Exploring Concreteness Through a Figurative Lens

Saptarshi Ghosh and Tianyu Jiang
University of Cincinnati
ghosh2si@mail.uc.edu, tianyu.jiang@uc.edu

Abstract

Static concreteness ratings are widely used in NLP, yet a word's concreteness can shift with context, especially in figurative language such as metaphor, where common concrete nouns can take abstract interpretations. While such shifts are evident from context, it remains unclear how LLMs understand concreteness internally. We conduct a layer-wise and geometric analysis of LLM hidden representations across four model families, examining how models distinguish literal vs. figurative uses of the same noun and how concreteness is organized in representation space. We find that LLMs separate literal and figurative usage in early layers, and that mid-to-late layers compress concreteness into a one-dimensional direction that is consistent across models. Finally, we show this geometric structure is practically useful: a single concreteness direction supports efficient figurative-language classification and enables training-free steering of generation toward more literal or more figurative rewrites.

1 Introduction

Concreteness has been widely studied in psychology, linguistics, and natural language processing (NLP). The degree of concreteness reflects the extent to which a word denotes a perceptible physical entity. Highly concrete words denote objects that can be directly experienced through the senses (e.g., *apple*, *chair*, *wood*). In contrast, words referring to intangible or abstract notions are considered less concrete (e.g., *justice*, *idea*, *freedom*). In linguistic studies, the concreteness of a lexical item is closely tied to whether its interpretation is literal or figurative (Lakoff and Johnson, 1980). While figurativity and concreteness are not equivalent, figurative expressions often involve applying a typically concrete word in a non-physical or non-literal sense. Consider the words denoted in bold in the following sentences:

*(a) There is a small **window** of opportunity.*
*(b) The diary opened a **door** into his past.*
*(c) The **church** joined the movement.*

Although *window*, *door*, and *church* each denote tangible entities in their canonical literal uses, the above sentences illustrate idiomatic (a), metaphorical (b), and metonymic (c) usage respectively. The intended objects are not directly tied to their physical meanings and refer to something else in the *figurative context*. Recent work highlighted this connection between concreteness and figurativity, showing that figurative usage is often associated with shifts in a word's concreteness (Chakrabarty et al., 2021; Stowe et al., 2021). Thus, the same lexical item can shift along the concreteness spectrum depending on its contextual interpretation.

In existing NLP work, concreteness is typically treated as a lexical semantic property, most notably through large-scale human annotations such as Brysbaert et al. (2014), which provides static concreteness ratings for words in isolation and have become a foundational reference across the field. Subsequent studies leverage these norms to predict concreteness from word-level embeddings (Charbonnier and Wartena, 2019; Tater et al., 2022; Wartena, 2024), consistently showing that contextualized embeddings reflect concreteness shifts between literal and figurative usage.

However, these studies primarily focus on prediction performance or embedding quality, and do not explicitly examine the internal computation and explainability of contextual concreteness in LLMs. Hence, we still lack a clear understanding of where and how concreteness is processed inside modern LLMs, particularly in relation to figurative text. Mainly, the following questions remain underexplored: (i) *Which layers of an LLM encode and contextualize concreteness?* (ii) *Is the concreteness distinction represented along a dedicated geometric subspace in the model's embedding space?* and

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15471–15490

(iii) *Can this geometric subspace of concreteness be used for steering?*

Addressing these questions, our work presents a systematic investigation by analyzing the internal representation of concreteness in terms of figurative text across layers and model architectures. First, we use a prompt-based method to generate a token carrying information about the contextual concreteness of a term. Using this, we find layer-wise correlation between the token embedding and concreteness scores from Brysbaert et al. (2014). Our results show that concreteness is encoded in the early layers. We also use a synthetic dataset where a term is used in a pair of high and low concrete sense and show that the early layers distinguish the high and low-concrete usage of the word. Second, we analyze the *internal geometric representation* of concreteness across layers. We compute DiffMean vectors (Marks and Tegmark, 2024) of high and low concrete terms and create a global concreteness axis using SVD, which we then use to classify sentences. We identify that *later layers* compress concreteness into a one-directional subspace. We exploit this geometry to classify and generate figurative text by showing that the one-directional subspace representation derived from the concreteness axis rivals a full dimensional supervised classifier in figurative language detection. Third, we use this geometry to steer hidden states and show that the concreteness axis can causally shift the figurative expression of the generated output without parameter updates. We conduct large-scale experiments using 25,000 sentences extracted from Wikipedia and evaluate four major model families—Llama-3.1-8B (Grattafiori et al., 2024), Qwen3-8B (Yang et al., 2025), Gemma2-9B (Gemma Team et al., 2024), and GPT-OSS-20B (OpenAI et al., 2025), ensuring the generalization of our findings. Our code is publicly available.[1] In summary, our findings are:

1. **Layer-wise representation of concreteness.** We find that *early layers* of an LLM encode concreteness of a term and distinguish between its high and low concrete use.

2. **Geometric subspace of concreteness.** We show that concreteness is compressed into a largely *one-directional subspace in later layers*, revealing a shared geometric organization across LLMs.

3. **Geometry-guided figurative language control.** We demonstrate that a one-directional concreteness axis supports both lightweight figurative classification and causal steering of generation, enabling representation-driven control over figurative expression.

[1] https://github.com/cincynlp/concreteness-interpretability

2 Related Work

Concreteness has long been a focus in psycholinguistics (Pollock, 2018). Early studies showed that the concreteness of a word's meaning in context emerges and changes depending on associative richness and imageability (Schwanenflugel and Shoben, 1983; Barsalou, 1999). Psycholinguistic research shows concrete words are processed faster and more accurately than abstract words in a variety of tasks, a phenomenon known as the *concreteness effect* (Jessen et al., 2000; West and Holcomb, 2000; Montefinese et al., 2025). Human-generated resources confirm that concreteness forms a continuous semantic dimension rather than a binary category (Brysbaert et al., 2014). In English, the most widely used resource is the Brysbaert et al. (2014) dataset of 40,000 words, where each word is rated on a scale of 1-5 rated by over 4,000 human annotators, making this dataset the standard for word-level concreteness. Similar concreteness norms have been extended to other languages, albeit in more limited scope, such as French (Bonin et al., 2018), Spanish (Guasch et al., 2016), and Italian (Montefinese et al., 2023; Puccetti et al., 2024). Muraki et al. (2023) collected concreteness judgments for 62k English multiword expressions.

While static concreteness norms are invaluable, they have inherent limitations (Troche et al., 2017), because many words can be concrete or abstract depending on usage (Frassinelli et al., 2017). Generally concreteness of words in isolation highly correlates with that of words in context (Montefinese et al., 2023). But in figurative language, the concreteness of a word significantly alters from its static concreteness (Lakoff and Johnson, 1980; Holyoak and Stamenković, 2018; Lai et al., 2019; Mon et al., 2021). Several works have previously leveraged this property of concreteness to identify figurative language such as metaphors and idioms (Tsvetkov et al., 2013, 2014; Beigman Klebanov et al., 2015; Maudslay et al., 2020). Concreteness has also been extended to multimodal settings, where studies show that it plays a crucial role in shaping how concepts are in figurative de-

pictions such as metaphor (Hessel et al., 2018; Su et al., 2021) and metonymy (Ghosh et al., 2026).

With the advent of powerful LLMs, researchers have studied language models behavior with respect to concreteness. Prior work has shown that contextual concreteness can be predicted using machine learning techniques (Tater et al., 2022). Wartena (2024) showed that dynamic concreteness estimated from BERT-based contextual embeddings had a very high correlation with human judgment, although the study was limited to encoder only models (Devlin et al., 2019; Zhuang et al., 2021). The recent work of Kewenig et al. (2025) exemplifies this trend, showing that textual context, visual features and emotional cues to automatically generate concreteness ratings. Interpretability studies have shown that concreteness can be identified as a direction in the word embedding space (Wartena, 2022). Recent studies demonstrate that the human-to-model alignment is substantially driven by concreteness (Iaia et al., 2025). In spite of these advancements, there exists a lack of explainability as to how current LLMs understand concreteness. We do not exactly know which layers, or the subspace geometry that is responsible for distinguishing if a term is used in a high or low concrete sense in the given context. Our study fills this gap by providing insight into model behavior across different LLMs.

3 Layer-wise Concreteness Representation

In this section, we discuss the datasets and methodology. We then show which layers of the LLM are responsible for encoding concreteness and differentiating high concrete (literal) usage of a word from its low concrete (figurative) sense.

3.1 Datasets

Wikipedia Extraction. In this work, we restrict our analysis to nouns for three reasons: (i) nouns are the primary carriers of referential meaning in figurative text, and concreteness is traditionally defined over nouns; (ii) nouns frequently exhibit systematic sense shifts between concrete and abstract usage (*root, window, door*), making them ideal for studying figurative contextual variation; (iii) nouns in Brysbaert et al. (2014) span a diverse range of concreteness values, enabling controlled evaluation across literal and figurative contexts. The statistics from the Brysbaert et al. (2014) concreteness dataset further support our decision to focus on nouns. As shown in Table 1, nouns exhibit higher variability in concreteness compared to adjectives and verbs. This suggests that nouns span a broader range from concrete to abstract meanings, while verbs and adjectives are more concentrated.

POS	Mean	SD	Count
Noun	3.52	1.01	19,056
Verb	2.92	0.76	5,369
Adjective	2.49	0.51	6,112

Table 1: Statistics of the different parts-of-speech identified from Brysbaert et al. (2014), showing mean concreteness score (Mean), standard deviation (SD), and number of samples (Count).

In total, we identify 19,056 nouns in the database using spaCy. We then extract *25,000 sentences from Wikipedia dumps* containing these target nouns.[2] Extraction proceeds in two stages. First, we collect an initial pool of 50,000 sentences, enforcing a cap of 20 occurrences per noun to maintain lexical diversity. Then, we downsample this pool to 25,000 sentences while preserving a broad distribution of concreteness values across nouns. This ensures that the resulting dataset is both lexically diverse and semantically balanced, providing robust supervision for contextual concreteness probing.

Synthetic Dataset Creation. To enable fine-grained and controlled evaluation, we construct a synthetic contrastive dataset using GPT-5.1. We first sample 600 highly concrete nouns (concreteness score > 4.5) from Brysbaert et al. (2014). For each noun, GPT-5.1 is prompted to generate two different sentences: (i) one where the noun is used in its normal high concrete (literal) sense, and (ii) one where it is used in a low concrete (figurative) sense. This results in 600 sentence pairs, 1,200 sentences total. To verify the generated sentences, two human annotators independently judged whether each sentence used the target noun at the expected level of concreteness. The annotators validated 600 literal sentences and 572 figurative sentences as correct. This portrayed GPT-5.1's strong ability in understanding concreteness and generation. The inter-annotator raw agreement is 93.7%. Full details about the prompts and annotation are given in Appendix H. We randomly sampled 500 pairs from 572, making the synthetic dataset. Table 2 provides

[2]https://huggingface.co/datasets/wikimedia/wikipedia

High Concreteness	Low Concreteness
The **roof** was damaged during the storm and needed repairs	The new policy put a **roof** on employee salaries to prevent overpayment.
The rusty **chain** bound the old gate shut.	The **chain** of events led to the company's downfall.
He climbed the **ladder** to reach the top shelf.	She's been climbing the corporate **ladder** for years.
The **window** on the second floor was broken.	The company's new product launch provided a brief **window** into their future plans.

Table 2: Example of synthetic GPT-5.1 generated sentences.

representative examples from this synthetic corpus. The motivation for using this synthetic dataset is that we want to analyze how LLMs uncover *contextual concreteness* in figurative usage.

3.2 Concreteness Token

The findings of Montefinese et al. (2023) and Wartena (2024) showed that concreteness of words in isolation highly correlates with that of words in context. However, it is well established that in figurative text, the concreteness of a term highly deviate from its static norm (Lakoff and Johnson, 1980). Previous concreteness probing methods typically extract embeddings of the target word in context (Wartena, 2024). However, in decoder-only LLMs, such embeddings are limited by left-to-right processing and may not fully reflect contextual or figurative interpretations (e.g., "*The **chain** of events led to his downfall*"). As a result, we use a prompt-based approach.

Methodology. To recover contextual semantics, we provide the model with the sentence containing the target noun, asking for its concreteness:

```
Sentence: [sentence]
On a scale of 1 to 5 (5 being the
highest), in the context of the sentence,
what is the concreteness of the word
"[target_word]"?
```

We adopt two complementary approaches using this prompt: (i) allowing the model to generate an explicit numerical estimate (Gen), and (ii) extracting the hidden states of the last token representation and train a regression model to predict the concreteness score (Tok). For comparison, we also predict concreteness scores without the context by removing the sentence and providing just the target word to the model. The prompt reads "*On a scale of 1 to 5 (5 being the highest), what is the concreteness of the word [target_word].*" In both cases, we compute and report the Pearson correlation between the predicted score and human-rated concreteness values from Brysbaert et al. (2014).

	With Context		W/o Context	
Model	**Gen**	**Tok**	**Gen**	**Tok**
Llama-3.1-8B	0.66	**0.88**	0.70	**0.98**
Qwen3-8B	0.60	**0.87**	0.65	**0.98**
Gemma2-9B	0.64	**0.92**	0.68	**0.98**
GPT-OSS-20B	0.58	**0.82**	0.63	**0.98**

Table 3: Pearson correlation between the human concreteness ratings and model predictions using two methods: Gen—numeric rating generated by the model; Tok—score derived from last token hidden states.

The hidden representations of the last token encode the contextual interpretation of the noun, enabling concreteness probing in generative architectures where direct noun embeddings are insufficient. We extract the hidden states of the last token across all layers. We train a multi-layer perceptron (MLP) regression model using an 80-20 train-test split and predict the concreteness score of the target noun (More details regarding MLP settings in Appendix G). This setup enables us to study how concreteness is encoded across layers, and how it differs with and without context. Since human concreteness ratings are defined over individual words without context, we expect the without context setting to have a higher correlation.

Results. Table 3 reports the Pearson correlation between human-rated concreteness scores and model predictions obtained from the two methods, with and without the sentence. When directly prompted to output a concreteness rating, the models achieve reasonable correlation with human judgments, while predictions derived from the hidden states of the last token representation show consistently higher correlation, demonstrating that the last-token representation encodes substantial information about the concreteness of the target noun.

Figure 1 presents the layer-wise (normalized across models) correlation between token representations and human concreteness ratings across all

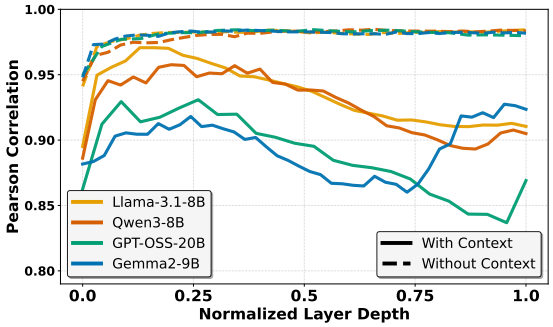


Figure 1: Layer-wise Pearson correlation between static human concreteness ratings and model predictions using the last token representation, with and without sentence context, across four LLMs. Contextual input yields an early peak followed by a later decline, indicating increasing deviation from static norms as deeper layers incorporate contextual interpretation.

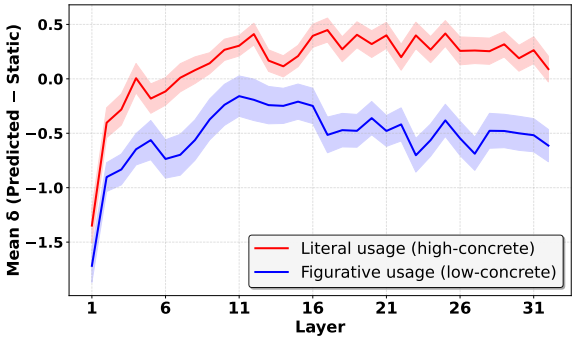


Figure 2: Mean δ across layers in Llama-3.1-8B, where $\delta = C_{\text{pred}} - C_{\text{static}}$. Positive values indicate higher predicted concreteness than the static norm (literal usage), while negative values indicate lower predicted concreteness (figurative usage). The early and consistent separation between high and low concrete curves shows that contextual shifts in concreteness are captured from early layers onward.

models. As we expected, the correlation score for the without context setting is higher than contextual setting. In both settings, correlation is already high in the earliest layers, suggesting that *concreteness is encoded in the very initial layers*. In without context settings, where the target noun is presented in isolation, this high correlation persists across layers. In contrast, when sentence context is provided, the correlation exhibits a distinct pattern: it peaks in the early layers and gradually decreases in later layers. This pattern suggests that early layers capture lexical concreteness, while later layers increasingly incorporate sentence-level context such as figurative interpretations.

3.3 Layers Discriminating Literal & Figurative Usage

A shift in a term's concreteness is often brought by the figurative use of that term (Stowe et al., 2021). If concreteness is encoded in the early layers, does that also mean these layers are responsible for distinguishing literal (high-concrete) and figurative (low-concrete) usage of a term? Here, we answer this question with a controlled experiment on the synthetic dataset which contains high and low concrete usage of the same noun.

Using the 25,000 Wikipedia-sentence corpus, we first extract the hidden states of the last token representation using our prompt, and train an MLP (same settings as the previous experiment) to predict concreteness for each layer-l. We then evaluate the trained regression model on the *synthetic dataset*, which contains pairs where the same noun appears in either a literal (highly concrete) usage (e.g., "the *window* was broken") or figurative (less concrete) usage (e.g., "the *window* of opportunity has passed").

For each layer l, we compute the difference between the predicted contextual concreteness score $C^{(l)}_{\text{pred}}$ and the static human score C_{static} from Brysbaert et al. (2014), and report the mean difference:

$$\delta^{(l)}_{\text{mean}} = \frac{1}{n}\sum_{i=1}^{n}\left(C^{(l)}_{\text{pred},i} - C_{\text{static},i}\right),$$

where n denotes the number of sentences (500 for each literal and figurative). The *magnitude* of $\delta^{(l)}_{\text{mean}}$ (i.e., $|\delta^{(l)}_{\text{mean}}|$) reflects how far the predicted contextual concreteness deviates from the static score. The *sign* indicates the direction of the shift (higher vs. lower than the static norm). We present both $\delta^{\text{high}(l)}_{\text{mean}}$ (literal high-concrete usage, red line) and $\delta^{\text{low}(l)}_{\text{mean}}$ (figurative low-concrete usage, blue line) for every layer to visualize how their separation evolves across the layers.

Results. Figure 2 shows the results for Llama-3.1-8B; results for the remaining models are reported in Appendix D. Across all models, we observe a consistent pattern in which early layers already differentiate literal high-concrete from figurative low-concrete noun usage, with $\delta^{\text{high}(l)}_{\text{mean}}$ and $\delta^{\text{low}(l)}_{\text{mean}}$ separating from as early as layer 2 and remaining distinct through the final layer.

Importantly, the direction of the shift is systematic: literal high-concrete usages yield positive $\delta^{\text{high}(l)}_{\text{mean}}$, while figurative low-concrete usages yield negative $\delta^{\text{low}(l)}_{\text{mean}}$, indicating that the model adjusts

concreteness estimates in the appropriate direction based on context. Appendix A shows the same experiment on verbs yields a lower separation.

Takeaway. LLMs encode concreteness and differentiate literal (high-concrete) and figurative (low-concrete) usage of a noun in the *early layers*.

4 Concreteness Geometric Subspace

In this section, we examine how LLMs internally represent concreteness. We create a global concreteness axis and use it to classify figurative text.

4.1 Creating Global Subspace

We now want to characterize the geometric structure of concreteness representations by deriving a global subspace. For this, we first construct *difference-in-means* (DiffMean) vector (Marks and Tegmark, 2024) that captures the dominant direction along which concreteness varies in the model's hidden space. DiffMean is a lightweight linear method that identifies directions associated with behavioral distinctions, offering mathematical simplicity along with direct interpretability and strong empirical performance (Vennemeyer et al., 2025; Yao et al., 2026).

Constructing DiffMean. We use the 25,000 Wikipedia sentences to create the DiffMean vector. Since DiffMean is the difference between the samples from positive and negative classes, we set two thresholds to gather high and low concrete sentences. Using static concreteness ratings (1–5) from Brysbaert et al. (2014), we classify nouns with scores ≥ 4 as high-concrete ($\mathcal{D}_{\text{high}}$) and scores ≤ 2 as low-concrete ($\mathcal{D}_{\text{low}}$). This yields a balanced sample of 2,256 high-concrete and 2,116 low-concrete sentence instances. Following our prompt-based probing setup, we extract the hidden representation of the last token that encodes concreteness information at every transformer layer l for both high and low concrete sets. For layer l, we compute the mean representation of the last token for the high- and low-concrete groups:

$$\mu_{\text{high}}^{(l)} = \mathbb{E}_{i \sim \mathcal{D}_{\text{high}}} \left[\mathbf{h}_i^{(l)} \right], \quad \mu_{\text{low}}^{(l)} = \mathbb{E}_{i \sim \mathcal{D}_{\text{low}}} \left[\mathbf{h}_i^{(l)} \right],$$

where $\mathbf{h}_i^{(l)}$ denotes the layer-l hidden representation of the last token for sentence instance i. We define the *DiffMean* at layer l as:

$$\mathbf{w}^{(l)} = \mu_{\text{high}}^{(l)} - \mu_{\text{low}}^{(l)}.$$

Global Concreteness Subspace via SVD. We seek to identify a compact set of linear directions that differentiate high and low concrete noun usage across the model's layers. To do this, we first stack all the DiffMean vectors from all the layers, creating a global matrix W. We then perform singular value decomposition (SVD) on this matrix. This yields a set of orthogonal directions in the hidden state space ($V^\top$), ranked by how strongly they separate high and low concrete usages. We select the top-k right singular vectors to form a global basis:

$$B_k = V_{1:k}^\top,$$

where B_k contains the k strongest directions along which high and low concrete usages differ. B_k defines a layer-independent subspace capturing the dominant geometry of concreteness. By selecting k in a single direction ($k = 1$), we test whether concreteness is encoded in a highly compressed, unidirectional space across layers of the LLM.

4.2 Layer-wise Projection & Results

To evaluate the effectiveness of this global subspace, we use our prompt-based probing technique on the synthetic dataset and obtain, for each sentence, the hidden state $h^{(l)}$ at layer l that carries contextual concreteness information. We then project these embeddings into the global subspace:

$$s(h^{(l)}) = B_k h^{(l)},$$

where B_k is the global concreteness subspace and $s(h^{(l)})$ are the k-dimensional projection scores. We then take the average of this scalar score across all dimensions in the model. Using these scores for high and low concrete sentences as positive and negative examples, we compute the area under the ROC curve (AUROC) at each layer, which quantifies how well the global subspace separates concrete usages across model depth.

Results. Figure 3 reports AUROC scores for a unidirectional concreteness subspace ($k = 1$) across all models with normalized layer depth. Individual model graphs and different k values are provided in Appendix E. The results show that *all models* compress concreteness into a **single global direction** in the later layers. The compression begins primarily in the middle layers. GPT-OSS-20B shows slightly different behavior as the compression begins in the early layers—we attribute this deviation due to its mixture-of-expert architecture (Lo et al., 2025). For all models, AUROC score reach around 0.90

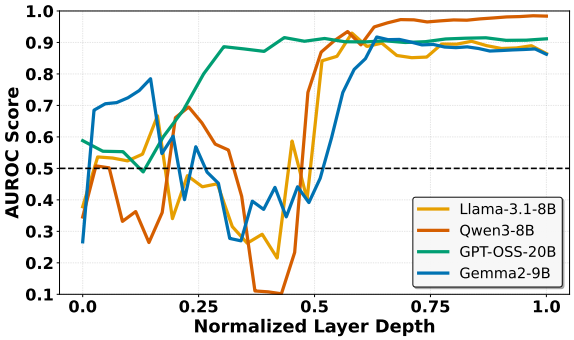


Figure 3: Layer-wise AUROC for separating high and low concrete noun usages using a global unidirectional concreteness axis across four LLMs, showing increasing separability and compression in middle and later layers.

Task	Dataset	AUROC Score	
		Subspace (Zero-shot)	Full Rep. (Trained)
Idioms	MAGPIE	95.2	98.5
	EPIE	95.3	99.2
Metaphor	VUA	95.7	97.6
	MUNCH	93.2	95.1
Metonymy	ConMeC	60.2	62.6
	MetFuse	85.7	96.3

Table 4: AUROC scores for figurative text classification with Llama-3.1-8B, comparing a unidirectional concreteness subspace learned from Wikipedia and applied to downstream datasets against a full-representation classifier trained separately on each dataset.

from the middle layers through to the final layer, indicating that a unidirectional linear subspace is sufficient to distinguish concrete from abstract usage at those stages.

4.3 Figurative Text Classification

Prior work on figurative text classification typically relies on training neural networks using the contextualized embeddings. Motivated by our finding that concreteness is encoded in a compressed one-directional subspace in the LLMs, we explore classifying figurative text using only this unidirectional concreteness axis *without training*, and successfully achieve comparable performance to full-dimensional supervised training.

Datasets. We use three forms of figurative text and two datasets from each: i) Idioms - EPIE (Saxena and Paul, 2020) and MAGPIE (Haagsma et al., 2020), ii) Metaphor - VUA (Steen et al., 2010) and MUNCH (Tong et al., 2024), iii) Metonymy - ConMeC (Ghosh and Jiang, 2025) and MetFuse (Ghosh and Jiang, 2026). We select samples where the figurative expression is conveyed through the noun. The data preparing, preprocessing and distribution is provided in Appendix F.

Methodology. Similar to the previous experiment, we use the 25,000 Wikipedia sentences to construct the DiffMean vector and get the one-directional concreteness subspace ($k = 1$) using SVD. We then evaluate this geometric subspace on different figurative text classification datasets. For each dataset, we select the hidden representation of the target noun from a layer where our analysis showed strong single directional separability (layer 20 for Llama-3.1-8B). We then project the representation onto the learned unidirectional axis to obtain a single scalar projection score and compute the AUROC score. As a comparison, we also train a regression classifier with the same parameters as in Appendix G on the last layer's *full hidden state* (4,096-dimensional for Llama-3.1-8B) using 80-20 train-test split, averaged over 5 runs. This baseline represents the performance without dimensionality constraints. This setup allows us to directly measure how much discriminative signal for concreteness is preserved in the one-directional subspace.

Results. Table 4 reports AUROC scores when using the concreteness subspace projection score to classify, compared to a classifier trained on the full 4,096-dimensional hidden state of Llama-3.1-8B. Results for other models are provided in Appendix J. All four models show consistent results. Across all figurative datasets, our subspace projection consistently recovers more than 95% of the performance of the full classifier, despite it being a scalar projection on a unidirectional axis compared to the its original 4,096 dimensional embeddings. The high AUROC for metaphor and idiom datasets using the concreteness axis aligns with previous linguistic studies which states that concreteness of a term alters highly in metaphors and idioms (Lakoff, 1993; Glucksberg, 2001). In comparison, metonymy keeps the interpretation tied to a concrete entity or situation and the concreteness of a term shifts less (Barcelona, 2003). This results in lower AUROC for metonymy compared to metaphors and idioms. MetFuse gets a relatively better score compared to ConMeC as it solely focuses on location metonymy, narrowing its scope. The gap between the unidirectional concreteness subspace with the fully trained classifier is more,

Original Literal Sentence	Unsteered Rewrite	Figuratively Steered Rewrite
She struggled to push the heavy door open.	She struggled to force the heavy door open.	The heavy door battled her attempts to push it open.
The dirt path wound through the forest, the hikers followed it to a hidden waterfall.	The dirt path wound through the forest, leading the hikers to a hidden waterfall.	A winding dirt path marched the hikers through the forest to a hidden waterfall.
The electoral results were to be announced later that day.	The electoral results were scheduled to be announced later that day.	Suspense engulfed the city for the electoral results later that day.
The factory was shut down in 1972.	The factory closed in 1972.	The factory closed its door in 1972.

Table 5: Example of concreteness-guided steering for rewriting literal sentences with Llama-3.1-8B. The original sentence and all the unsteered rewrites are literal. The steered rewrites are figurative.

tying to previous linguistic work regarding limited concreteness shift in metonymic interpretation of a term (Panther and Radden, 1999). These results indicate that the low-dimensional concreteness subspace can be used for figurative text identification.

Takeaway. LLMs *compress concreteness into a unified one-directional subspace in the late layers*, and this compression *emerges in the middle layers*. This subspace can be used for figurative text classification with significantly low latency and almost equal performance as a fully trained classifier.

5 Concreteness Controlled Steering

The results above establish that LLMs encode literal-figurative variation along a highly compressible one-directional subspace in its deeper layers. We now assess whether this latent direction is merely representational or whether it plays a causal role in generative behavior. If this semantic dimension truly governs how the model interprets a noun in context, then directly manipulating hidden states along this axis should influence the figurativity of the generated text. We turn the concreteness subspace into a controllable knob, enabling post-hoc control over how literal or figurative the model's generations are.

Methodology. We use the previous method to derive the unidirectional concreteness subspace using SVD over hidden representations of 25,000 Wikipedia sentences. This learned direction is subsequently applied to multiple figurative language datasets, enabling us to evaluate the cross-dataset transferability of the learned concreteness representation. This gives a unit direction u along which concreteness varies in the model's latent space. We then provide a sentence to the model and use a prompt to rewrite the input: "*Rewrite the following sentence clearly and naturally:*". As noted, the prompt does not contain any instructions about figurativeness or concreteness of the rewrite. During decoding, we intervene on the model's hidden state $h^{(\ell)}$ at a selected layer ℓ—chosen from layers where the one-directional subspace reliably encodes concreteness information (Figure 3). At this layer, we modify the hidden representation by adding an offset α along the concreteness axis:

$$h^{(\ell)}_{\text{steer}} = h^{(\ell)} + \alpha \cdot u,$$

where α controls the strength and direction of the intervention. For a more *literal* rewrite, we steer in the direction of high concrete representations ($\alpha > 0$), and for a more *figurative* rewrite, we steer in the opposite direction ($\alpha < 0$). The perturbed hidden state is then propagated through subsequent layers to produce the rewritten continuation.

To evaluate steering effectiveness, we use the synthetic dataset. We sample 100 figurative (low concrete) sentences when steering toward literal rewrites, and 100 literal (high concrete) sentences when steering toward figurative rewrites. For each sentence, we generate both steered ($\alpha = \pm 40$) and unsteered rewrites. In both cases, we use the same rewriting prompt without any hint regarding the intended expression of the rewrite. Then two independent human annotators rate whether the output exhibits the intended literal or figurative expression. We intervene at layer 20 for Llama-3.1-8B, 25 for Qwen3-8B, 27 for Gemma2-9B and 15 for GPT-OSS-20B.

Results. Table 5 shows some qualitative examples of unsteered vs steered rewrites going from literal to figurative. Table 6 shows the quantitative result of the steering-based rewriting experiment. We used human annotators to evaluate 100 sentences under both figurative and literal steering conditions, as well as their unsteered counterparts. For literal inputs, all models produce literal paraphrases in the unsteered condition. Steering towards the low-concrete direction ($\alpha = -40$) induces figurative reinterpretations for around 10%

	Lit → Fig		Fig → Lit	
Model	**Steer**	**Unsteer**	**Steer**	**Unsteer**
Llama-3.1-8B	**12**	0	**71**	42
Qwen3-8B	**10**	0	**68**	45
Gemma2-9B	**9**	0	**67**	52
GPT-OSS-20B	**15**	0	**75**	39

Table 6: Number of sentences judged by humans to be successfully steered in a sample of 100. *Lit → Fig* indicates original sentence was literal, the LLM rewrite is figurative. *Fig → Lit* is when input is figurative, rewrite is literal.

of the sentences. For figurative inputs, unsteered rewriting already produces literal paraphrases for 40–50% of the cases, reflecting a strong literal bias in LLMs (Chakrabarty et al., 2022). Our concreteness-based steering consistently improves this figure to around 70%, showing that shifting hidden states along the identified concreteness axis makes the model more likely to rewrite the figurative sentence in a literal sense. Table 8 in Appendix I shows qualitative examples of steering from figurative to literal.

The results reveal that steering from figurative to literal usage is more effective than steering from literal to figurative usage. This aligns with the observation that transforming literal text to figurative is difficult due to semantic and meaning abstraction challenges (Stowe et al., 2021; Chakrabarty et al., 2021). Figurative language requires controlled conceptual mapping while preserving meaning, which makes generation harder than literal instances (Lai and Nissim, 2024).

Nevertheless, the ability to induce even moderate figurative reinterpretation from literal inputs without modifying model parameters or providing a figurativity prompt suggests that the learned concreteness direction is a *causally relevant* control dimension. We believe that this result opens opportunities for future work on controllable figurative generation, including designing steering vectors derived from richer figurativity phenomena (e.g., metonymy vs. metaphor distinctions) and combining causal steering with explicit task prompts to further enhance stylistic expressivity.

Takeaway. The one-directional geometric axis of concreteness could be used as a causally manipulable handle for modifying the figurative-literal interpretation of text, enabling controlled rewriting without model retraining or task-specific prompts.

6 Conclusion

Our work provides a comprehensive analysis of how LLMs represent and process concreteness internally. We show that the early layers encode concreteness, and differentiate high-concrete (literal) and low-concrete (figurative) usage of a term. In later layers, concreteness is compressed into a shared direction in representation space, and the compression emerges from the middle layers. This structure generalizes across models and supports zero-shot figurative language classification with significantly low latency and close to a fully trained classifier. Finally, we demonstrate that manipulating this concreteness direction causally steers generation between literal and figurative interpretations.

Limitations

While our work analyzes the internals of different LLMs in understanding concreteness and shows promising future research direction, certain limitations exist. First, our study does not use human-annotated judgments for contextual concreteness. We use a trained regression model and concreteness information carrying token to predict the concreteness score. The predicted scores show very high correlation with the human annotated scores (≈0.98). However, incorporating human annotations would provide a stronger grounding for evaluating contextual variation and remains an important direction for future work.

We construct a unidirectional concreteness subspace using DiffMean vectors derived from high and low concrete noun usages. This subspace yields competitive performance for figurative language classification and enables controllable rewriting between figurative and literal expressions. However, it may be possible that this subspace does not isolate concreteness in a strictly independent manner. The identified direction may also encode other correlated semantic or linguistic factors, and intervening along this axis does not guarantee that only concreteness-related properties are affected. Identifying overlapping factors in a geometric subspace and disentangling them remains challenging. We view this as an interesting and important direction and leave it for future work.

While concreteness is strongly related to figurative language, the two are not equivalent. A reduction in contextual concreteness does not uniquely imply figurative usage. For instance, word sense

disambiguation phenomena can also involve shifts from highly concrete nouns to less concrete contextual usage without invoking figurative language (*root* of a tree vs. square *root* of a number). To mitigate this confound, we explicitly identify and remove such cases when constructing our synthetic dataset, ensuring that concreteness shifts are primarily driven by figurative usage. Consequently, our experiments evaluate the concreteness axis in a controlled setting where concreteness-based distinctions align with figurative contrasts. However, this does not imply that all figurative language can be reduced to concreteness differences, nor that concreteness alone is sufficient to fully characterize figurative meaning.

Ethical Considerations

The natural language data used in our experiments originates from Wikipedia, which may contain historical and demographic biases; however, our study focuses exclusively on representational differences in word concreteness and does not involve any applications with social impact. Synthetic evaluation examples were generated using a commercial LLM (GPT-5.1), and human annotation was limited to verifying concreteness usage rather than sensitive attributes.

Acknowledgments

We thank the CincyNLP group for their suggestions and feedback. We also thank the anonymous ACL reviewers for their insightful suggestions.

References

Antonio Barcelona. 2003. *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective*. Mouton de Gruyter.

Lawrence W Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

P. Bonin, A. Méot, and A. Bugaiska. 2018. Concreteness norms for 1,659 french words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods*, 50(6):2366–2387.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Cristina Cacciari and Patrizia Tabossi, editors. 1993. *Idioms: Processing, Structure, and Interpretation*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2021)*.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187, Gothenburg, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte m Walde. 2017. Contextual characteristics of concrete and abstract words. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Saptarshi Ghosh and Tianyu Jiang. 2025. ConMeC: A dataset for metonymy resolution with common nouns. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025)*.

Saptarshi Ghosh and Tianyu Jiang. 2026. Metfuse: Figurative fusion between metonymy and metaphor. *Preprint*, arXiv:2604.12919.

Saptarshi Ghosh, Linfeng Liu, and Tianyu Jiang. 2026. A computational approach to visual metonymy. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*.

Sam Glucksberg. 2001. *Understanding Figurative Language: From Metaphor to Idioms*. Oxford University Press, New York, NY, USA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

M. Guasch, P. Ferré, and I. Fraga. 2016. Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, 48(4):1358–1369.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*.

Keith J. Holyoak and Dušan Stamenković. 2018. Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, 144(6):641–671.

Cosimo Iaia, Bhavin Choksi, Emily Wiebers, Gemma Roig, and Christian J. Fiebach. 2025. The representational alignment between humans and language models is implicitly driven by a concreteness effect. *Preprint*, arXiv:2505.15682.

F. Jessen, R. Heun, M. Erb, D.-O. Granath, U. Klose, A. Papassotiropoulos, and W. Grodd. 2000. The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, 74(1):103–112.

Viktor Kewenig, Jeremy I. Skipper, and Gabriella Vigliocco. 2025. A multimodal transformer-based tool for automatic generation of concreteness ratings across languages. *Communications Psychology*, 3.

Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Comput. Surv.*, 56(10).

Vicky T. Lai, Odessa Howerton, and Rutvik H. Desai. 2019. Concrete processing of action metaphors: Evidence from erp. *Brain Research*, 1714:202–209.

George Lakoff. 1993. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*, 2 edition, pages 202–251. Cambridge University Press, Cambridge, UK.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. 2025. A closer look into mixture-of-experts in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025 (Findings of NAACL 2025)*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *Preprint*, arXiv:2310.06824.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226, Online. Association for Computational Linguistics.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *Preprint*, arXiv:2302.07842.

Serena K. Mon, Mira Nencheva, Francesca M.M. Citron, Casey Lew-Williams, and Adele E. Goldberg. 2021. Conventional metaphors elicit greater real-time engagement than literal paraphrases or concrete sentences. *Journal of Memory and Language*, 121:104285.

M. Montefinese, L. Gregori, A. A. Ravelli, R. Varvara, and D. P. Radicioni. 2023. Concretext norms: Concreteness ratings for italian and english words in context. *PLOS ONE*, 18(10):e0293031.

M. Montefinese, A. Visalli, A. Angrilli, and E. Ambrosini. 2025. Fine-grained concreteness effects on word processing and representation across three tasks: An erp study. *Psychophysiology*, 62(5):e70074.

Emiko J Muraki, Summer Abdalla, Marc Brysbaert, and Penny M Pexman. 2023. Concreteness ratings for 62,000 english multiword expressions. *Behavior research methods*, 55(5):2522–2531.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, and others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Klaus-Uwe Panther and Günter Radden, editors. 1999. *Metonymy in Language and Thought*, volume 4 of *Human Cognitive Processing*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Lewis Pollock. 2018. Concepts and concreteness in psycholinguistics.

Giovanni Puccetti, Claudia Collacciani, Andrea Amelio Ravelli, Andrea Esuli, and Marianna Bolognesi. 2024. ABRICOT - ABstRactness and inclusiveness in COntexT: A CALAMITA challenge. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1161–1167, Pisa, Italy. CEUR Workshop Proceedings.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. *Preprint*, arXiv:2006.09479.

Paula J. Schwanenflugel and Edward J. Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1):82–102.

Gerard J. Steen, Agnes G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Thomas Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing Company, Amsterdam.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL 2021)*.

Chang Su, Weijie Chen, Ze Fu, and Yijiang Chen. 2021. Multimodal metaphor detection based on distinguishing concreteness. *Neurocomputing*, 429:166–173.

Tarun Tater, Diego Frassinelli, and Sabine Schulte im Walde. 2022. Concreteness vs. abstractness: A selectional preference perspective. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 92–98, Online. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Joshua Troche, Sebastian Crutch, and Jamie Reilly. 2017. Defining a conceptual topography of word concreteness: Clustering properties of emotion, sensation, and magnitude among 750 english words. *Frontiers in Psychology*, 8:1787.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia. Association for Computational Linguistics.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. 2025. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. *Preprint*, arXiv:2509.21305.

Christian Wartena. 2022. On the geometry of concreteness. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 204–212, Dublin, Ireland. Association for Computational Linguistics.

Christian Wartena. 2024. Estimating word concreteness from contextualized embeddings. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 81–88, Vienna, Austria. Association for Computational Linguistics.

W. C. West and P. J. Holcomb. 2000. Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. *Journal of Cognitive Neuroscience*, 12(6):1024–1037.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, and others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Louie Hong Yao, Vishesh Anand, Yuan Zhuang, and Tianyu Jiang. 2026. Rhetorical questions in llm representations: A linear probing study. *Preprint*, arXiv:2604.14128.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *Preprint*, arXiv:2301.07069.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Verb Concreteness Discrimination

We further extend this analysis to verbs, seeing if the low concrete usage of a verb is identified similarly. We first extract sentences from Wikipedia based on verbs from Brysbaert et al. (2014) and train a probe. We then construct a synthetic dataset of action verbs used in a literal vs figurative sense. For instance, "*He **grasped** the ball*" for high-concrete literal usage, "*He **grasped** the opportunity*" for low-concrete figurative usage.

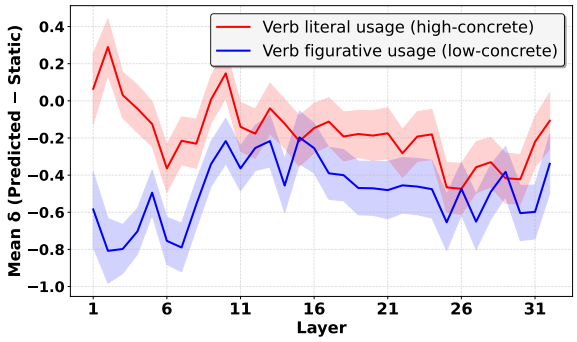


Figure 4: Mean δ across layers in Llama-3.1-8B, for verbs. Early high separation is followed by moderate to low separation in the middle to later layers.

Figure 4 shows the results for verbs in Llama-3.1-8B. Although the high-concrete verbs have a higher predicted score in all layers, the delta value is not as prominent as nouns. Hence the predicted concreteness scores for literal and figurative verb usages were similar across the layers. These observations agree with previous linguistic work which shows that verbs undergo subtler semantic extensions, while nouns exhibit clearer concreteness shifts between literal and figurative usage (Lakoff and Johnson, 1980). Intuitively, from the example above (*He **grasped** the ball* vs. *He **grasped** the opportunity*), we can see that the literal vs. figurative sense is intuitively clear. But the difference in concreteness between the two usages of the word "*grasped*" is much less explicit than in noun pairs (*She broke the **window*** vs ***window** of opportunity*).

B Dataset Distribution

Figure 5 shows the frequency distribution in the 25,000 sentences extracted from Wikipedia. There are a total of 15,853 unique nouns. The X-axis shows the number of times a word occurs in a sentence in the extracted corpus.

Figure 6 shows the probability density distribution of concreteness scores in the 25,000 extracted sentences from Wikipedia. The distribution shows the concreteness scores being spread out which ensures the diversity of the extracted sentences.

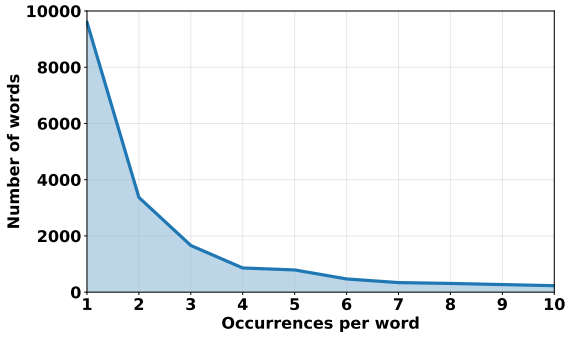


Figure 5: Word frequency distribution in the 25,000 sentences extracted from Wikipedia that we use in our experiments.

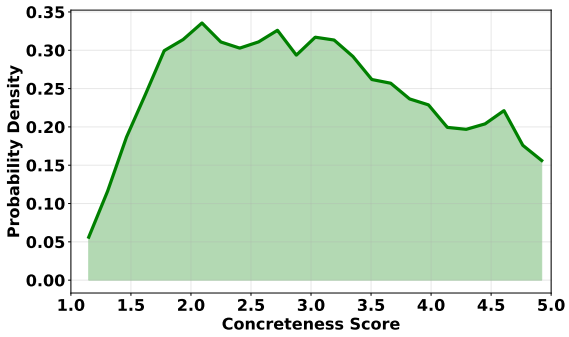


Figure 6: Concreteness score density distribution in the 25,000 sentences extracted from Wikipedia.

C Embedding Correlation Individual Results

Figure 10 shows the correlation per layer for static and contextual concreteness for Llama-3.1-8B, Qwen3-8B, Gemma2-9B and GPT-OSS-20B separately. The overall trend remains consistent. Static correlation remains consistently high across all layers. When provided with context, the correlation is highest in the early layers, and then gradually decreases as we move with layers. The trend confirms that all models have similar behavior.

An interesting observation is that the contextual correlation for all models increase in the later layers with varying degree. For Llama-3.1-8B, this rise is the least among the four models. It is more prominent in Qwen3-8B, GPT-OSS-20B and especially in Gemma2-9B.

Models	Static	Contextual
Llama-3.1-8B	0.98	0.91
Qwen3-8B	0.98	0.90
Gemma-2-9B	0.98	0.92
GPT-OSS-20B	0.98	0.86

Table 7: Pearson correlation (r) between the last token embedding and concreteness scores using different LLMs using the last hidden layer. The high correlation across all models shows the last token representation carries information about the concreteness of the term.

C.1 Correlation Table

Table 7 reports the Pearson correlations between the predicted scores and human concreteness ratings from Brysbaert et al. (2014) for 25,000 sentences from Wikipedia, using only the final layer representation of the last input token. The static setting yields extremely high correlation ($r = 0.98$), confirming that the last token in our prompt-based method robustly captures lexical concreteness. The contextual setting shows slightly lower correlations ($r = 0.90$), consistent with the fact that contextual concreteness has a high correlation with static scores, but less than the use of the word in isolation (Montefinese et al., 2023). The results support the validity of our probing method for measuring contextual concreteness in decoder-only LLMs.

D Contextual Concreteness Individual Results

Figure 11 shows the layers differentiating high concrete (literal) and low concrete (figurative) usage of the noun for the remaining three models: Qwen3-8B, Gemma2-9B and GPT-OSS-20B. As mentioned in the main body, the behavior of the models are quite similar, with contextual concreteness being discriminated at the very early layers, and remains consistent through all layers.

E Geometric Subspaces Individual Results

Figure 12 shows the one-directional subspace performance in differentiating high-concrete (literal) and low-concrete (figurative) usage of the noun for the remaining models: Qwen3-8B, Gemma2-9B and GPT-OSS-20B. Like the results in the main body, we see that concreteness is compressed into a linear subspace universally in the later layers, with AUROC scores around 0.9. The one-directional compression for Qwen3-8B and Gemma2-9B is similar to Llama, starting around the late-mid layers. For GPT-OSS-20B the compression begins in the early-mid layers. We believe GPT-OSS shows slightly different behavior due to its mixture-of-expert architecture (Lo et al., 2025). The results confirm that LLMs universally compress concreteness into a singular directional subspace in the later layers.

E.1 Effect of Subspace Dimensionality

We further examine the effect of increasing the dimensionality of the concreteness subspace. Figure 7 reports results for k = 2, 3, and 4 using Llama-3.1-8B; we observe the same qualitative trends across all evaluated models. As the subspace dimensionality increases, the AUROC scores in the middle through to the later layers consistently decrease. This degradation indicates that the concreteness signal becomes less discriminative when spread across multiple dimensions, rather than being aligned along a single dominant direction.

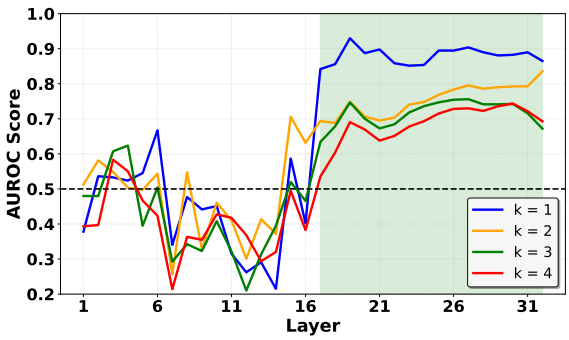


Figure 7: AUROC score for classifying high and low concrete nouns using $k = 1, 2, 3, 4$ for Llama-3.1-8B.

This behavior provides additional evidence that concreteness is primarily encoded along a unidirectional axis in deeper layers. While higher-dimensional subspaces can capture more total variance, they also incorporate secondary directions that are less consistently associated with concreteness and may reflect noise or task-irrelevant variation. Consequently, expanding the subspace dilutes the discriminative signal, leading to weaker separation between literal and figurative usages. These findings further support our central claim that concreteness is compressed into a single, coherent direction in the later layers of large language models, and that this compact representation is most effective for downstream probing and control.

F Figurative Text Datasets

In this section, we describe how we process the datasets for our experiments. Our objective is to extract sentences where the figurative meaning is realized by a noun.

Metaphor. We use two human-annotated metaphor datasets: VUAMC (VUA) (Steen et al., 2010) and MUNCH (Tong et al., 2024). We chose these two datasets because both datasets provide token-level metaphor annotations, indicating which word or phrase in a sentence is used metaphorically. We first identify the metaphor-annotated token and then apply part-of-speech tagging using SpaCy to determine whether the metaphorically used word is a noun. If so, we extract the sentence as a metaphorical sample and record the noun as the target word. From each of VUA and MUNCH, we collect 1,000 such noun-metaphor sentences. To construct a balanced dataset for classification, we additionally sample 1,000 noun usages where the target noun is annotated as *non-metaphorical*. This results in an equal distribution of metaphor and literal examples.

Idioms. We use two human-annotated idiom datasets: EPIE (Saxena and Paul, 2020) and MAGPIE (Haagsma et al., 2020). Just like the metaphor datasets, both these datasets include span-level annotations that mark idiomatic expressions within sentences. Following the same pipeline as in the metaphor setting, we identify the annotated idiomatic span and apply SpaCy POS analysis to determine whether the idiomatic expression involves a noun. Sentences meeting this criterion are extracted as positive idiom samples, with the noun recorded as the target word. We collect 1,000 idiomatic sentences from each dataset. To form a balanced dataset, we additionally sample 1,000 noun usages from instances that are explicitly marked as non-idiomatic, yielding an equal distribution of positive versus negative examples.

Metonymy. For metonymy, we use the ConMeC (Ghosh and Jiang, 2025) and MetFuse (Ghosh and Jiang, 2026), both of which are human-annotated resources. ConMeC has instances of common noun metonymy spread across six categories. MetFuse contains a combination of common noun and named entity metonymy, spread across locational or institution based metonymic mapping. In these datasets, the metonymic shift is realized through a noun, and the target noun is explicitly provided alongside the sentence. We sample 1,000 metonymic and 1,000 literal noun instances from ConMeC to obtain a balanced set. MetFuse consists of 1,000 parallel sentences consisting of literal, metonymic, metaphoric and hybrid expressions. We incorporate all the samples, using the metonymic sentences as positive and literal sentences as negative cases.

G Token Generation Prompt & MLP Details

G.1 Prompts Used

> sentence: *[sentence]*
>
> On a scale of 1 to 5 (5 being the highest), in the context of the sentence, what is the concreteness of the word "*[target_word]*"?

Figure 8: Prompt for generating contextual concreteness token. The last token representation is used to carries information regarding the concreteness of the term.

> **Static Concreteness Token Generation Prompt:**
>
> On a scale of 1 to 5 (5 being the highest), what is the concreteness of the word *[target_word]*?

Figure 9: Prompt for generating static concreteness token.

Figure 9 shows the prompt used for static concreteness information. We simply provide the target word without context. Figure 8 shows the prompt used to generate the a singular token that carries the contextual concreteness information in our prompt-based contextual probing technique. We provide the sentence, and then ask the model to predict the concreteness score of the target word in context of that sentence.

G.2 Prompt Sensitivity Analysis

We perform further prompt sensitivity analysis as our experiments show that the structure of the prompt affects the concreteness information in the last token representation.

While our prompt-based probing method reliably captures contextual concreteness, we observe sensitivity to the placement of the target noun within the prompt structure. When the target noun appears as the *final* lexical item in the prompt, the predicted concreteness scores show extremely high correlation with human norms (Pearson $r \approx 0.98$). In

contrast, when the noun occurs earlier in the sentence—followed by several contextual tokens—the correlation decreases ($r \approx 0.80 \pm 0.10$). This suggests that decoder models prioritize the most recent context in their internal representations during generation, consistent with recency-weighted attention and local next-token prediction behavior. Similar effects have been reported in studies showing that token salience increases toward the rightmost position during decoding in autoregressive LLMs, with downstream tasks benefiting when key information is placed near the model's prediction point (Mialon et al., 2023; Zhang et al., 2023; Brown et al., 2020). Our prompt was designed accordingly and we selected the prompt that had the maximum correlation score for our experiments.

G.3 MLP Details

Our MLP follows the same settings that of Wartena (2024). It consists of three hidden layers (dimensions: $512 \rightarrow 256 \rightarrow 128$), with ReLU activations and a dropout rate of 20% at each hidden layer. AdamW optimizer was used with a learning rate of 1×10^{-5}, weight decay of 1×10^{-4}, trained for 50 epochs using a batch size of 15. Using this trained regression model, we predict the concreteness score of the target noun using 10-fold cross-validation. In each fold, the model is trained on 9 folds and used to predict concreteness scores on the held-out fold. We compute Pearson correlations between the predicted and human-rated concreteness values from Brysbaert et al. (2014), averaged across folds.

H Synthetic Data Generation

H.1 Prompt Used

Figure 13 shows the prompt used to generate the synthetic data using GPT-5.1. We use three few-shot exemplars of metonymic, metaphorical and idiomatic example.

H.2 Annotation Details

For the annotation task, we recruited students and colleagues. Their participation was completely voluntary. Figure 14 provides the annotation guidelines provided to the annotators. For the synthetic dataset annotation, two annotators judged 600 high-concrete and 600 low-concrete sentences. All 600 sentences were judged to be literal by both the annotators, and 572 were judged to have the noun used in a low-concrete sense. We identified that 14 sentences were marked to low-concrete in a word-sense disambiguation sense and not in a figurative sense. We removed those cases from the final pool of 500 sentences to ensure the dataset is consistent.

For the steering task, we used two different human annotators to avoid bias. The guidelines were different, the annotators were asked to judge if the sentence as a whole is literal or figurative and not focus on an individual noun.

I Additional Steering Examples & Discussion

Table 8 are examples of steering from figurative to literal. Interestingly, we observed that when using the steering method, the literal examples *resolved* the figurative instance rather than write it write it out, which was more common in the general prompt without steering. For example, in "*The **city** attacked the new policy,*" the metonymy occurs through the noun *city*. The steered output directly resolves the figurative usage of the noun, rendering it literal. Therefore, the rewrites of the steered generation tended to resolve the figurative noun, while the non-steered generation tended to rewrite the sentence to get rid of the figurative instance.

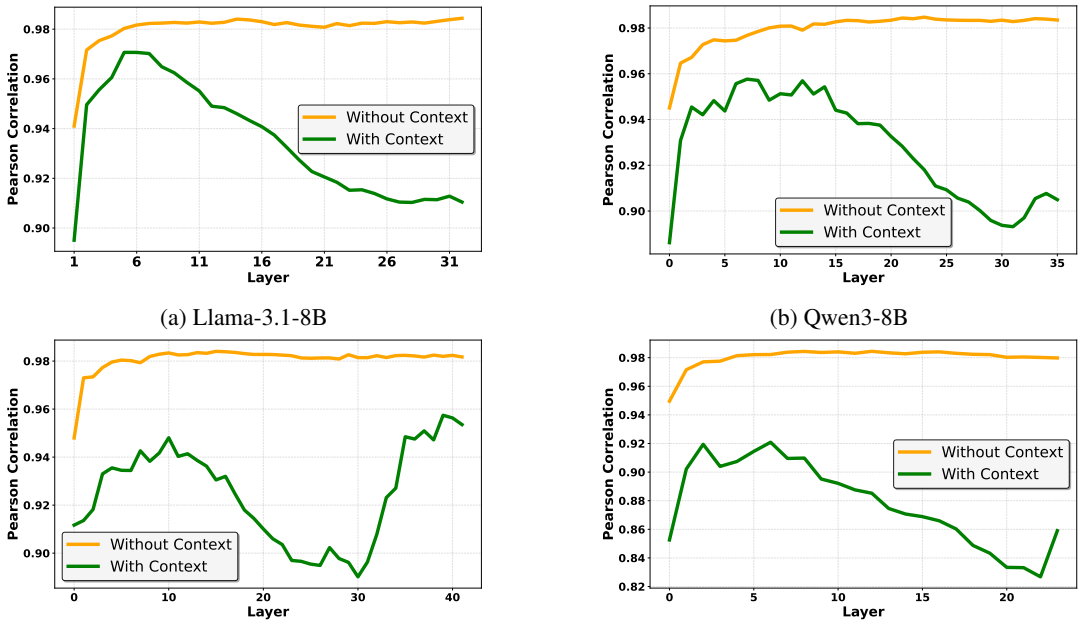
Table 5 are examples of steering from literal to figurative. As we mentioned earlier, previous work have shown that LLMs find it hard to go in this direction. Higher α values often led to unwanted noise of randomness.

J Figurative Text Classification - Remaining Results

Table 9 shows the results for remaining models: Qwen3-8B, Gemma2-9B and GPT-OSS-20B. The same trend aligns: Idioms and metaphors have better classification score, while metonymy has a lower score. As mentioned in the main body, this is due to idioms and metaphors altering concreteness more than metonymy (Lakoff, 1993; Cacciari and Tabossi, 1993; Barcelona, 2003). The consistent results across different LLM families indicate that the concreteness axis can be universally used for figurative text classification.

K Computational Resources

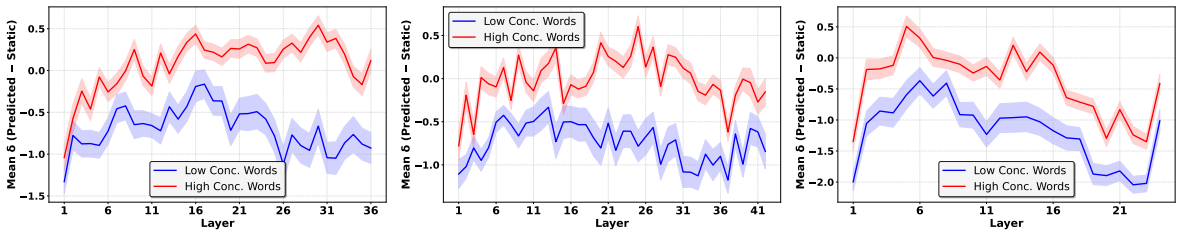
Computational resources are also an important factor in this work. Our analysis requires extracting layer-wise embeddings from large LLMs and running plenty of probing operations, which is computationally demanding. We rely heavily on GPU ac-



(a) Llama-3.1-8B (b) Qwen3-8B

(c) Gemma2-9B (d) GPT-OSS-20B

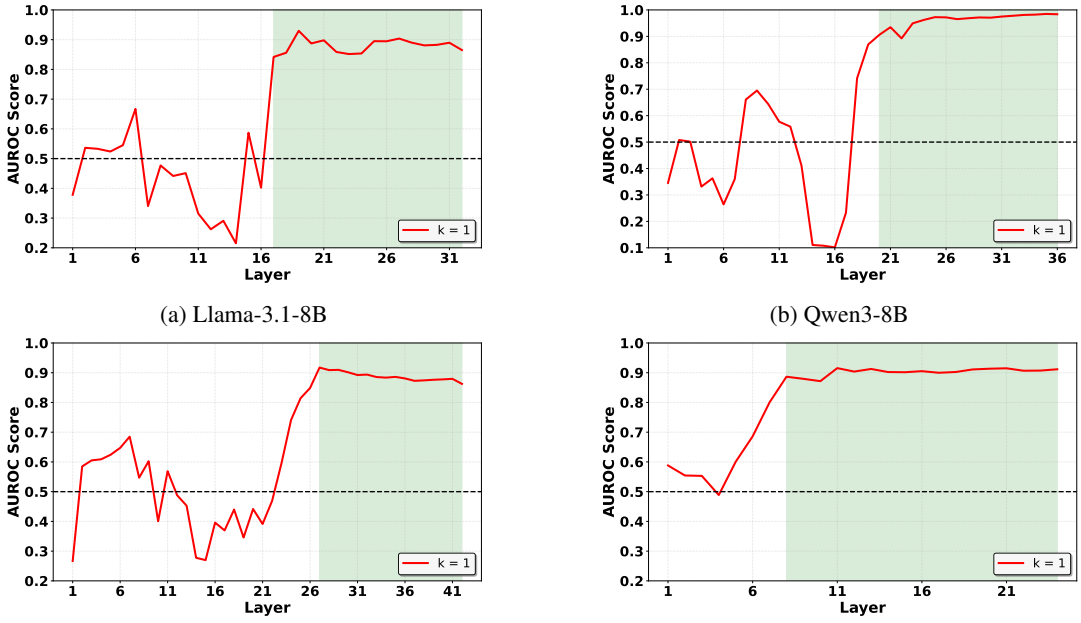
Figure 10: Pearson correlation between embeddings of models and concreteness scores from Brysbaert et al. (2014) for every layer.



(a) Qwen3-8B (b) Gemma2-9B (c) GPT-OSS-20B

Figure 11: Mean δ for all layers for remaining models.

celeration (multiple NVIDIA A100 GPUs) as well as substantial CPU time for mathematical operations such as DiffMean analysis. Runtime is a practical constraint: for example, layer-wise MLP probing with 10-fold cross-validation on larger models can take several hours to complete.



(a) Llama-3.1-8B

(b) Qwen3-8B

(c) Gemma2-9B

(d) GPT-OSS-20B

Figure 12: AUROC score for classifying high and low concrete nouns using one-directional geometric subspace for remaining models.

Original Figurative Sentence	Unsteered Rewrite	Steered Literal Rewrite
Her apology served as a bridge to mend their broken relationship.	Her apology helped mend their broken relationship. F	Her apology provided a way to heal their relationship. F
His actions served as a mirror to his true intentions.	His actions reflected his true intentions. F	His actions revealed his true intentions. F
She was the gatekeeper of her family's secrets.	She stood as the gatekeeper of her family's secrets. L	Her family's secrets were her responsibility to keep. F
The city attacked the new policy.	The city strongly opposed the new policy. L	The city's residents and businesses were against the new policy. F

Table 8: Example of concreteness guided steering for Figurative → Literal. L indicates the rewritten sentence was literal and F means figurative.

		Qwen		Gemma		GPT	
Task	**Dataset**	**Subspace**	**Full Rep.**	**Subspace**	**Full Rep.**	**Subspace**	**Full Rep.**
Idioms	MAGPIE	94.3	97.2	97.4	98.7	93.6	98.8
	EPIE	93.1	97.9	97.6	98.6	91.7	98.5
Metaphor	VUA	93.1	97.0	96.7	98.8	94.1	98.8
	MUNCH	92.0	96.3	96.5	98.6	93.2	97.9
Metonymy	ConMeC	55.4	59.4	61.2	61.5	59.3	66.5
	MetFuse	85.5	97.0	86.2	98.2	85.3	97.5

Table 9: AUROC scores for figurative text classification using Qwen3-8B, Gemma2-9B and GPT-OSS-20B comparing a unidirectional concreteness subspace learned from Wikipedia and applied to downstream datasets against a full-representation classifier trained separately on each dataset.

Synthetic Data Generation Prompt

System Prompt: You are a helpful assistant. You are given an English word that is concrete but can also be used abstractly depending on context. Here is your task:
1. Generate a sentence where this word is used in a highly concrete (or sensory) way.
2. Generate a sentence where this word is used in a less concrete (figurative) way.

Few-shot exemplar 1
Word: city
highly concrete = The *city* is 5 km away from this location.
less concrete = He was overwhelmed by the *city*.

Few-shot exemplar 2
Word: window
highly concrete = she opened the *window* as it was getting hot.
less concrete = she saw a *window* of opportunity and pounced

Few-shot exemplar 3
Word: door
highly concrete = He painted the wooden *door*.
less concrete = He closed the *door* on their relationship.

Input
Word: *[target_word]*

Figure 13: Prompt for generating static concreteness token.

Annotation guidelines

Concreteness denotes how physical or abstract a term is. Physical entities that can be perceived or touched are *concrete* terms, while words referring to intangible or abstract notions are less-concrete (or abstract term).
Your task is binary classification. You will be given a sample, each sample contains a pair of sentences where the *noun* is used in a different way, highly-concrete or less-concrete than its usual norm. The noun is usually highly-concrete in isolation, but it can also be used in a low-concrete sense when used figuratively. The pair of sentences in each sample are:

1. Literal Sentence. The noun is used in its usual high-concrete sense.
Example: *He climbed the* ***ladder*** *to reach the roof*. Here, *ladder* refers to a physical ladder, and is therefore used in its usual high-concrete sense.

2. Figurative Sentence. The noun is used in a high-concrete sense than its lexical norm, mainly in a figurative sense.
Example: *He climbed the corporate* ***ladder*** *to the position of VP*. Here, *ladder* does not refer to its physical self, but used in a metaphorical sense, therefore is less-concrete than its lexical norm.
(Note that the term does not have to be abstract. It just needs to be used in a sense than its usual norm).

Primary Guidelines:
1. If the noun carries its original concrete sense in the literal sentence, it is considered correct (mark as 1, else 0).
2. If the noun is used in a *less-concrete* sense in the supposedly figurative sentence, it is considered correct (mark as 1, else 0).

Additional Guidelines:
The intended figurative expressions are primarily metaphors, metonymy and idiom. If the noun is used in a less-concrete sense through word-sense disambiguation (eg. - square *root* of 7), please mark that sentence.

Figure 14: Annotation guidelines.