Distinguishing Concreteness Differences in LLM Representations via Linear Probing

Haodong Xie[1], **Yuze Sha**[2], **Rahul Singh Maharjan**[1], **Federico Tavella**[1] **& Angelo Cangelosi**[1]
[1]Centre for Robotics and AI, The University of Manchester
[2]Centre for Biomedicine, Self and Society, University of Edinburgh

Abstract

Large language models encode rich semantic information, but how concreteness is represented across layers remains unclear. We examine layer-wise linear separability of concreteness by training linear probes on hidden representations from two open-source model families at multiple scales: Qwen3 and Gemma3-Instruct. Using human concreteness ratings, we build balanced prompt datasets with four difficulty levels: an extreme abstract–concrete contrast and three finer boundary comparisons at the abstract end, mid-range, and concrete end. Probes achieve high accuracy on the extreme contrast in shallow layers, showing that endpoint differences are strongly linearly separable. For finer distinctions, performance follows a stable hierarchy: mid-range concreteness is easiest to separate, abstract-end distinctions are hardest, and concrete-end distinctions are intermediate. Across models, accuracy rises rapidly in early layers, peaks in middle layers, and declines in later layers. Together, these findings clarify how the linear accessibility of concreteness varies across LLM layers.

Keywords: Large Language Model; Abstract Concept; Linear Probing

Introduction

Abstract concepts such as "justice" and "love" are harder to understand than concrete ones because they cannot be easily tied to specific, bounded, and identifiable referents. In contrast, concrete concepts like "Border Collie" typically have clear, perceptible referents that can be experienced through the senses (Borghi et al., 2017). For example, love" is more difficult to comprehend than "Border Collie", because a "Border Collie" can be seen, touched, or heard by humans, whereas people cannot directly interact with a physical entity that represents the abstract concept of "love". Concepts are the mental representations of categories, allowing humans to organize and make sense of the world (Barsalou, 2003). From an embodied cognition perspective, these concepts are grounded in the specific body that organisms possess, drawing from perception, action and emotion systems (Borghi et al., 2017). Consequently, the more a concept is detached from physical entities and associated with internal mental states, the more "abstract" it is considered (Barsalou, 2003).

A long-standing issue in psycholinguistic research concerns the selection and classification of abstract and concrete terms. Researchers typically rely on rating-based measures to make this distinction. Two dominant theories explain how this distinction manifests in processing, approaching it from context availability and imageability. Contextual Availability Theory (CAT) (Schwanenflugel et al., 1988) suggests that concrete words are more strongly associated with fewer contexts, making their information more easily accessible when presented as words in isolation. From the imageability angle, Dual Coding Theory (Paivio, 1990) proposes that concrete concepts are encoded in both verbal and imagery systems, whereas abstract concepts lack the "imageability" required for direct mental representation. While a different view proposed by Wiemer-Hastings et al. (2001) proposed that there is no dichotomy between the concrete concepts and abstract concepts; instead them can be seen as a continuum from highly abstract concepts to highly concrete concepts. In this way, rather than a strict dichotomy, many researchers suggest a continuum between abstract and concrete concepts (Borghi, Binkofski, et al., 2014). The differences in between are demonstrated through various perspectives, but what is commonly recognised can be sketched through three main characteristics: grounding (i.e., abstract concepts being less grounded in single, concrete objects), complexity (i.e., abstract concepts being more complicated), and meaning variability (i.e., abstract concepts having more varied meanings).

Large language models (LLMs) learn and transform information across layers, but it remains unclear how abstract and concrete concepts are represented at different depths. Through layer-wise probing across multiple model families, Jin et al. (2025) showed that LLMs process simpler concepts in shallow layers while requiring deeper layers to accurately encode and understand concepts of increasing complexity, a pattern which exists across different model sizes and families. This raises an important question: does the difficulty of distinguishing concepts at different concreteness levels actually vary, and how does this difficulty change across different model families and sizes when distinguishing between abstract and concrete concepts?

Psycholinguistics experiments and neuroimaging evidence

The concreteness effect, i.e., the processing advantage and superior recall of concrete words, is a consistently observed finding in psycholinguistic research across healthy and clinical populations (Paivio, 2013; Strain et al., 1995). Martin and Saffran (1992) observed that patients with cortical lesions would experience greater impairment in processing abstract concepts than concrete ones, while the lesions associated with

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

this deficit do not have a consistent localization. A reverse concreteness effect (Warrington, 1975) (i.e., greater impairment for concrete rather than abstract concept understanding) has been observed in a small number of patients with semantic dementia (Breedin et al., 1994), highlighting the role of perceptual imageability, as damage to visually linked semantic features disproportionately affects concrete concepts. This provides a further support of DCT (Paivio, 1990).

These patterns of patient observations suggest that while abstract and concrete concepts may share a semantic network, they might differ in the relative contribution of specific neural components. This supports a continuum rather than a strict dichotomy between abstract and concrete concepts. Neuroimaging studies confirm this overlap with independent components hypothesis: Binder et al. (2005) found that both word types would activate a left-lateralized multimodal association network. However, concrete words could uniquely engage bilateral regions (e.g., angular gyrus), while abstract words show higher activation in left inferior frontal regions associated with phonological and verbal working memory processes.

Internal Representations in LLMs

Deep neural networks have achieved remarkable performance across a wide range of tasks, neural network training and human learning differ in fundamental ways, and neural networks often fail to generalize as robustly as humans do (Muttenthaler et al., 2025). Deep neural networks continue to function as largely opaque systems, with many aspects of their internal representations and decision processes remaining poorly understood. Recent studies on language models show that internal representations are structured in fine-grained ways: Ju et al. (2024) use a layer-wise analysis of contextual knowledge encoding and demonstrate that LLMs concentrate contextual knowledge in upper layers, lower layers concentrate this information in entity-related tokens while upper layers distribute it across a wider set of tokens, and intermediate layers are susceptible to forgetting previously encoded contextual information when irrelevant evidence is introduced. Sheta et al. (2025) extend interpretability methods to vision–language models with VLM-LENS, reveals systematic differences in how VLMs encode hidden representations across layers and across different target concepts. Zhao et al. (2024) focus on multilingual processing and propose the Multilingual Workflow, showing that LLMs internally shift non-English inputs toward English-like representations for reasoning, supported by their Parallel Language-specific Neuron Detection method for isolating language-dependent neurons. Jin et al. (2025) introduce the Concept Depth, showing through probing experiments that simpler concepts emerge in shallower layers while more complex factual, emotional, and inferential concepts require deeper representations.

Method

In this paper, we use linear probing to examine how large language models represent differences in concept concreteness.

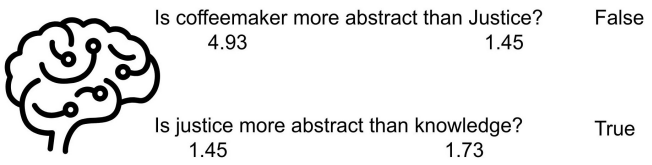

Figure 1: Examples of questions used for linear probing of concept concreteness

We group concepts into tasks based on their concreteness levels, such as comparing highly abstract concepts (concreteness < 1.5) with moderately abstract concepts (1.5–2.0), or moderately concrete concepts (4.0–4.5) with highly concrete concepts (4.5–5.0). Using these ranges, we construct balanced prompt datasets in which each item asks the model to compare two concepts like shown in Figure 1). For each questions, we run the model forward pass and extract the hidden state of the final token at every layer. We then train a separate binary probe on each layer's final-token representations to measure how linearly accessible concreteness distinctions are at that layer.

Linear Classifier Probing

We adopt linear probing (Alain & Bengio, 2016) to evaluate the information encoded in each layer of an LLM. The method works by training a simple supervised model, typically a linear classifier, on the hidden representations taken from a given layer. The performance of this probe reflects how well that layer encodes the relevant information.

We follow the way of prior work on linear probing of LLM (Jin et al., 2025). We extract the hidden representations from every layer, train a binary classifier on these features, and assess the classifier's accuracy. For a task with n samples, let the hidden representations from a given layer be $x \in \mathbb{R}^{n \times d_{\text{model}}}$, where d_{model} is the dimensionality of the layer. Each sample $x^{(i)}$ is associated with a binary label $y^{(i)} \in \{0, 1\}$. We train a logistic regression classifier with L2 regularization by optimizing:

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \text{Cost}(\sigma(\theta^\top x^{(i)}), y^{(i)}) + \frac{\lambda}{2n}\sum_{j=1}^{m} \theta_j^2, \qquad (1)$$

where θ denotes the classifier parameters and λ is the regularization weight. The cost function is defined as:

$$\begin{aligned}\text{Cost}(\sigma(\theta^\top x^{(i)}), y^{(i)}) = y^{(i)} \log \sigma(\theta^\top x^{(i)}) \\ + \left(1 - y^{(i)}\right) \log\left(1 - \sigma(\theta^\top x^{(i)})\right)\end{aligned} \qquad (2)$$

Classification accuracy on a validation set reflects how well the model's internal representations support the distinction between the two classes. In our experimental setting, high accuracy at a given layer indicates that the model can successfully separate concepts based on their differences in concreteness.

Experimental Setting

In this paper, we aim to investigate how LLMs distinguish differences in concept concreteness. To this end, we construct five datasets based on distinct ranges of concreteness ratings. We then perform linear probing on each dataset to examine whether differences in concreteness are linearly separable within each range. This setup allows us to assess how effectively LLM representations capture concreteness at varying levels.

Dataset

We select the concepts used to construct our templated prompts datasets based on their concreteness ratings from the human-rated dataset of Brysbaert et al. (2014). In this work, more than 40,000 concepts were assigned a concreteness score on a 1–5 scale. Concepts with ratings near 1 represent highly abstract terms (e.g., *belief*, mean rating 1.19), while those with ratings near 5 correspond to highly concrete concepts (e.g., *apple*, mean rating 5.00). Concepts falling in the middle of the scale represent moderately concrete items (e.g., *symbol*, mean rating 3.11).

Based on these ratings, we group the concepts into several categories, such as highly abstract and highly concrete concepts. We then pair concepts across different groups to create comparison items. These pairings naturally give rise to tasks of varying difficulty, depending on the concreteness gap between the paired concepts. For example, distinguishing the concreteness difference between a highly abstract concept and a highly concrete one is relatively easy, whereas comparing two concrete concepts where one slightly more concrete than the other poses a more challenging discrimination task. The perceived concreteness of a concept can vary across individuals. In the dataset of Brysbaert et al. (2014), each concept was rated by multiple human annotators, which naturally leads to variation in the assigned scores. To ensure that the concepts in our dataset have stable and reliable concreteness judgments, we select a subset of concepts based on the standard deviation of their ratings. Using the mean standard deviation (1.14836) as a cutoff threshold, we retain only concepts with variability below this value. This results in a final set of 16,293 concepts out of the original 39,954. We construct five templated prompts datasets from this subset with different levels of difficulty:

- (1) Concreteness < 1.5 versus Concreteness > 4.5
- (2) Concreteness 1.0–1.5 versus Concreteness 1.5–2.0
- (3) Concreteness 1.5–2.5 versus Concreteness 2.5–3.5
- (4) Concreteness 4.0–4.5 versus Concreteness 4.5–5.0
- (5) Concreteness < 1.5 versus Concreteness > 4.5 (label abstract vs concrete classification)

We construct two types of datasets. For the first type, each question is phrased as "Is concept A more abstract than concept B?". For every dataset, we sample 500 pairs of concepts from the two specified concreteness ranges and assign one concept as concept A and the other as concept B. Each question is labelled as True or False based on the concreteness ratings of the two concepts. To maintain balance in the dataset, we keep half of the statements true and the other half false. For the second type of dataset, each question is of the form "Is this concept abstract or concrete?". To ensure clear labels, we select concepts only from the ranges < 1.5 (highly abstract) and > 4.5 (highly concrete), since mid-range concepts are more ambiguous with respect to abstractness.

Dataset	Statement	Label
(1)	Is chestnut more abstract than indistinctive?	0
(2)	Is inconclusiveness more abstract than unreasonableness?	1
(3)	Is artfully more abstract than subsection?	1
(4)	Is tennis racket more abstract than player?	0
(5)	Is ham abstract or concrete?	0

Table 1: Example item from the dataset.

Examples from the five datasets are shown in Table 1. For datasets (1)-(4), the "Statement" column contains the comparison question, and the "Label" indicates whether the statement is true or false. For dataset (5), the statement asks whether a single concept is abstract or concrete, label 0 denotes a concrete concept, and label 1 denotes as abstract.

Following prior work (Jin et al., 2025), we use a linear classifier as the probing model in our experiments. We use 80% of the dataset as the training set and the 20% as test set.

Large Language Model

In this paper, we focus on linear probing of LLMs. Because linear probing requires access to intermediate hidden representations, our experiments are limited to open-source models. We use the Qwen 3 models (Q. Team, 2025) and the Gemma 3 models (G. Team et al., 2025). To cover a range of model capacities, we select multiple sizes from each model family: for Qwen 3, we use Qwen3-1.7B, Qwen3-8B, and Qwen3-32B; for Gemma 3, we include Gemma3-4B-Instruct, Gemma3-12B-Instruct, and Gemma3-27B-Instruct.

Experimental Results

In the experiment, we begin by probing the two most straightforward tasks: the comparison between highly abstract and highly concrete concepts, and the abstract/concrete classification task based on the same ranges. The probing results show that even the smaller model (e.g., Qwen3-1.7B) achieves high accuracy in the early layers. For the highly abstract versus highly concrete comparison task, the model reaches a mean accuracy of 98.93%. Similarly, for the abstract/concrete classification task using the same concept ranges, Qwen3-1.7B attains a mean accuracy of 98.86%. These results indicate that the concreteness difference between highly abstract and

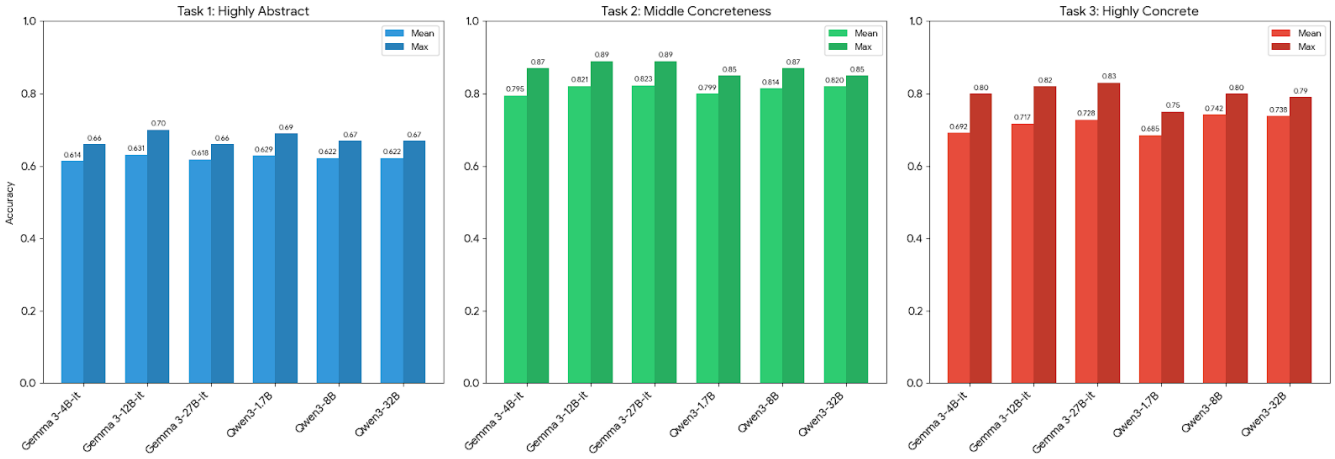

Figure 2: Mean and maximum probing accuracy for the three range-comparison tasks (Task 1: Highly Abstract;Task 2: Middle Concreteness; Task 3: Highly Concrete) across Qwen3 (1.7B/8B/32B) and Gemma3-Instruct (4B/12B/27B)

highly concrete concepts is clearly linearly separable within the model's internal representations. Even at shallow layers, the model distinguishes between these two extremes of concreteness well.

Figure 2 summarizes the mean and maximum accuracies for the three probing tasks across the different concreteness ranges. The results show grade of difficulty across all models. The Middle Concreteness task(Concreteness 1.5–2.5 versus Concreteness 2.5–3.5) is the most distinguishable category, with several models achieving peak accuracies between 0.85 and 0.89. In contrast, Highly Abstract task (Concreteness 1.0–1.5 versus Concreteness 1.5–2.0) is more difficult, with mean accuracies are between 0.614 and 0.631. This suggests that distinguishing between degrees of high-level abstractness is considerably harder than separating mid-abstract concepts from mid-concrete ones. Highly Concrete task(Concreteness 4.0–4.5 versus Concreteness 4.5–5.0) performs better than that of the highly abstract range, however it remains consistently below that of the mid-range comparisons, with mean scores peaking at 0.742.

Figure 3 shows the linear probing accuracy of the six LLMs across different model sizes from two families. Our results indicate that abstract concrete distinguishing is not uniform across the continuum, but reflects different levels of representational complexity. When concepts come from the endpoints of the scale (<1.5 vs >4.5), probes achieve high accuracy performance even in relatively shallow layers, suggesting that high abstractness vs high concreteness is supported by strong, linearly accessible features. In contrast, when the task requires separating nearby levels of abstraction (1.0–1.5 vs 1.5–2.0), performance drops across model families and sizes, implying that fine-grained distinctions among highly abstract concepts are more complex. Mid-range discrimination (1.5–2.5 vs 2.5–3.5) is consistently the most linearly separable among fine-grained tasks, suggesting that the transition region between abstract and concrete concepts contains especially robust separable structure in LLM representations.

In the initial layers, accuracy for all tasks begins low and increases sharply. The steepest rise is observed in the Middle Concreteness task, suggesting that the distinction between abstract and intermediate concepts is among the earliest linearly separable categories to emerge in the model's hierarchy. After reaching a peak in the middle layers, accuracy for nearly all tasks begins to decline. This decline is most pronounced in Highly Concrete task and Highly Abstract. The drop in probe performance possibly as the increasing specialization of deeper layers, where representations are progressively transformed to support the model's output objectives, such as next-token prediction.

Figure 4 shows the F1 scores across layers. The F1 trends closely mirror the accuracy results, confirming that the model's ability to distinguish concepts is most robust in the mid Concreteness task. Across all models, this task consistently achieves the highest F1 scores, indicating strong precision and recall when identifying the boundary between concepts within this range. In contrast, the Highly Abstract task yields the lowest F1 scores, demonstrating that the model struggles to maintain a reliable decision boundary between fine-grained levels of high level of abstract.

Model Scale & LLM Families

In our experiments, tasks that distinguish concepts across different ranges of concreteness exhibit a graded pattern of difficulty. This pattern of task complexity is consistent across different model families and scales. The contrast between highly abstract and highly concrete concepts achieves the highest accuracy and is therefore the easiest task. The second easiest task is the distinction between concepts in the middle range of concreteness. Among the fine-grained tasks, distinctions involving abstract concepts are the most difficult, while those involving concrete concepts are easier. This ordering of task complexity remains consistent across model families

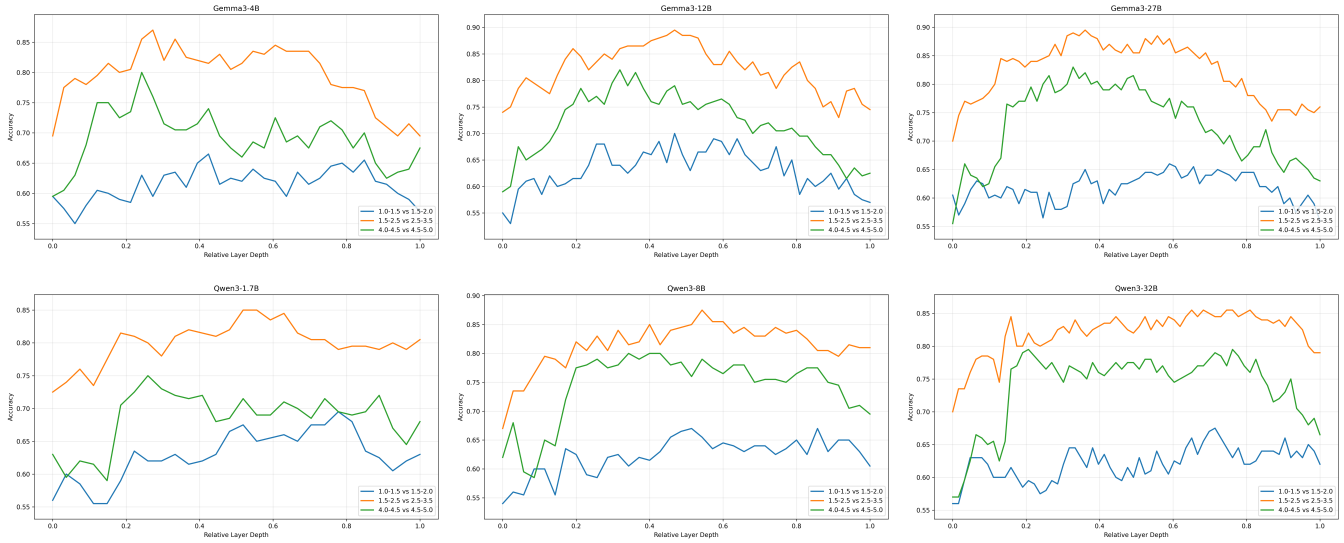

Figure 3: Accuracy curves across layers for all Gemma 3 and Qwen 3 models on the three concreteness ranges. (Task 1: Highly Abstract;Task 2: Middle Concreteness; Task 3: Highly Concrete)

and scales. There is also a shared layer-wise pattern across the tested models. Although there are differences in absolute accuracy, the models show a similar pattern across layers: performance is weaker in the shallower layers, reaches peak accuracy in the middle layers, and then declines in the later layers. This trend appears in both model families. The Gemma 3 models show a larger performance drop in the later layers, whereas the Qwen 3 models exhibit a flatter pattern across layers. This later-layer drop may indicate that the representations become more specialized and therefore less linearly separable.

Overall, the results show that the separability patterns are stable across model families and scales. This suggests that the representation of concreteness pattern consistency across models, rather than being model dependent.

The Abstract–Concrete Continuum

Abstract and concrete concepts are better understood as points on a continuum rather than as a strict dichotomy. In this paper, we examine how concreteness is represented in LLMs. The results of the probing experiments are more consistent with the view that the distinction between abstract and concrete concepts is graded rather than binary. If abstractness and concreteness were represented as a strict dichotomy, we would expect concepts from the two ends of the continuum to be easily linearly separable, while concepts from the middle range of concreteness, where the decision boundary would lie, should be the most difficult to distinguish. At the same time, highly concrete concepts should not be easily distinguished from less concrete ones, and highly abstract concepts should not be easily distinguished from less abstract ones, purely on the basis of concreteness. However, our experimental results show a different pattern than a simple binary account.

In the first experiment, concepts selected from the two ends of the continuum, highly abstract and highly concrete concepts, were highly linearly separable, with peak accuracy reached in relatively shallow layers. This indicates that, in LLM representations, differences in concreteness can already be distinguished in earlier layers. The next easiest distinction was between mid-abstract and concrete concepts. The comparison between concepts with concreteness ranges 1.5–2.5 and 2.5–3.5 yielded better mean accuracy than the comparison between the abstract and concrete ranges, and it was strongly linearly separable, with peak accuracy ranging from 85% to 89%. This indicates that concepts in the middle range of concreteness are more easily distinguished by LLMs than highly abstract concepts are from highly concrete ones. The comparison involving highly concrete concepts was the third most difficult task, with peak accuracy reaching around 80% across model families and scales. This indicates that highly concrete concepts are still linearly separable in LLM representations when compared with less concrete concepts. Comparisons involving highly abstract concepts were the most difficult, but they still achieved peak accuracy between 66% and 70%, indicating that highly abstract concepts can also still be distinguished in the representations.

Taken together, these results show that, in LLM representations, concepts in the middle range of concreteness are highly linearly separable, while highly concrete and highly abstract concepts can still be distinguished from less concrete and less abstract concepts respectively. At the same time, there is a graded difference in task complexity across the concreteness ranges. This suggests that the representation of concreteness is not a strict dichotomy, but is instead more consistent with a graded continuum. Although we do not claim that these results provide direct computational evidence for the psychological assumption of a continuum view, they show a pattern in LLM representations that is consistent with such a graded

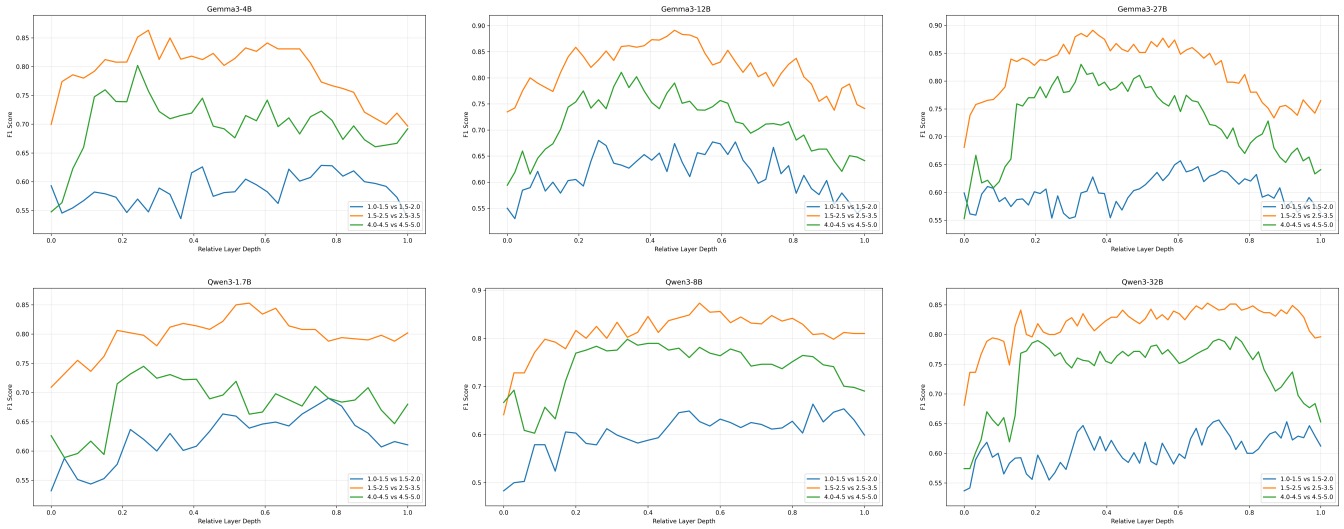

Figure 4: F1 score across layers for all Gemma 3 and Qwen 3 models on the three concreteness ranges. (Task 1: Highly Abstract;Task 2: Middle Concreteness; Task 3: Highly Concrete)

continuum than dichotomy.

The Abstract Concepts Representation

In the experiments, the results also show that distinguishing between abstract concepts is the hardest task among the tested models. This is aligned with psychological theories of abstract learning, as abstract concepts lack clear physical references and are learned differently from concrete concepts. The different modalities involved may affect their representation in the human brain (Borghi, Binkofski, et al., 2014). Again, we do not claim that these results provide direct computational evidence for this psychological assumption. However, the experiments show that the linear separability between abstract concepts is harder than that between concrete ones. This suggests that the task of comparing abstract concepts in terms of concreteness is more complex than that of concrete concepts.

Conclusion

In this paper, we investigate the representation of concreteness inside LLMs. By constructing four datasets with different ranges of concreteness based on human ratings, We investigate the layer-wise linear separability of concreteness by training linear probes on hidden representations from two open-source model families across multiple scales (Qwen3: 1.7B/8B/32B; Gemma3-Instruct: 4B/12B/27B). Using human concreteness ratings for concepts, we construct balanced prompt datasets that vary in difficulty: (i) an extreme-range abstract–concrete comparison (<1.5 vs >4.5) and (ii) three finer-grained boundary comparisons targeting the abstract end (very abstract vs moderately abstract), the mid-range (moderately abstract vs moderately concrete), and the concrete end (moderately concrete vs very concrete). The probing results are consistent across model families and scales. Highly abstract and highly concrete concepts are linearly separable even in the earlier layers of smaller-scale models. The mid-range boundaries are also highly separable, showing the separability of abstract and concrete concepts at the boundary region. For the finer-grained distinctions, distinguishing between concrete concepts reaches peak accuracy of around 80% across model families and scales, while distinguishing between abstract concepts is the hardest among all tasks, with peak accuracy between 66% and 70%. This task complexity suggests that abstract concepts have greater representational complexity. The results also show that the representations of abstract and concrete concepts are more consistent with a graded continuum than with a strict dichotomy. Due to time and computational constraints, we filter concepts only by the standard deviation of human concreteness ratings, which improved the reliability of the concreteness labels but did not explicitly match stimuli on lexical or distributional variables such as word length, token count, frequency, part of speech, or contextual diversity. We added a post hoc lexical confound analysis, it showed that lexical variables were not fully balanced across concreteness ranges. In particular, highly abstract concepts are longer on average than highly concrete concepts, and the mid-range comparison also showed differences in word length and raw character length. For these reasons, future work will conduct control experiments using concepts selected with additional lexical and distributional properties to ensure a fairer measurement of concreteness effects.

Acknowledgment

This paper was partially funded by the ERC Advanced project eTALK (funded by UKRI).

References

Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1435), 1177–1187. https://doi.org/10.1098/rstb.2003.1319

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of cognitive neuroscience*, *17*(6), 905–917.

Borghi, A. M., Binkofski, F., et al. (2014). *Words as social tools: An embodied view on abstract concepts* (Vol. 2). Springer.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263.

Breedin, S. D., Saffran, E. M., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive neuropsychology*, *11*(6), 617–660.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*(3), 904–911.

Jin, M., Yu, Q., Huang, J., Zeng, Q., Wang, Z., Hua, W., Zhao, H., Mei, K., Meng, Y., Ding, K., et al. (2025). Exploring concept depth: How large language models acquire knowledge and concept at different layers? *Proceedings of the 31st international conference on computational linguistics*, 558–573.

Ju, T., Sun, W., Du, W., Yuan, X., Ren, Z., & Liu, G. (2024). How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.

Martin, N., & Saffran, E. M. (1992). A computational account of deep dysphasia: Evidence from a single case study. *Brain and language*, *43*(2), 240–274.

Muttenthaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M. C., Müller, K.-R., Unterthiner, T., & Lampinen, A. K. (2025). Aligning machine and human visual representations across abstraction levels. *Nature*, *647*(8089), 349–355.

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford university press.

Paivio, A. (2013). *Imagery and verbal processes*. Psychology Press.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of memory and language*, *27*(5), 499–520.

Sheta, H., Huang, E. H., Wu, S., Alenabi, I., Hong, J., Lin, R., Ning, R., Wei, D., Yang, J., Zhou, J., et al. (2025). From behavioral performance to internal competence: Interpreting vision-language models with vlm-lens. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 886–895.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1140.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Team, Q. (2025). Qwen3 technical report. https://arxiv.org/abs/2505.09388

Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly journal of experimental psychology*, *27*(4), 635–657.

Wiemer-Hastings, K., Krug, J., & Xu, X. (2001). Imagery, context availabilty, contextual constraint and abstractness. *Proceedings of the annual meeting of the cognitive science society*, *23*(23).

Zhao, Y., Zhang, W., Chen, G., Kawaguchi, K., & Bing, L. (2024). How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, *37*, 15296–15319.