structure, thereby affirming the robustness of the feature representation learned by the SAE.**

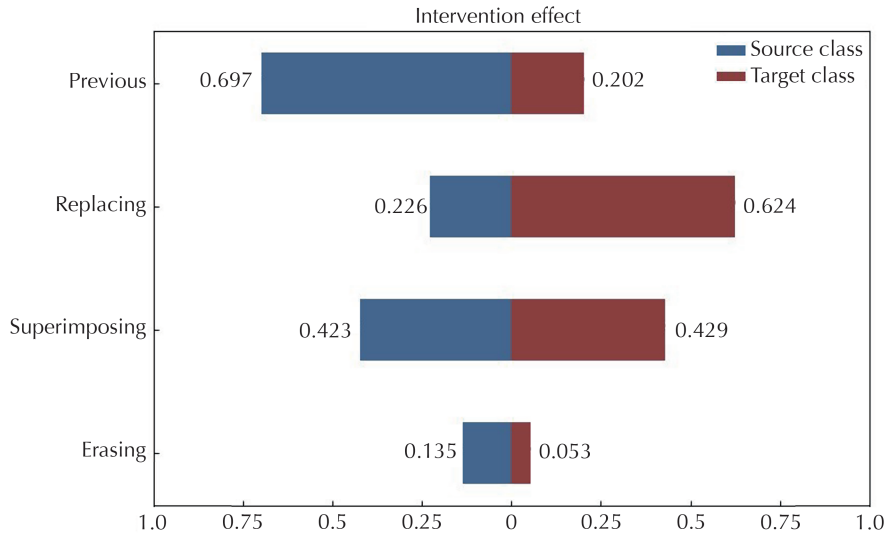


Figure 3. A figure shows the overall result of feature intervention.

The left side displays the distribution of predicted probability values for the original knowledge both before and after intervention. In contrast, the right side shows the distribution of predicted probability values for the target knowledge before and after the intervention.

In the contrastive experiment, we utilize a dataset comprising negative sentences. These sentences incorporate entities from our ontology schema but deliberately omit expressing the specific ontological relationships defined within it. This design serves to test the model’s ability to distinguish between genuine ontological relationships and mere entity co-occurrence. For reference, the features captured from our ontology dataset form the ground feature set against which we compare. We then calculated the average F1 scores for the feature sets extracted from this negative dataset, which are detailed in Table 1. The results indicate a marked distinction between the two types of feature sets. Clearly, there is a significant difference between these two types of feature sets, and **this also illustrates the monosemantic nature of the ontological knowledge feature sets.**

Table 1. F1 scores of negative input data about different ontology classes.

subject type	Object type	F1 score
instance	bottom class	0.04
instance	top class	0.02

5.3.2 Ontology feature is of multi-dimension

We sort the ontology features based on the magnitude of their activation values and select the top k features for intervention experiments—different values of k yield different intervention effects, as shown in

4. We calculate the probability change ratio when intervening with the top k features. Figure 4 illustrates the relationship between the mean overall probability change ratio and k . From the figure, we observe that the intervention effect shows a significant change with the increase of k when k is less than 50, and no significant change is observed when k is greater than 50. Therefore, we posit that **ontology feature is of multi-dimension** and approximately 50 features in the feature dictionary are sufficient to clearly express an ontological relationship.

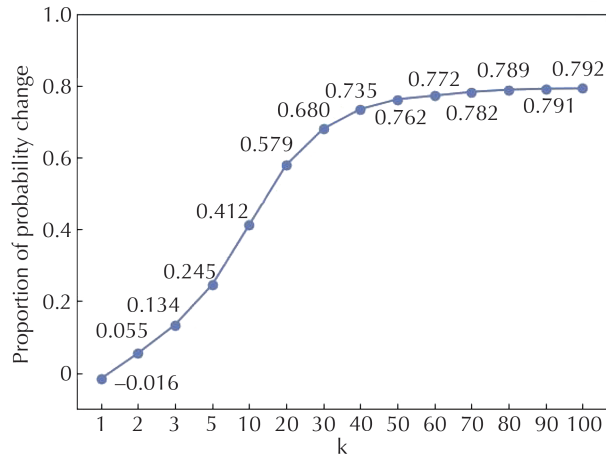


Figure 4. The relationship between the mean overall probability change ratio and k .

horizontal **axis** of the graph represents the number of features to be intervened, while the vertical axis displays the degree of change in probability. This change is calculated by dividing the actual probability change by the target confidence.

5.3.3 Learning (Mastering) Mechanism of Ontology Knowledge

Figure 5 shows the relationship between feature vector similarity and prediction probability differences. When the model receives different inputs, it assigns varying probabilities to the correct answers while activating distinct feature vectors in the feature dictionary. We compare the activation patterns of the SAE for the same class in pairs, calculating the difference in their predicted probabilities and the similarity of their feature vectors. It can be observed from the figure that the closer the probabilities, the more similar the corresponding features vectors. This suggests that when acquiring feature vectors associated with a specific class, greater disparity in the model's understanding of ontological relationships leads to more unstable (or inaccurate) feature vectors. This variability in feature representation can be attributed to the **model's inconsistent grasp of the hierarchical and relational aspects within the ontology, leading to a less reliable representation of the underlying concepts**. It should be noted that among the data sets with small probability differences, there are several outliers. This is because for some inputs, the model fails to generate the correct token, and the generated incorrect tokens can be quite diverse. Consequently, the corresponding feature vectors also differ significantly. In these cases, the predicted probabilities for the correct answers are very low, resulting in lower calculated probability differences.

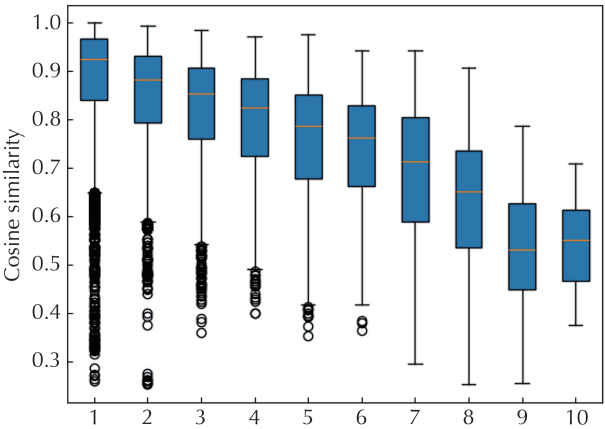


Figure 5. Relationship between feature vector similarity and prediction probability differences. The horizontal axis of the graph is scaled from 0 to 1, divided into ten intervals of 0.1, representing the range of the probability differences. The vertical axis indicates the similarity between feature vectors.

Figure 6 delineates the correlation between feature concentration and the probability of knowledge prediction. Here, we focus on the probability of actual predicted token, which is the maximum likelihood generated by the model, and we posit that the magnitude of this maximum likelihood reflects the model's confidence in its knowledge predictions. Meanwhile, feature concentration, represented by the variance of the feature vectors, indicates the sharpness of the distribution of feature activation values. The graph shows that, generally, as the predicted probability increases, the corresponding feature variance also increases. This suggest that **the sharper and sparser the activated features are, the more explicit the knowledge they represent**, and consequently, the higher the model's "confidence" in its predictions.

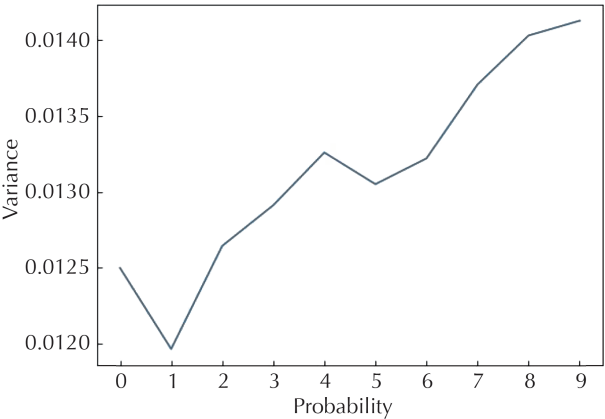


Figure 6. Correlation between feature concentration and the probability of knowledge prediction. The horizontal axis of the graph is scaled from 0 to 1, divided into ten intervals of 0.1, representing the range of predicted probability values. The vertical axis denotes the variance of the feature vectors, where the variance for each point is the mean of a set of variances.

To clarify the relationship between feature similarity and the degree of association in ontology knowledge, we defined an index called “ontological distance” to quantify the distance between entities within our ontology schema. This metric is based on the hierarchical structure of our four-layer ontology:

Distance 1: Entities in the same bottom layer. For example, Lionel Messi and Cristiano Ronaldo are both “football players,” so their ontological distance is 1.

Distance 2: Entities in different bottom layers but the same middle layer. For instance, Messi (a football player) and Kobe Bryant (a basketball player) are both “athletes” but in different sports, giving them a distance of 2.

Distance 3: Entities in different middle layers but the same top layer. For example, Messi (an athlete) and Isaac Newton (a scientist) are both “people” but in different professions, resulting in a distance of 3.

Distance 4: Entities in different top layers. For example, Messi (a person) and Dolly the sheep (an animal) belong to different top layers, so their ontological distance is 4.

Figure 7 plots the ontological distance between two instances on the horizontal axis and the similarity of the ontological features captured when considering these two instances as head entities on the vertical axis. It is evident from the figure that as the ontological distance increases, the similarity of the captured feature vectors decreases. This observation suggests that **the cognitive structure of the model aligns to some extent with the structure of the ontology, indicating that the abstract hierarchical relationships present within the ontology are reflected within the model’s representation.**

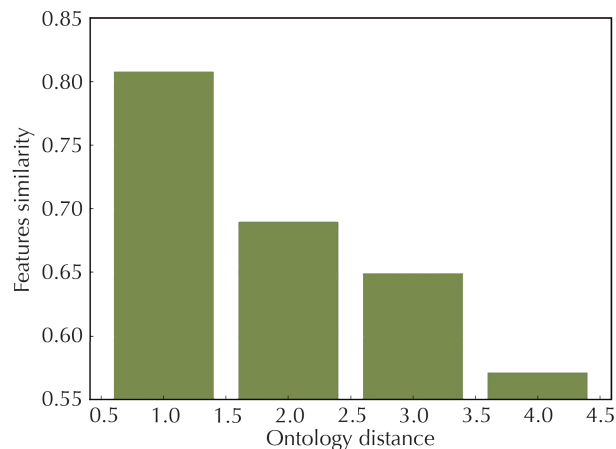


Figure 7. Correlation between ontological distance and feature similarity.

The horizontal axis of the graph represents the ontological distance between two instances and the vertical axis denotes the similarity of the ontology features captured when considering these two instances as head entities.

Furthermore, we perform t-SNE dimensionality reduction on the feature vectors associated with different ontological knowledge and visualize them as a two-dimensional distribution, as shown in Figure 8. The same color represents the feature distribution in the same top class. The larger the dot, the higher the ontological hierarchical layer. From the visualization, it is observable that ontological knowledge entities on the same layer of the ontology schema tend to cluster in the same region. Within each region, entities in the same top class further aggregate, indicating a **discernible pattern that reflects the hierarchical structure of the ontology schema**.

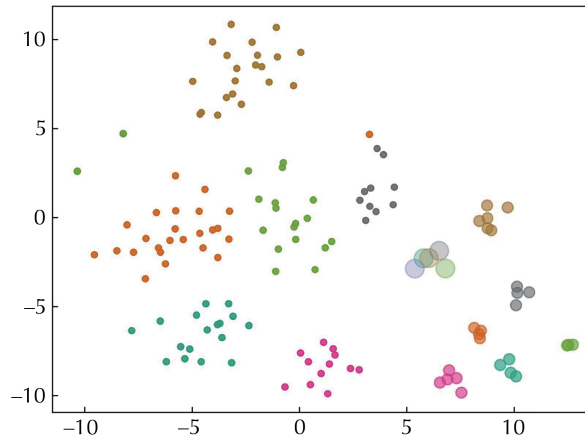


Figure 8. t-SNE visualization of ontological feature vectors.

A dot represents an ontology feature vector related to one class. The same color represents the feature distribution under the same top class. The larger the dot, the higher the ontology hierarchical layer.

These findings are not only a validation of the model's ability to capture the ontological relationships but also provide insights into how LLMs can internalize and process structured knowledge. The correlation between ontology similarity and feature similarity highlights the importance of considering the hierarchical structure of knowledge when designing and training language models. It opens avenues for further research into how different ontological structures might influence model performance and how models can be fine-tuned to better align with specific ontological frameworks. This alignment could potentially lead to more accurate and interpretable language models, enhancing their applicability in knowledge-intensive domains.

5.3.4 Intervention Mechanism of Ontology Knowledge

We define the model's predicted probability for the correct answer as the confidence of the corresponding ontology feature.

Figure 9 illustrates the relationship between the effectiveness of feature intervention and feature confidence. Following the intervention method of replacing, the output probability vector undergoes

certain changes. We calculate the sum of the increases in the target output probability and the decreases in the original correct output as the statistical data for probability change. Generally, the darker colors in the upper right and lighter colors in the lower left of the graph suggest that intervention is most effective when both the source and target confidences are high. From this, we can conclude that **the accuracy of capturing ontological features is related to the predicted probability of the correct answer**: the higher the predicted probability, the more precise the feature capture, and the more accurately the feature set represents the corresponding ontology knowledge.

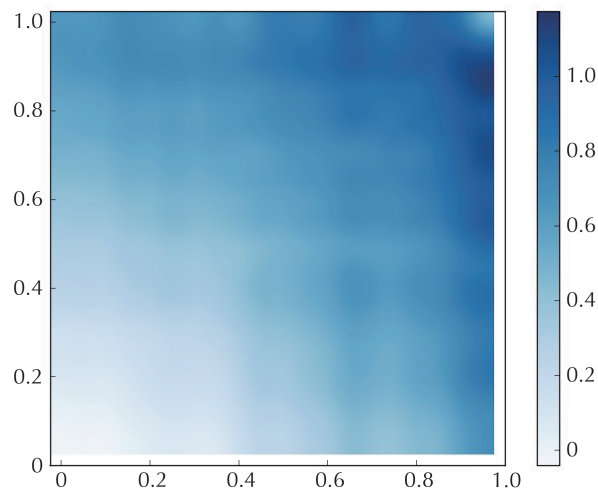


Figure 9. Relationship between the effectiveness of feature intervention and feature confidence.

The horizontal axis of the graph represents the confidence in the original correct output, and the vertical axis represents the confidence in the target output (confidence values are rounded to two decimal places), with the color of the points indicating the magnitude of the probability change.

In the preceding chapter, we learn that ontological features are superpositions of multi-dimensional features, with approximately 50 dimensions of features required to represent a single piece of ontological knowledge. However, the actual situation varies for different specific knowledge bases. We define the effective k number as the k value when the probability change reaches 95% of its maximum. Through experimentation, we found that the effective k value for specific knowledge ranges from 10 to 90. We present the distribution relationship between the intervention probability change and the effective k value in Figure 10. We observe that larger effective k values correspond to smaller probability changes. This suggests that when feature capture is not accurate enough, more features are required to manipulate the model's effect; whereas sufficiently accurate features only need a few features with larger activation values to essentially encompass the relevant information.

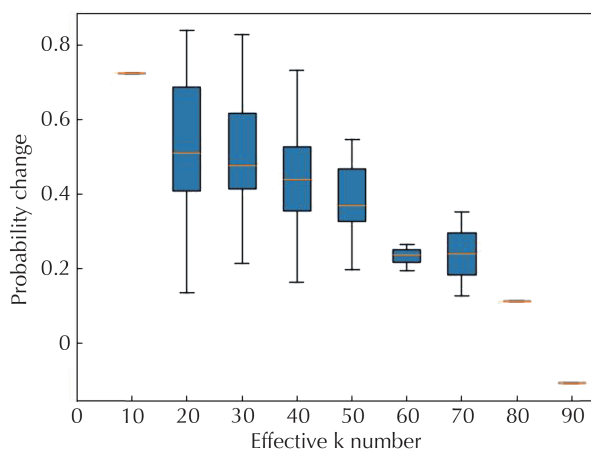


Figure 10. Relationship between the effectiveness of feature intervention and effective k numbers. The horizontal axis of the graph represents the effective k number, and the vertical axis represents the distribution of probability change of target tokens when intervening the features.

6. CONCLUSION AND FUTURE WORK

In this study, we explore how ontology knowledge is represented within large language models (LLMs) through feature-based analysis, leveraging Sparse Autoencoders. Our experiments demonstrate that LLMs are capable of encoding multidimensional representations of ontological knowledge, with specific feature sets capturing hierarchical relationships between entities. By intervening in these features, we were able to manipulate the models' outputs, highlighting their sensitivity to structured knowledge.

Key findings of this research include the identification of ontological knowledge as being encoded through feature superposition, the multidimensional nature of this knowledge, and the significant role of feature concentration in the models' understanding and prediction confidence. Additionally, we showed that effective manipulation of model behavior can be achieved by intervening in these feature sets, confirming the robustness and accuracy of the knowledge representation.

This work contributes to enhancing the interpretability and control of LLMs, particularly in domains requiring deep understanding of structured and abstract knowledge. Future research should aim to generalize these findings across different model architectures and further optimize SAE configurations for improved feature extraction.

While our study provides valuable insights into the representation and manipulation of ontological knowledge within LLMs, it is of many limitations. Firstly, our analysis is primarily based on a single model architecture, which may not be generalizable across different model sizes or families. Future work should explore the variability in ontological knowledge encoding across a broader range of models. Secondly, the SAEs used in this research could be further optimized. We acknowledge the need for a more extensive exploration of SAE configurations, including different training regimes and settings, to enhance the fidelity

of feature extraction and representation. Additionally, our focus has been on the structural aspects of ontology, with less emphasis on the properties associated with the entities. A more comprehensive analysis that includes the properties could reveal deeper insights into how LLMs process and utilize ontological knowledge. Lastly, while our interventions demonstrate the models' sensitivity to ontological features, the long-term stability and consequences of such interventions require further investigation. Expanding the scope to include a wider range of interventions and their long-term impacts will be crucial for developing more robust and controllable LLMs.

AUTHOR CONTRIBUTIONS

Xiaotong Qin: designed experiments, performed experiments, acquired and organized data, created visual charts, summarized conclusions, and drafted the paper.

Tong Zhou: designed experiments, analyzed results, and participated in paper revision.

Yubo Chen: Proposed the topic, conducted preliminary analysis, and participated in topic summary and paper revision.

Kang Liu: Conducted preliminary analysis and participated in topic summary and paper revision.

Jun Zhao: Conducted preliminary analysis and participated in topic summary and paper revision.

ACKNOWLEDGEMENTS AND FUNDINGS

This work is supported by the National Natural Science Foundation of China (No. U24A20335, No. 62176257). This work is sponsored by Beijing Nova Program (No. 20250484750) and supported by Beijing Natural Science Foundation (L243006). This work is also supported by the Youth Innovation Promotion Association CAS.

REFERENCES

- [1] J. F. Nilsson, "Ontological constitutions for classes and properties," in Proc. Int. Conf. Conceptual Struct. (ICCS), 2006, pp. 37–53.
- [2] C. K. Ogden and I. A. Richards, *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. New York, NY, USA: Harcourt, Brace & Co., 1923.
- [3] R. Neches, et al., "Enabling technology for knowledge sharing," *AI Mag.*, vol. 12, no. 3, pp. 36–56, 1991.
- [4] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.
- [5] R. Studer, R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data Knowl. Eng.*, vol. 25, no. 1–2, pp. 161–197, 1998.
- [6] N. Zhang, Y. Yao, and S. Deng, "Knowledge editing for large language models," in Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resources Eval. (LREC-COLING), Tutorial Summaries, 2024, pp. 33–41.

- [7] D. Powell, W. Gerych, and T. Hartvigsen, "TAXI: Evaluating categorical knowledge editing for large language models," in *Findings Assoc. Comput. Linguist. (ACL)*, 2024, pp. 15343–15352.
- [8] F. Petroni, et al., "Language models as knowledge bases?" in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP) and 9th Int. Joint Conf. Nat. Lang. Process. (IJCNLP)*, 2019, pp. 2463–2473.
- [9] Y. Lin, Y. C. Tan, and R. Frank, "Open sesame: Getting inside BERT's linguistic knowledge," in *Proc. ACL Workshop BlackboxNLP: Anal. Interpreting Neural Netw. NLP*, 2019, pp. 241–253.
- [10] C. Wang, X. Liu, and D. Song, "Language models are open knowledge graphs," *arXiv preprint arXiv:2010.11967*, 2020.
- [11] N. Elhage, et al., "Toy models of superposition," *arXiv preprint arXiv:2209.10652*, 2022.
- [12] "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet," *Transformer Circuits*. [Online]. Available: <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>. Accessed: May 21, 2024.
- [13] Z. Jiang, et al., "How can we know what language models know?" *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 423–438, 2019.
- [14] X. Qin, et al., "A comprehensive ontology knowledge evaluation system for large language models," in *China Conf. Knowl. Graph Semantic Comput. (CCKS)*, 2024, pp. 318–328.
- [15] W. Wu, et al., "Do PLMs know and understand ontological knowledge?" in *Proc. 61st Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2023, pp. 3080–3101.
- [16] M. Jullien, M. Valentino, and A. Freitas, "Do transformers encode a foundational ontology? Probing abstract classes in natural language," *arXiv preprint arXiv:2201.10262*, 2022.
- [17] K. Wang, G. Qi, J. Chen, et al., "Embedding Ontologies via Incorporating Extensional and Intensional Knowledge," *Data Intelligence*, vol. 6, no. 4, pp. 1–20, 2024.
- [18] H. Chen, "Large knowledge model: Perspectives and challenges," *Data Intelligence*, vol. 6, no. 3, pp. 587–620, 2024.
- [19] H. Zhao, et al., "Explainability for large language models: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 2, pp. 1–38, 2024.
- [20] "Language models can explain neurons in language models," *OpenAI*. [Online]. Available: <https://openai.com/index/language-models-can-explain-neurons-in-language-models/>. Accessed: Sep. 16, 2024.
- [21] D. Dai, et al., "Knowledge neurons in pretrained transformers," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2022, pp. 8493–8502.
- [22] H. Cunningham, et al., "Sparse autoencoders find highly interpretable features in language models," *arXiv preprint arXiv:2309.08600*, 2023.
- [23] "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits*. [Online]. Available: <https://transformer-circuits.pub/2023/monosemantic-features/index.html>. Accessed: Oct. 5, 2023.
- [24] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. NAACL-HLT*, 2013, pp. 746–751.
- [25] N. Nanda, A. Lee, and M. Wattenberg, "Emergent linear representations in world models of self-supervised sequence models," in *Proc. 6th BlackboxNLP Workshop: Anal. Interpreting Neural Netw. NLP*, 2023, pp. 16–30.
- [26] J. Engels, et al., "Not all language model features are linear," *arXiv preprint arXiv:2405.14860*, 2024.
- [27] S. Auer, et al., "DBpedia: A nucleus for a web of open data," in *Proc. Int. Semantic Web Conf. (ISWC)/ Asian Semantic Web Conf. (ASWC)*, 2007, pp. 722–735.
- [28] "SAELens." [Online]. Available: <https://github.com/jbloomAus/SAELens>.