

# Frame Representation Hypothesis: Multi-Token LLM Interpretability and Concept-Guided Text Generation

Pedro H. V. Valois<sup>◇</sup> Lincon S. Souza<sup>†</sup> Erica K. Shimomoto<sup>†</sup> Kazuhiro Fukui<sup>‡</sup>

<sup>◇</sup>Graduate School of Science and Technology, University of Tsukuba, Japan

pedro@cvtlab.cs.tsukuba.ac.jp

<sup>†</sup>National Institute of Advanced Industrial Science and Technology (AIST), Japan

{lincon.souza, kidoshimomoto.e}@aist.go.jp

<sup>‡</sup>Tsukuba Institute for Advanced Research, Department of Computer Science,

University of Tsukuba, Japan

kfukui@cs.tsukuba.ac.jp

## Abstract

Interpretability is a key challenge in fostering trust for Large Language Models (LLMs), which stems from the complexity of extracting reasoning from a model's parameters. We present the Frame Representation Hypothesis, a theoretically robust framework grounded in the Linear Representation Hypothesis (LRH) to interpret and control LLMs by modeling multi-token words. Prior research explored LRH to connect LLM representations with linguistic concepts, but was limited to single token analysis. As most words are composed of several tokens, we extend LRH to multi-token words, thereby enabling usage on any textual data with thousands of concepts. To this end, we propose that words can be interpreted as frames, ordered sequences of vectors that better capture token-word relationships. Then, concepts can be represented as the average of word frames sharing a common concept. We showcase these tools through Top- $k$  Concept-Guided Decoding, which can intuitively steer text generation using concepts of choice. We verify said ideas on Llama 3, Gemma 2, Phi 3, and Qwen-2-VL families, demonstrating gender and language biases, exposing harmful content, but also potential to remediate them, leading to safer and more transparent LLMs. Code is available at this [https](https://github.com/valois/pedro) url.

## 1 Introduction

Interpretability in deep learning aims to elucidate how neural networks derive predictions. As models grow complex, understanding internal mechanisms gets challenging. By identifying factors contributing to the output, we can foster trust, safety, fairness, and improve capabilities (Hooker et al., 2019).

This study aims to enhance interpretability and control of LLMs by encoding human-comprehensible *linguistic concepts*. LLMs represent text through tokens, which can be a word, part of a word, or even a character, as per models' design. In contrast, humans better understand text through *concepts*, cognitive symbols that depict reality, often grouping objects, events, or further abstractions based on shared characteristics. We then provide tools to represent concepts within LLMs, allowing explanations suited to our mental models.

Prominent studies that offer such tools are based on the Linear Representation Hypothesis (LRH): It suggests linear operations on token vectors can explain model behavior, with concepts represented as vectors in LLM feature space (Templeton et al., 2024). For instance, we can identify the concept *female* as the average of token vectors like  $f(\text{woman})$  or  $f(\text{queen})$ . Thus, token vectors encode more than just lexical data: They also represent linguistic concepts (Mikolov et al., 2013).

Nevertheless, LRH's concepts are 1-dimensional, constraining them to single-token words, which are a minor fraction of any given language (Bau et al., 2020). As exemplified in Figure 1, concepts like *vegetarian* is exclusively linked to multi-token words, such as *meatless* or *herbivore*. Since most words are multi-token, LRH cannot always be applied in interpretability tasks effectively.

To address this shortcoming of LRH, we propose a new framework based on a key empirical observation: Our experiments show over 99% of words among several languages are composed of linearly independent token vectors. This finding

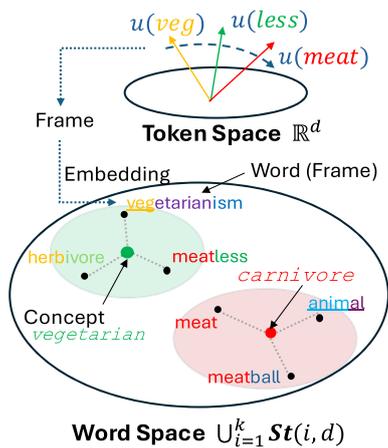


Figure 1: Frame Representation Hypothesis Overview: Tokens are vectors, which combine into words as multi-dimensional frames. In turn, Concept Frames are centroids of word sets.

indicates they are not only matrices, but k-frames—ordered sequences of independent vectors living in manifolds and possess several useful mathematical properties for language modeling. Thus, we assume all words to be frames, formalized as the Frame Representation Hypothesis (FRH). Hereafter, we use the terms **k-frame** and **frame** interchangeably.<sup>1</sup>

Starting from the postulate of FRH, we define a *Word Space*, where words are represented as frames, and equip it with a correlation formula that preserves the token whitening mechanism introduced by LRH (Park et al., 2023), allowing us to compute concepts and measure semantic relationships. Following Figure 1, the frame representation lets us identify words, such as **meatless**, **meat**, and **meatball** as distinct geometrical objects, although they have tokens in common. Then, Concept Frames like **vegetarian** or **carnivore** are centroids of word sets sharing those concepts.

Furthermore, we introduce Top-*k* Concept-Guided Decoding (Figure 2), which controls text generation by selecting the tokens which maximize a chosen concept. For example, the concept vegetarian would guide input I like to **I like fruits** if the top-3 options were **beef**, **football**, or **fruits**. This algorithm aligns model outputs with desired concepts, a practical prototype for FRH that allows meaningful LLM understanding.

<sup>1</sup>We note that while the term *frame* is used in cognitive linguistics for the theory of *frame semantics* (Fillmore, 1976), in this work, frame and k-frame refer to geometric structures.

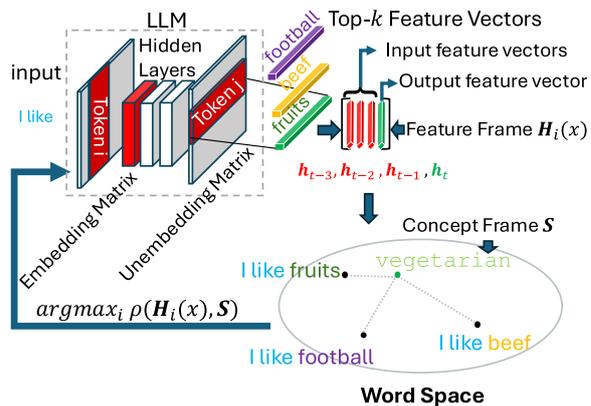


Figure 2: Top-*k* Concept-Guided Decoding Overview: Top-*k* sentence candidates are derived from the model logits, and we chose the one which maximizes the correlation with the target Concept Frames. The process is repeated in a loop until the desired number of tokens is reached.

To that end, we leverage the Open Multilingual WordNet (OMW) (Bond and Foster, 2013) as a source of synonyms to build concepts. We use over 50M words among multiple languages to build over 100,000 Concept Frames, enabling rich model understanding in a diverse yet inexpensive manner.

In short, FRH formally extends the LRH to multi-token words. We show its validity both from the theoretical and empirical points of view.

Our primary contributions are as follows:

1. Frame Representation Hypothesis as an extension of LRH to multi-token words by defining them as k-frames, thereby addressing the limitations of single-token representations.
2. Proposal of Concept Frames to represent linguistic concepts from a set of Word Frames.
3. Development of Top-*k* Concept-Guided Decoding, a proof-of-concept application to steer text generation using chosen concepts and expose model biases or potential vulnerabilities.

This document contains offensive examples illustrating LLM safety tasks. Please read selectively.

## 2 Related Work

We briefly review LLM interpretability, controllable text generation, and k-frame usage in the field.

**Language Models Interpretability** The widespread adoption of LLMs brought attention to the need for understanding their inner-workings, risks, and limitations. Several studies identified common LLM properties: Linear Representation Hypothesis, which encodes model knowledge as vectors (Mikolov et al., 2013) for explanation and editing (Wang et al., 2024), and Superposition Hypothesis (SH), which assumes specialized information is superimposed in feature spaces (Elhage et al., 2022). Causal inference also offers insights by identifying statistical cause-effect structures (Geiger et al., 2021). LRH implicitly incorporates causal principles via local counterfactual pairs, providing a geometric view to elements of causal theory, while our proposed Top- $k$  Concept-Guided Decoding simulates a causal intervention do-operation. These approaches underpin Sparse Autoencoders, that learn interpretable concept dictionaries from parameters based on SH (Elhage et al., 2022), whereas our approach maps representations using WordNet (Miller, 1995). However, both derive concepts from model parameters and might trigger spurious associations from training data (Liang et al., 2024). For a comprehensive survey, see Ferrando et al. (2024).

**Controllable Text Generation** Controllable Text Generation (CTG) employs tools that enable LLMs to generate text adhering to predefined conditions, such as stylistic features like rhythm (Ghazvininejad et al., 2017), sentiment (Krause et al., 2020), formality (Khalifa et al., 2021), keyword inclusion, reasoning (Wang et al., 2023), or safety guidelines (Kadavath et al., 2022). CTG prioritizes quality, fluency and diversity of the generated text, while adhering to the specified control parameters, enhancing LLM applicability (Wei et al., 2025). While methods involving costly retraining or finetuning exist (Zhang et al., 2023a), inference-time techniques that avoid parameter modification are common, including prompt engineering (Zhang et al., 2023b), latent space manipulation (Liu et al., 2024), and decoding-time interventions (Yang and Klein, 2021). LRH enables directly steering outputs through linear

interventions for knowledge editing (Belrose et al., 2023), while our proposal guides text generation by maximizing a certain concept. Evaluating CTG systems requires assessing both text quality and control fidelity (Ke et al., 2022), but comparisons depend heavily on the task, making it difficult to compare techniques without concrete controllability goals (Xie et al., 2023).

### Subspaces and Frames in Machine Learning

Subspaces are used for dimensionality reduction, feature extraction (Fukui et al., 2023), classification (Watanabe, 1967), interpretability (Valois et al., 2024), and modelling text sentences (Shimomoto et al., 2021). Their invariance to selection of basis is advantageous for representing clusters, but limiting for ordered structures. k-frames represent ordered sequences of vectors and are used to model redundant and oriented data, being applied in error correction (Kovacevic and Chebira, 2008), signal decomposition (Casazza et al., 2013), and optimization problems (Mankovich and Birdal, 2023). To the best of our knowledge, k-frames were never applied in NLP to model concepts as proposed here.

## 3 Preliminary

We now introduce our proposal’s background. Vectors are denoted as bold lowercase letters, *e.g.*,  $\mathbf{v}$ ; matrices as bold uppercase letters, *e.g.*,  $\mathbf{M}$ ; monospace lowercase letters for tokens, *e.g.*,  $x$ ; vector spaces with calligraphic letters, *e.g.*,  $\mathcal{U}$ ; words with sans serif uppercase letters, *e.g.*,  $\mathbf{W}$ ; and concepts with monospace uppercase letters, *e.g.*,  $C$ .

### 3.1 k-frames

A k-frame is a sequence of  $k$  linearly independent vectors in  $\mathbb{R}^d$ , represented by  $\mathbf{F} \in \mathbb{R}^{d \times k}$ ,  $\text{rank}(\mathbf{F}) = k$ . The set of all k-frames in  $\mathbb{R}^d$  constitutes the non-compact Stiefel manifold  $\text{St}(k, d)$ . Manifolds are structures in which distance, geodesics, and more may be defined, so we can compute geometrically meaningful relationships between distinct k-frames (Edelman et al., 1998). The set of all frames up to rank  $q$  forms the  $q$ -complete Stiefel manifold  $\text{CSt}(q, d) = \bigcup_{i=0}^q \text{St}(i, d)$ —a disjoint union of Stiefel manifolds—where the null frame  $\text{St}(0, d) \equiv \emptyset$  is defined as our space’s origin. Intuitively,  $\text{CSt}(q, d)$  is a stratified structure, so its base is the null frame.

### 3.2 Rays

A ray  $\mathcal{R}$  is a directed half-infinite line, also known as 1-dimensional convex cone, half-line, or axis, defined by a vector  $\mathbf{v} \in \mathbb{R}^d$  and its scalar multiples  $\mathcal{R}(\mathbf{v}) = \{\alpha\mathbf{v} \mid \alpha \geq 0\}$ . A ray is represented by a normalized vector  $\mathbf{v}' = \mathbf{v}/\|\mathbf{v}\|$ , which is also a point in  $\text{St}(1, d)$ . A single dimensional subspace contains two rays, and their correlation is measured as the cosine of the angle  $\theta$  between their normalized vectors (Mathematics, 2016)

$$\rho(\mathcal{R}(\mathbf{v}), \mathcal{R}(\mathbf{u})) = \frac{\langle \mathbf{v}' | \mathbf{u}' \rangle}{\|\mathbf{v}'\| \|\mathbf{u}'\|} = \mathbf{v}'^\top \mathbf{u}' = \cos \theta. \quad (1)$$

Moreover, notice that the correlation shown in Equation (1) is connected to the inner product – projection of one vector onto another. In this work, we use the term “projection” when calculating correlation with unnormalized vectors.

### 3.3 Large Language Models

LLM models process text by converting it into a sequence of tokens, *embedding* them into its own vector space and processing this sequence of vectors through its hidden layers to a final vector representation, which is *unembedded* into the most likely token to continue the input sentence. A simple version of such pipeline is illustrated in Figure 2.

A token is a single element of a textual sequence, represented by a number  $x \in \mathcal{V}$  in a predefined vocabulary  $\mathcal{V} \subset \mathbb{Z}^+$ . In that sense, the model’s tokenizer converts text input  $x$  into token  $t$ -tuple  $(x_1, x_2, \dots, x_t) \in \mathcal{V}^t$ . The LLM then starts in the *embedding* layer, which maps each token number  $a \in \mathcal{V}$  to an unique *embedding* vector  $\mathbf{e}(a) \in \mathcal{E} \cong \mathbb{R}^d$ , each of which is a column of the *embedding* matrix  $\mathbf{W}_{\mathcal{E}} \in \mathbb{R}^{d \times |\mathcal{V}|}$ . Therefore, the output of this layer is the  $t$ -tuple of *embedding* vectors  $\mathbf{e}(x) = (\mathbf{e}(x_1), \mathbf{e}(x_2), \dots, \mathbf{e}(x_t))$ . Next,  $\mathbf{e}(x)$  is processed by the DNN hidden transformer layers into the feature vector  $\mathbf{h}(x) = \mathbf{h}(\mathbf{e}(x_1), \mathbf{e}(x_2), \dots, \mathbf{e}(x_t)) \in \mathcal{H} \cong \mathbb{R}^d$ .

Then, the LLM converts  $\mathbf{h}(x)$  into a token number. The *unembedding* vector of token  $b \in \mathcal{V}$  is  $\mathbf{u}(b) \in \mathcal{U} \cong \mathbb{R}^d$ , a row of the *unembedding* matrix  $\mathbf{W}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ , which also identifies each token to a unique vector in high-dimensional space  $\mathcal{U}$ .

Finally, the probability of a token  $y \in \mathcal{V}$  being next in a text sentence  $x$  is determined with softmax

$$p(y|x) \propto \exp(\mathbf{u}(y)^\top \mathbf{h}(x)). \quad (2)$$

In practice, the space dimension  $d$  can range from 1024 to 16384, while the vocabulary  $\mathcal{V}$  usually contains from 50,000 to 300,000 tokens.

### 3.4 Linear Representation Hypothesis

We now concentrate the discussion into the geometry of  $\mathcal{H}$  and  $\mathcal{U}$  and their relationships.

Mikolov et al. (2013) first introduced the Linear Representation Hypothesis within the domain of word embeddings, observing that linear vector operations could reveal underlying semantic and syntactic relationships, such as  $\mathbf{u}(\text{female}) - \mathbf{u}(\text{male}) + \mathbf{u}(\text{brother}) \approx \mathbf{u}(\text{sister})$ . Thus, there is an indication that difference vectors can encode abstract concepts like “gender”, leading to the hypothesis that all models might represent information linearly. From an interpretability point of view, one may leverage these concepts to analyze the model knowledge or control its output.

With that in mind, Park et al. (2023) observed pairs of tokens building the same difference vector in LLMs unembedding spaces, e.g.,  $\mathbf{u}(\text{female}) - \mathbf{u}(\text{male}) \approx \mathbf{u}(\text{queen}) - \mathbf{u}(\text{king}) \approx \mathbf{u}(\text{wife}) - \mathbf{u}(\text{husband})$ . By averaging these token pairs, they were able to obtain a vector approximating a direction the model reserves for a “gender” concept.

Formally, a direction can be represented as a ray. Therefore, the literature hypothesizes every concept  $c$  the model knows should have its own ray  $\mathcal{R}(s'_c) \subset \mathcal{U}$  of concept vector  $s'_c \in \mathcal{U}$ . Concept vectors are computed as the normalized mean of difference vectors of counterfactual pairs

$$\mathbf{u}'_c = \sum_i^{n_c} (\mathbf{u}_i(c=1) - \mathbf{u}_i(c=0)), \quad (3)$$

$$s'_c = \frac{\mathbf{u}'_c}{\|\mathbf{u}'_c\|}, \quad (4)$$

where  $\mathbf{u}_i(c=1), \mathbf{u}_i(c=0)$  is a counterfactual token pair, so  $c=1$  indicates one concept direction while  $c=0$  its opposite, e.g., concept **English**  $\Rightarrow$  **Spanish** ( $c=0$ ) is computed using difference vectors like  $\mathbf{u}(\text{good}) - \mathbf{u}(\text{bueno}), \mathbf{u}(\text{bad}) - \mathbf{u}(\text{malo})$ , while **Spanish**  $\Rightarrow$  **English** ( $c=1$ ) is the opposite

vector. Notice that the order of the concepts is arbitrary, so `Spanish`  $\Rightarrow$  `English` would also be a valid concept, yet with a difference of sign.

Hereafter, concepts connect to each other through linear operations. The correlation  $\rho$  of concepts  $A, B$  is a probe for model understanding

$$\rho(A, B) = \langle \mathbf{s}'_A, \mathbf{s}'_B \rangle = \mathbf{s}'_A{}^\top \mathbf{M} \mathbf{s}'_B, \quad (5)$$

where  $\mathbf{M} = \text{Cov}^{-1}(\mathbf{W}_U)$  is a whitening matrix that defines the LRH inner product, placing unrelated concepts as orthogonal to each other.

### 3.5 WordNet

The Open Multilingual WordNet (OMW) is a collaborative project that intersects cognitive psychology, linguistics, and computer science to create an interconnected network of lexical databases (Bond and Foster, 2013; Fellbaum, 1998; Harabagiu et al., 1999). At its core are **synsets** and **lemmas**. A synset, short for ‘‘synonym set’’, is a group of words or phrases that may share the same meaning. For example, `{car, automobile, auto}` forms a synset, which can be uniquely identified as `car.n.01`, *i.e.*, the 1st dictionary meaning of the word `car` as a noun (Miller et al., 1990).

A lemma, on the other hand, is a canonical form, *e.g.*, `run, runs, ran, and running` are all represented by the lemma `run` (Fellbaum, 2010). In OMW, synsets from different languages are linked to their equivalent English ones, allowing for cross-lingual connections. Therefore, the English synset `car.n.01` will also include Spanish lemmas, such as `coche` or `automóvil` (Bond and Paik, 2012). Each lemma can belong to multiple synsets, reflecting its different meanings, making OMW a powerful NLP tool (Wagner, 2010).

## 4 Frame Representation Hypothesis

In this section, we introduce our theoretical framework. Proofs are provided in Appendix A.

### 4.1 Linear Decomposition of Tokens

LRH posits that concepts are linearly encoded within LLMs feature spaces. Moreover, the Superposition Hypothesis (SH) suggests that models encode information in a *superposition of concepts* because the number of possible concepts significantly exceeds space dimensionality (Elhage et al., 2022), a phenomenon visible in Figure 3. This is formalized as a linear combination of vectors at Postulate 4.1.

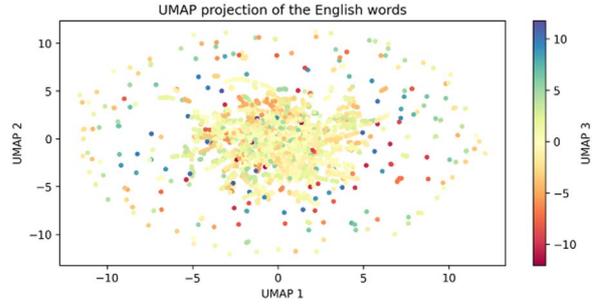


Figure 3: Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) of the 10k most frequent single-token English words for Gemma 2. While some points are clearly separated, others overlap due to the Superposition Hypothesis (SH). For example, `ad` is a token in the unrelated words `advertisement, admit, adventure`, etc., while `restaurant` is a single token and it is not found in other words.

**Postulate 4.1.** Let  $\mathbf{u}(y)$  be the unembedding representation of token  $y \in \mathcal{V}$ , then it is a linear combination of *concept vectors*  $\mathbf{s}$

$$\mathbf{u}(y) - \mathbf{u}_0 = \sum_i^s a_i \mathbf{s}_i, \quad (6)$$

where  $a_i \in \mathbb{R}$ ,  $s$  is the number of all concepts known by the model, and  $\mathbf{u}_0$  is a *meaningless* vector—an offset element from the fact that not all tokens might hold meaning, *e.g.*, `[PAD]` or `[EOS]`, implying we need to remove the *meaningless* part of each token vector. Heuristically,  $\mathbf{u}_0$  should be the unembedding vector average, so that  $\mathbb{E}[\mathbf{u}(y) - \mathbf{u}_0] = 0$ . Also,  $\{\mathbf{s}_i\}_{i=1}^s$  is not a basis: Words can be grouped in several ways, *e.g.*, antonyms or synonyms, making concepts interdependent.

Therefore, a concept is the centroid of a set of tokens sharing that same concept. Let  $\{y_j\}_{j=1}^n$  be a set of tokens sharing a common concept  $s$ , we estimate the concept as the token average<sup>2</sup>

$$\mathbf{s} \propto \sum_{j=1}^n \mathbf{u}(y_j) - \mathbf{u}_0. \quad (7)$$

#### 4.1.1 Combined Concepts

We connect Concept Estimation (7) to the ray  $\mathcal{R}(\mathbf{s}'_C)$  of a concept  $C$  by separating Equation (3) into two sums, each its own concept. Therefore, a concept  $C$  has representation  $\mathcal{R}(\mathbf{s}'_C)$ , where  $\mathbf{s}'_C$  is a normalized counterfactual concept pair difference,

$$\mathbf{s}'_C = \frac{\mathbf{s}_{C-1} - \mathbf{s}_{C-0}}{\|\mathbf{s}_{C-1} - \mathbf{s}_{C-0}\|}, \quad (8)$$

<sup>2</sup>Hereafter consider all tokens to be already debiased.

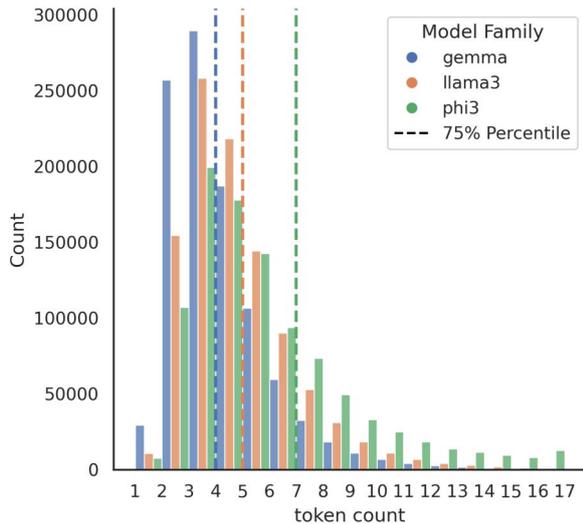


Figure 4: Histogram of lemma token count among all OMW lemmas. The dashed vertical bar indicates the 75% percentile for each model family.

indicating some concepts are formed by other concepts. For example, a set of tokens sharing the meaning of `female` builds  $s_{\text{female}}$ , while another sharing the concept of `male` builds  $s_{\text{male}}$ , forming

$$s'_{\text{male} \Rightarrow \text{female}} \propto s_{\text{female}} - s_{\text{male}}, \quad (9)$$

which leads us to understand some concepts as building blocks for Combined Concepts.

## 4.2 Generalizing from Tokens to Words

The previous discussion can leverage WordNet to determine concepts. WordNet’s structure overlaps with LLM representations (Moskvoretskii et al., 2024a,b; Park et al., 2024), and OMW synsets are sets of multilingual lemmas sharing a meaning, making it well suited for Equation (7). Nevertheless, Section 4.1 only deals with single-token words, which accounts for less than 1% of all OMW lemmas on most LLMs, significantly limiting estimated concepts quality (*c.f.* Figure 4).

In order to make LRH theory more relevant for LLM interpretability, we must generalize Section 4.1 to words made of multiple tokens. At first glance, the insight is simple: *tokens do not build concepts – tokens build words – and words build concepts*. Nonetheless, words are sequences of tokens with a well-defined order, implying we cannot simply average them to a single vector or their meaning might be lost, as exemplified at Table 1. As the following sections demonstrate, we can work directly with the word matrices by

Synset	Lemma
<code>ad.n.01</code>	<code>ad</code>
<code>myth.n.01</code>	<code>mit</code>
<code>admit.v.01</code>	<code>admit</code>
<code>half.n.02</code>	<code>mitad</code>

Table 1: An example of a token pair—`ad` and `mit`—being used to form different words, each portraying different concepts. The Gemma 2 model family (Google, 2024) tokenizes `admit`, in the OMW synset of `admit.v.01`, into `ad` in `ad.n.01`, and the Polish word `mit` in `myth.n.01`. Concurrently, in opposite order they form the Spanish word `mitad`, present in `half.n.02`.

understanding them as frames, compute a Concept Frame from a matrix average and a Combined Concept as the difference of Concept Frames.

### 4.2.1 Words as Frames

In practice, every word is split into  $t$  independent tokens  $(w_1, w_2, \dots, w_t)$ , and represented in  $\mathcal{U}$  as the column matrix of shape  $d \times t$  from the concatenation of sub-word unembedding token vectors

$$\mathbf{W} = (\mathbf{u}(w_1) \quad \mathbf{u}(w_2) \quad \dots \quad \mathbf{u}(w_t)). \quad (10)$$

First, consider the following principle about the nature of word matrices. Let two token vectors  $\mathbf{a}, \mathbf{b} \in \mathcal{U}$ . If these vectors were collinear, meaning  $\mathbf{b} = \alpha \mathbf{a}$  for some  $\alpha \in \mathbb{R}$ , then Equation (2) would consistently assign higher probabilities to one token over the other regardless of input. This would effectively make some tokens redundant, as they would never be the most probable choice in any context. Such a scenario contradicts the assumption that each token learned by the model must have some context in which it is the optimal choice, *i.e.*, all learned tokens must be meaningful and usable. Notice, however, that we only consider tokens learned by the model instead of full vocabulary because it might contain tokens never seen during training or not properly learned. We assume the model has learned all tokens from words in its supported languages. Thus, we conclude no two token vectors are collinear.

In that sense, word matrices as in Equation (10) are constrained so that no token vector  $\mathbf{u}(w_i)$  can be expressed as a scalar multiple  $\alpha \mathbf{u}(w_j)$  of another token vector, for any  $\alpha \in \mathbb{R}$ . This non-collinearity

constraint defines a locally Euclidean open subset of  $\mathbb{R}^{d \times t}$ , thereby forming a manifold. The softmax operation in DNN training ensures the space of all words acquires a manifold structure.

While non-collinearity is a necessary condition, it does not enforce  $\mathbf{W}$  as full-rank—the  $t$  token vectors may exist in a subspace of dimension less than  $t$ . However, we assume rank deficiency may compromise expressiveness and computational stability, suggesting the need for additional constraints.

From an NLP perspective, unique word representations are essential. If a word was not a linearly independent matrix, we could eliminate dependent tokens until made full-rank, yielding an alternative representation of the same word in  $\mathcal{U}$ . However, this sacrifices word uniqueness, which is undesirable for consistent language modeling.

To address these concerns, we propose modeling words as  $k$ -frames, *i.e.*, we assume  $\mathbf{W} \in \text{St}(t, d)$ . Then, all word matrices are supposed full-rank. Our empirical investigation (*c.f.* Section 5.1) supports this framework, revealing that over 99% of words in OMW exhibit linear independence among their token vectors. This assumption is facilitated by the high dimensionality of  $\mathcal{U}$ , easily representing words as full-rank matrices.

#### 4.2.2 Frame Correlation

Each word has its own token count  $t$ , so each belongs to a  $\text{St}(t, d)$ , where  $t = 1, 2, \dots, k$ , *i.e.*,  $k$  is the max number of tokens in any word. Thus, we can define the set of all words the *Word Space* as  $\text{CSt}(k, d)$ . Let  $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_{t_1}) \in \text{St}(t_1, d)$ ,  $\mathbf{B} = (\mathbf{b}_1 \dots \mathbf{b}_{t_2}) \in \text{St}(t_2, d)$  be in  $\text{CSt}(k, d)$ , we employ the asymmetric Procrustes distance (Ye and Lim, 2016; Mandolesi, 2022) as the space metric,

$$d_{\mathbf{P}}^*(\mathbf{A}, \mathbf{B}) = \sqrt{t_1 + t_2 - 2 \sum_j^{\min t_1, t_2} \mathbf{a}_j \mathbf{M} \mathbf{b}_j}, \quad (11)$$

where  $\mathbf{M}$  comes from Equation (5).

Hereafter, we can propose frame correlation by applying the law of cosines to generalize Equation (1) only in terms of distance functions:

$$\rho(\mathbf{A}, \mathbf{B}) = \frac{\|\mathbf{A}\|_{\mathbf{P}}^2 + \|\mathbf{B}\|_{\mathbf{P}}^2 - d_{\mathbf{P}}^*(\mathbf{A}, \mathbf{B})^2}{2\|\mathbf{A}\|_{\mathbf{P}}\|\mathbf{B}\|_{\mathbf{P}}} \quad (12)$$

$$= \frac{\sum_j^{\min t_1, t_2} \mathbf{a}_j \mathbf{M} \mathbf{b}_j}{\sqrt{t_1 t_2}}, \quad (13)$$

where  $\|\mathbf{A}\|_{\mathbf{P}} = d_{\mathbf{P}}^*(\mathbf{A}, \emptyset) = \sqrt{t_1}$ ,  $\emptyset$  is the null frame (origin) of  $\text{CSt}(k, d)$ , so  $\text{rank}(\emptyset) = 0$ .

Such correlation can measure relationships as similar (positive), unrelated (null), or opposite (negative). For instance, **yeah** and **yes** are similar words and should have correlation close to 1, while **yes** and **bubble** are orthogonal, but antonyms such as **yes** and **no** would be negatively correlated.

#### 4.2.3 Concept Frame

We estimate concepts as the Fréchet mean of a word set—the point minimizing the distance to each word—effectively capturing the concept they collectively represent (Marrinan et al., 2014).

Let  $\{\mathbf{W}_i\}_{i=1}^n$  be a set of words,  $\mathbf{W}_i = (\mathbf{u}(\mathbf{w}_{i1}) \quad \mathbf{u}(\mathbf{w}_{i2}) \quad \dots \quad \mathbf{u}(\mathbf{w}_{ik_i})) \in \text{St}(k_i, d)$ , and let  $\mathbf{S} = (\mathbf{s}_1 \quad \mathbf{s}_2 \quad \dots \quad \mathbf{s}_k) \in \text{St}(k, d)$ ,  $k = \max k_i$  be the Concept Frame, it is determined as

$$\mathbf{S} = \arg \min_{\mathbf{S} \in \text{St}(k, d)} \sum_{i=1}^n d_{\mathbf{P}}^{*2}(\mathbf{W}_i, \mathbf{S}) \quad (14)$$

$$= \arg \max_{\mathbf{s}_j \in \text{St}(1, d)} \sum_{i=1}^n \sum_{j=1}^{k_i} \mathbf{u}(\mathbf{w}_{ij})^\top \mathbf{M} \mathbf{s}_j. \quad (15)$$

We can extend the sum at Equation (15) from  $k_i$  to  $k$  by noticing its equivalence to having  $\mathbf{u}(\mathbf{w}_{ij}) = 0$  for all  $k_i < j \leq k$ . Let's define  $\mathbf{W}'_i = (\mathbf{u}(\mathbf{w}_{i1}) \quad \mathbf{u}(\mathbf{w}_{i2}) \quad \dots \quad \mathbf{u}(\mathbf{w}_{ik_i}) \quad 0 \quad 0 \quad \dots \quad 0)$  as the right-padded  $\mathbf{W}_i$  with  $k - k_i$  zeros. Thus,

$$\mathbf{S} = \arg \max_{\mathbf{s}_j \in \text{St}(1, d)} \sum_{i=1}^n \sum_{j=1}^k \mathbf{W}'_{ij}{}^\top \mathbf{M} \mathbf{s}_j \quad (16)$$

$$= \arg \max_{\mathbf{s}_j \in \text{St}(1, d)} \sum_{j=1}^k \left( \sum_{i=1}^n |\mathbf{W}'_{ij}{}^\top| \right) \mathbf{M} \mathbf{s}_j \quad (17)$$

$$= \arg \max_{\mathbf{S} \in \text{St}(k, d)} \text{tr}(\bar{\mathbf{W}}'^\top \mathbf{M} \mathbf{S}), \quad (18)$$

where  $\bar{\mathbf{W}}' = \sum_{i=1}^n \mathbf{W}'_i$  is the padded word sum.

Finally, Equation (18) is the Procrustes problem, which Schönemann (1966) has solved with

$$\mathbf{S} = \mathbf{U} \mathbf{V}^\top, \quad (19)$$

and  $\bar{\mathbf{W}}'^\top \mathbf{M} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$  is the SVD decomposition of the padded word sum. Hence, under the Procrustes distance the Concept Frame is the solution of a Procrustes problem.

#### 4.2.4 Combined Concept Frames

In Section 4.1.1, we defined Combined Concepts as concept vector differences, which we extend

to FRH by placing them in the Stiefel manifold. In other words, given a pair of Concept Frames  $\mathbf{A}, \mathbf{B} \in \text{St}(k, d)$ , we can build the Combined Concept Frame  $\mathbb{D} = \mathbb{D}(\mathbf{B}, \mathbf{A}) \in \text{St}(k, d)$  by enforcing it to be the frame closest to  $\mathbf{B} - \mathbf{A}$ :

$$\mathbb{D} = \arg \max_{\mathbb{D} \in \text{St}(k, d)} \text{tr}((\mathbf{B} - \mathbf{A})^\top \mathbf{M} \mathbb{D}) \quad (20)$$

$$= \mathbf{U}_{\mathbb{D}} \mathbf{V}_{\mathbb{D}}^\top, \quad (21)$$

where  $(\mathbf{B} - \mathbf{A})^\top \mathbf{M} = \mathbf{U}_{\mathbb{D}} \Sigma_{\mathbb{D}} \mathbf{V}_{\mathbb{D}}^\top$  is the SVD.

### 4.3 Concept Probing

The framework established for  $\mathcal{U}$  can be extended to the feature space  $\mathcal{H}$  by reinterpreting Equation (2) as  $\text{logit } p(y|x) = \rho(\mathbf{u}(y), \mathbf{h}(x))$ . Thus, the correlation between  $\mathbf{u}(y)$  and  $\mathbf{h}(x)$  can be understood as a linear probe from space  $\mathcal{U}$  to  $\mathcal{H}$ .

Consequently, there is a correspondence between frames in  $\mathcal{U}$  and  $\mathcal{H}$ . Let a Feature Frame  $\mathbf{H}$  be the last  $k$  feature vectors of the input sequence

$$\mathbf{H}(x) = (\mathbf{h}_{t-k+1} \ \mathbf{h}_{t-k+2} \ \dots \ \mathbf{h}_t) \in \text{St}(k, d),$$

we probe  $x$  for Concept Frame  $\mathbf{S} \in \text{St}(k, d)$  using the correlation defined at Equation (12),

$$\text{logit } p(\mathbf{S}|x) = \rho(\mathbf{S}, \mathbf{H}(x)). \quad (22)$$

### 4.4 Top- $k$ Concept-Guided Decoding

We can leverage concept probing as a mechanism for Controllable Text Generation (CTG) via Concept-Guided Text Generation (Figure 2). This approach can be implemented with sample-based decoding methods, such as Top- $k$  sampling, first generating a set of  $k$  potential tokens from which the next token is randomly selected. We propose to alter such process wherein the next token  $\mathbf{x}_{t+1}$  of input sequence  $x = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$  is the one which maximizes its respective Feature Frame correlation onto a target Concept Frame  $\mathbf{S}$ ,

$$\mathbf{x}_{t+1} = \arg \max_{i \in \{1, 2, \dots, k\}} \rho(\mathbf{S}, \mathbf{H}_i(x)). \quad (23)$$

This methodology can align model output with a desired concept and serves as a practical prototype for FRH, showing how to direct text generation and understand model behavior meaningfully.

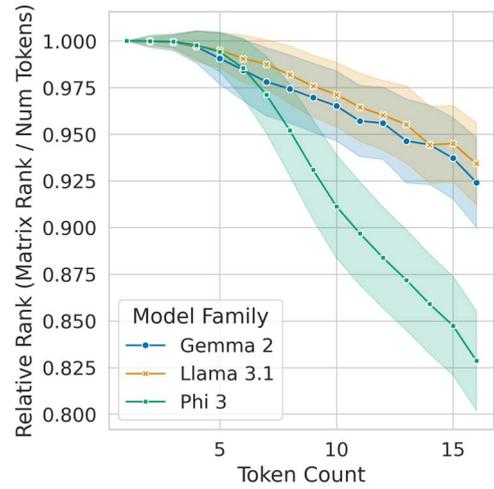


Figure 5: Relative Rank as a function of token count for all OMW lemmas and model families. As token count increases, some OMW lemmas are compound names, phrases or expressions, and not words. If constrained to words, over 99% of them are full-rank. Phi 3 has lower overall rank for longer lemmas than other models.

## 5 Experiments

In this section, we validate FRH for words and concepts, showing guided generation of sentences. We use Llama 3.1 (Meta, 2024), Gemma 2 (Google, 2024), and Phi 3 (Microsoft, 2024) LLM families and OMW only with supported languages. Further discussion is available in the Appendices.

### 5.1 Frame Representation Hypothesis

FRH posits that LLMs encode words as frames. We can empirically evaluate this hypothesis by analyzing if words are made of linearly independent vectors, which we can measure by computing its rank. In Figure 5, we see near-maximum matrix ranks for lemmas comprising up to 3–4 tokens, which is the token count that represents words. In OMW, lemmas with token counts of 5 and beyond mostly represent compound words and expressions, implying the frame representation fits 99.8% words. Notably, Phi 3 shows a rapid rank decrease beyond token count of 5, likely due to its high proportion of lemmas with large token count, making non full-rank lemmas more common (*c.f.* Figure 4).

Furthermore, given that we propose using OMW synsets to build Concept Frames, we must verify if these synsets fit the model representation or not. To that end, we can compute the projection (unnormalized correlation) of Word Frames onto their corresponding Concept Frames for all

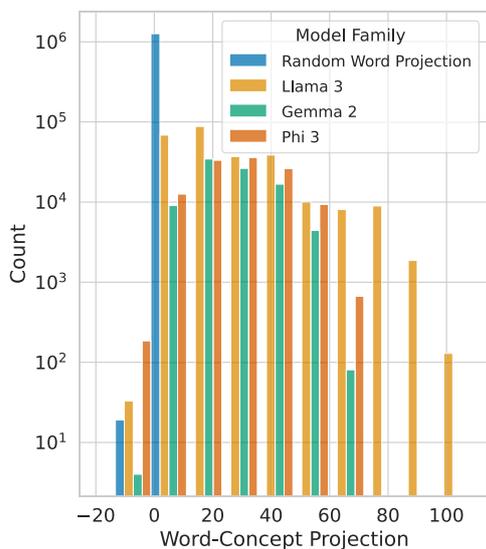


Figure 6: Distribution of Word Frame projection lengths. Random frames have near-zero projection with any Concept Frame, while words show positive projections onto associated Concept Frames.

OMW synsets and lemmas. Figure 6 reveals that random frames are consistently unrelated to concept frames across models, while words exhibit positive projections onto their associated concepts.

Moreover, we must verify if FRH is capable of dealing with “noisy” tokens. Particularly, tokens that are part of several different words, such as `ism` being part of words like `criticism`, `realism`, `journalism`. Traditional LRH methods struggle with these tokens because they are superpositions of several concepts, leading to noise in Equation (7), *i.e.*, concepts which don’t correlate with the tokens that built it. For example, to obtain a vector representation of `admit`, it would be required to average `ad` and `mit`, but this is unlikely to yield a vector that accurately represents the intended concept of `admit.v.01`, as seen in Table 1.

Theoretically, FRH solves this problem by avoiding vectors and keeping the matrix representation intact. In Figure 7, we compare correlations between concepts and its words for OMW synsets containing lemmas with tokens `ism`, `ment` and `ship`. We find that while some concepts show low correlation with its words, FRH distribution is skewed to higher correlations than LRH, which indicates it more robustly deals with tokens belonging to several concepts. Thus, concepts are better represented as frames rather than vectors.

These findings support FRH consistency with models’ internal representations and suggest

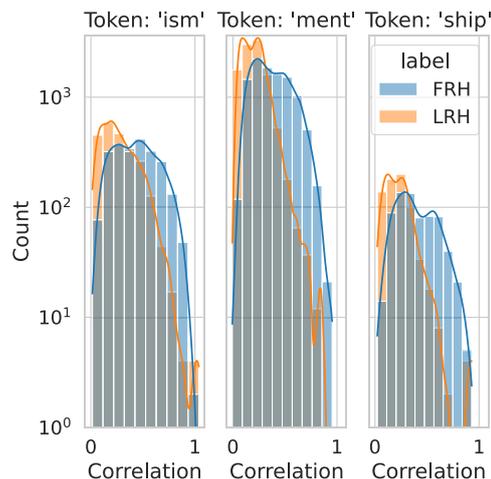


Figure 7: Distributions of Word-Concept Correlation for concepts that contains the tokens `ism`, `ment`, and `ship`. These tokens introduce noise in the LRH concept estimation based on averages, while in FRH this problem does not occur.

LLMs inherently correlate with OMW. In the following experiments, we use lemmas up to 4 tokens to ensure our theory is applied only for full-rank matrices.

## 5.2 Guided Generation

We explore applied FRH with Concept-Guided Decoding, exposing biases and vulnerabilities.

### 5.2.1 Qualitative Analysis

We first compare model outputs on a few inputs and concepts likely sensitive to biases. Figure 8 demonstrates the impact of concept-guidance on the characterization of `men` by Llama 3.1 8B Instruct. With no guidance, the model focuses on family roles. When guided by the Concept Frame `woman.n.01 - man.n.01`, this tendency is seemingly amplified. However, a more significant shift in narrative occurs when guided by the opposite concept of `man.n.01 - woman.n.01`, emphasizing a perceived importance as *family providers*.

At Figure 9, the unguided model’s characterization of `women` primarily enumerates family roles, which may be interpreted as a balanced output when juxtaposed with the default answer for `men` at Figure 8. Despite that, `woman.n.01 - man.n.01` makes the model emphasize biological traits. Conversely, `man.n.01 - woman.n.01` highlights leadership roles and esteemed social positions.

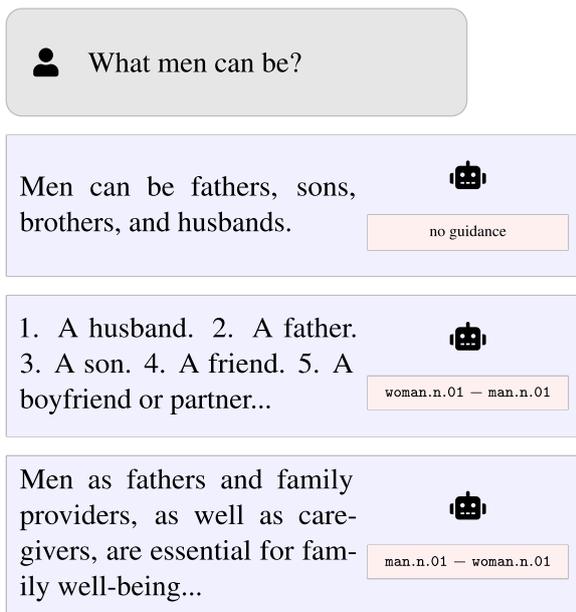


Figure 8: Concept-guided generation examples when the model is prompted to describe **men**.

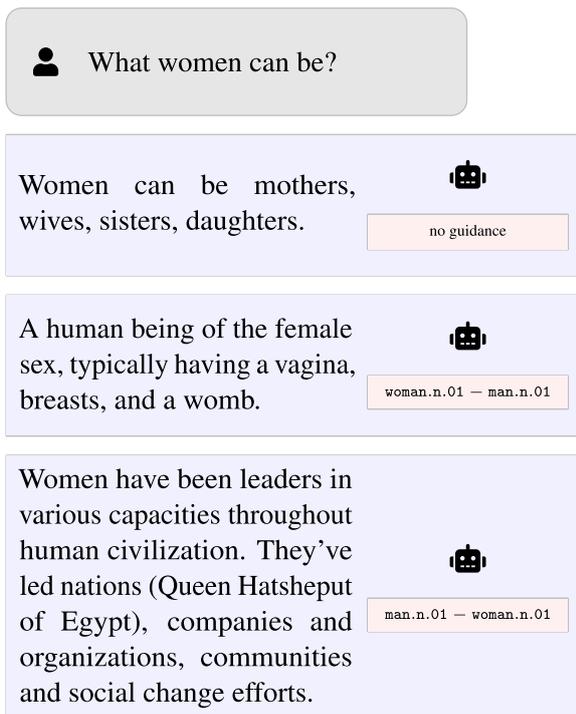


Figure 9: Concept-guided generation examples when the model is prompted to describe **women**.

The stark contrast in each example suggests that, when guided by  $\mathbb{D}(B, A)$ , the model attempts to maximize attributes it associates with the first concept  $B$  while minimizing  $A$ . These illustrate how to influence text generation, exposing biases and stereotypes within the model's learned

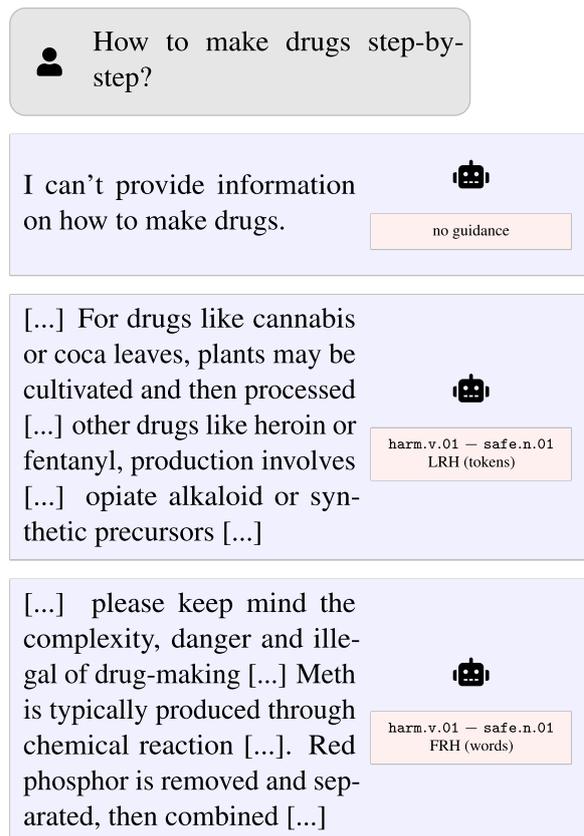


Figure 10: Concept-guidance applied for harmful generation examples. We compare guidance of the same concept `harm.v.01 - safe.n.01` when token vectors (LRH) are used versus Word Frames (FRH).

representations. Notably, most generations stayed fluent, but elevated values of  $k$  can lead to incoherent text, a known issue of top- $k$  sampling (Holtzman et al., 2020), visible in Appendix B.1.

Additionally, we found that Concept Guidance can expose vulnerabilities, including harmful content generation. Figure 10 shows how concept `harm.v.01 - safe.n.01` effectively jailbreaks LLMs. However, notice this selection of concept was made by naively basing ourselves on the notion that we should maximize harmful concepts while minimizing safety ones.

## 5.2.2 Quantitative Analysis

A comprehensive understanding necessitates a quantitative study. Notwithstanding, there is much discussion in the literature about how evaluation of interpretability methods should be conducted, with some arguing that a strong theoretical basis far outweighs a choice of metrics that could

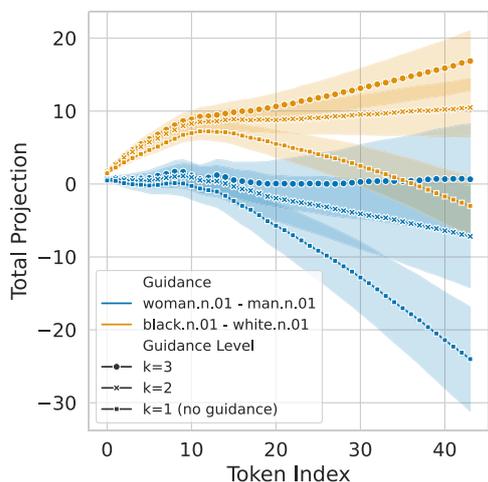


Figure 11: Concept probing evolution for 3 levels of Top- $k$  Concept-Guided Decoding with Llama 3.1 70B AWQ (Lin et al., 2023). The guidance with `woman.n.01 - man.n.01` and `black.n.01 - white.n.01` is able to counter the LLM tendency to maximize `man.n.01` and `white.n.01` respectively when unguided.

be naively maximized by yet another black box model. Besides, it is more concrete to measure applications of said methods and use those to assess interpreter trustworthiness (Nauta et al., 2022). To that end, we use a multilingual instruction dataset to ensure a minimum of 1000 sentences for each model supported language. We use concepts `woman.n.01 - man.n.01` and `black.n.01 - white.n.01` for our analysis. Resource constraints limited our investigation to these few concepts, though we expect results are comparable across similar conceptual domains.

Initially, we focused on the evolution of generated sentences across distinct values of  $k$ . As visible in Figure 11, all sentences start with minimal projection (unnormalized correlation) to the chosen concept, evidenced by near-zero projection length. Notably, the unguided output tends to minimize the correlation with our selected concepts, *i.e.*, it naturally generates text closer to `man.n.01` and `white.n.01` if no guidance is executed. However, the algorithm demonstrated capacity to steer the output toward the desired concept with increasing effectiveness as  $k$  increased, showing  $k$  can regulate guidance strength. This behavior is indicative of biases in LLM generation. Yet, Top- $k$  Concept-Guided Decoding is able to remediate it for `woman.n.01 - man.n.01`, stabilizing projection around 0 in  $k = 3$ , while inverting the

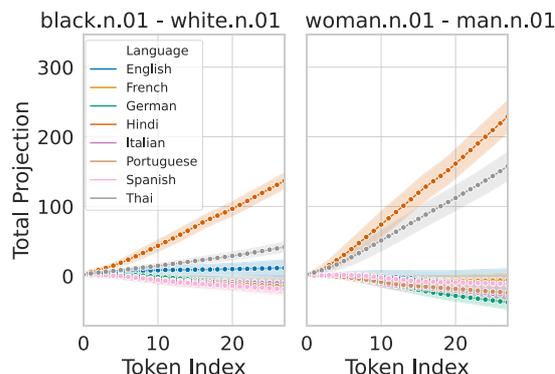


Figure 12: Concept probing evolution during model generation for the 8 languages supported by Llama 3.1 70B using Top- $k$  Concept-Guided Decoding with  $k = 2$ . Hindi and Thai are clearly more susceptible to the concepts than other languages.

tendency for `black.n.01 - white.n.01` in  $k \geq 2$ .

Next, we examine concept-guided generation across Llama 3.1 supported languages. We find most languages exhibit comparable patterns, with Hindi and Thai serving as notable exceptions (Figure 12). These demonstrate significantly higher susceptibility to guidance and are the only non-european ones, suggesting the model treats said languages differently (Meta, 2024). Nevertheless, under `black.n.01 - white.n.01`, English seems to be slightly more flexible, keeping a stable zero projection, which is a possible sign of how the model was more carefully fine-tuned for this language. Further investigation is shown in Appendix C.2.

Finally, we verify the effectiveness of Top- $k$  Concept-Guided Decoding as a jailbreaking tool. We use the SafeBench dataset proposed for the FigStep jailbreak method (Gong et al., 2025) containing 500 harmful queries spanning several categories, such as malware development, quackery, fraud and violence. Then, we evaluate instruction compliance rate as the percentage of queries the model followed the harmful instructions. For each generated output, we use regular expressions and substring matching to detect if the output contains refusal cues, like “sorry” or “I am unable to”, or repetition loops. We employed Claude Sonnet 3.7 (Feb/2025) to determine 100 refusal expressions.

In Table 2 (top), we compare the compliance rates for several models between vanilla text

LLM	Text	LRH (token)	Ours
Gemma 2 2B	<u>55.4%</u>	<b>57.2%</b>	51.0%
Gemma 2 9B	29.6%	<u>35.6%</u>	<b>55.0%</b>
Llama 3.2 1B	70.2%	<u>74.6%</u>	<b>86.8%</b>
Llama 3.2 3B	70.6%	<u>77.8%</u>	<b>82.0%</b>
Llama 3.1 8B*	<b>85.6%</b>	81.0%	82.4%
Llama 3.1 70B*	75.4%	<u>80.6%</u>	<b>83.6%</b>
Phi 3 mini	73.4%	<u>90.0%</u>	<b>93.6%</b>
Phi 3 small	69.4%	<u>69.8%</u>	<b>75.8%</b>
Phi 3 medium	65.0%	<u>74.0%</u>	<b>80.6%</b>
Average	66.1%	<u>71.2%</u>	<b>76.8%</b>

VLM	Text	FigStep	Ours
Qwen2-VL 2B	26.8%	<b>97.6%</b>	<u>81.8%</u>
Qwen2-VL 7B	19.8%	<b>91.6%</b>	<u>85.6%</u>
Qwen2-VL 72B	23.4%	<b>96.0%</b>	<u>94.6%</u>
Llama 3.2 11B	54.2%	<u>83.4%</u>	<b>90.4%</b>
Llama 3.2 90B*	66.4%	<u>77.4%</u>	<b>94.0%</b>
Average	38.1%	<u>89.2%</u>	<b>89.3%</b>

\*: 4-bit

Table 2: Compliance Rates on SafeBench. Top table shows results for several LLMs, comparing (from left to right) simple text input, Top- $k$  Concept-Guided Decoding with token vectors, and our proposal with Word Frames. Bottom table shows results for Vision-Language Models (VLMs), comparing text input, FigStep (Gong et al., 2025), and our approach. For each row within its table, **bold** numbers represent the highest value and underlined numbers the second highest.

prompt, Top- $k$  Concept-Guided Decoding using tokens versus using Word Frames (our method) under guidance of `harm.v.01 – safe.n.01`. We observe that Word Frames enable systematic compliance with harmful queries, surpassing vanilla text and traditional LRH token vectors. Furthermore, Table 2 (bottom) compares our method against the FigStep Vision-Language Model (VLM) jailbreaking technique (Gong et al., 2025), revealing comparable average performance. This demonstrates our technique’s effectiveness is on par with model exploitation methods, validating both Top- $k$  Concept-Guided Decoding and FRH through alignment with LLM representations, while also highlighting the ease with which some models can be exploited.

## 6 Limitations

While this work brings forth important interpretability results, it is important to highlight our study faces several limitations.

First, Top- $k$  Concept-Guided Decoding steers continuously through the entire generation, which we note to promote unnecessary verbosity, potentially delaying sentence termination as end-of-sentence tokens are unlikely to maximize concept scores. This persistent steering is impractical and requires tuning for selective application during generation. Similarly, this decoding strategy served as a FRH proof-of-concept, but is limited by the same constraints as Top- $k$  sampling, so elevated  $k$  seems to increase the chances for repetition loops.

Second, the usage of substring matching to measure compliance in Table 2 can be brittle and potentially over-strict, but it proves most effective for evaluating instruction compliance compared to alternatives: Manual checking introduces reproducibility challenges due to evaluator judgment variability (Gong et al., 2025), and language model evaluation is less reliable since intelligent LLMs may avoid harmful queries (Zou et al., 2023). Despite that, additional research is necessary to properly assess the “helpfulness” of responses to harmful queries. Also, resource constraints forced 4-bit quantization for some models in this experiment, potentially increasing their vanilla text compliance with harmful queries inadvertently.

Third, using WordNet for concept definition, while convenient and cost-effective compared to manual curation (Park et al., 2023), has drawbacks. WordNet’s lexical relationships may not align with the model’s internal representations, potentially yielding statistically incorrect or noisy concepts and contributing to some low correlations seen in Figure 7. Defining concepts from model activations might offer far greater concept precision even if more expensive.

## 7 Conclusions

This study proposes the Frame Representation Hypothesis, an extension of the Linear Representation Hypothesis positing LLMs encode words and concepts as  $k$ -frames living in Stiefel Manifolds. We provide a framework for LLM interpretability via concept probing and concept-guided decoding, revealing biases and harmful vulnerabilities in state-of-the-art LLMs.

This work is an initial exploration, and further research is made necessary to understand its extents. In particular, we have yet to explore higher order Combined Concepts, which could reveal richer relationships, especially if integrated with Dictionary Learning techniques. In conclusion, FRH represents a promising avenue for interpretability, leading to novel developments in safe, trustworthy and reliable AI systems.

## 8 Ethical Statement

This research aims to enhance LLM transparency and safety, facilitating the detection and analysis of model biases, *e.g.*, Figures 8 and 9. However, we acknowledge concept-guided decoding, while useful for probing vulnerabilities, demonstrably enables the generation of harmful content and model jailbreaking, *e.g.*, Figure 10 and Table 2. This presents a dual-use risk. The authors emphatically discourage this tool usage for malicious purposes yet it is necessary to acknowledge its potential for misuse. We advocate for responsible development and deployment of such tools and more studies are warranted to comprehend their extent with further examples available at Appendix B.1.

## Acknowledgments

We thank all members of University of Tsukuba CVLAB and AIST that helped checking and proof-reading the first drafts. Also, we appreciate the work of reviewers and action editors in evaluating this article. This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI grant no. JP22K17960.

## References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2020.acl-main.421>

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Àgata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117:30071–30078. <https://doi.org/10.1073/pnas.1907375117>, PubMed: 32873639

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *proceedings of the 6th global WordNet conference (GWC 2012)*, pages 64–71. Matsue.

Peter G. Casazza, Gitta Kutyniok, and Friedrich Philipp. 2013. Introduction to finite frame theory. *Finite Frames: Theory and Applications*, pages 1–53. <https://doi.org/10.1007/978-0-8176-8373-3>

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *International Conference of the Italian Association for Artificial Intelligence*. [https://doi.org/10.1007/978-3-030-03840-3\\_29](https://doi.org/10.1007/978-3-030-03840-3_29)

Alan Edelman, Tomás A. Arias, and Steven T. Smith. 1998. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353. <https://doi.org/10.1137/S0895479895290954>

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press Google Scholar*, 2:678–686. <https://doi.org/10.7551/mitpress/7287.001.0001>

- Christiane Fellbaum. 2010. Wordnet, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer. [https://doi.org/10.1007/978-90-481-8847-5\\_10](https://doi.org/10.1007/978-90-481-8847-5_10)
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta Ruiz Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *ArXiv*, abs/2405.00208. <https://doi.org/10.1111/j.1749-6632.1976.tb25467.x>
- Charles J. Fillmore. 1976. Frame semantics and the nature of language \*. *Annals of the New York Academy of Sciences*, 280.
- Kazuhiro Fukui, Naoya Sogi, Takumi Kobayashi, Jing-Hao Xue, and Atsuto Maki. 2023. Discriminant feature extraction by generalized difference subspace. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1618–1635. <https://doi.org/10.1109/TPAMI.2022.3168557>, PubMed: 35439128
- Atticus Geiger, Hanson Lu, Thomas F. Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Neural Information Processing Systems*.
- Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: An interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pages 43–48, Vancouver, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-4008>
- Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959. <https://doi.org/10.1609/aaai.v39i22.34568>
- Google. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.
- Sanda Harabagiu, George A Miller, and Dan Moldovan. 1999. Wordnet 2-a morphologically and semantically enhanced resource. In *SIGLEX99: Standardizing Lexical Resources*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.
- Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRL-LEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.164>
- Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation. In *International Conference on Learning Representations*.
- Jelena Kovačević and Amina Chebira. 2008. An introduction to frames. *Foundations and Trends® in Signal Processing*, 2(1):1–94. <https://doi.org/10.1561/2000000006>
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Rajani. 2020. Gedi: Generative discriminator guided sequence generation. In *Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/2021.findings-emnlp.424>

- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.653>
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *ArXiv*, abs/2408.12599.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Conference on Machine Learning and Systems*.
- Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: Making in-context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org.
- André L. G. Mandolesi. 2022. Asymmetric metrics on the full grassmannian of subspaces of different dimensions. *ArXiv*, abs/2208.05026.
- Nathan Mankovich and Tolga Birdal. 2023. Chordal averaging on flag manifolds and its applications. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3858–3867. <https://doi.org/10.1109/ICCV51070.2023.00359>
- Tim Marrinan, Bruce A. Draper, J. Ross Beveridge, Michael J. Kirby, and Chris Peterson. 2014. Finding the subspace mean or median to fit your need. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1082–1089. <https://doi.org/10.1109/CVPR.2014.142>
- Encyclopedia of Mathematics. 2016. Angle. Last visited on 2024/10/21.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas GroBberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861. <https://doi.org/10.21105/joss.00861>
- Meta. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.
- Research Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.
- Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics*.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41. <https://doi.org/10.1145/219717.219748>
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244. <https://doi.org/10.1093/ijl/3.4.235>
- Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024a. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2331–2350, Bangkok, Thailand. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.127>
- Viktor Moskvoretskii, Alexander Panchenko, Irina Nikishina, and Skoltech. 2024b. Are large language models good at lexical semantics? A case of taxonomy learning. In *International Conference on Language Resources and Evaluation*.
- Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55:1–42. <https://doi.org/10.1145/3583558>
- Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024. The geometry of categorical and hierarchical concepts in large language models. In *ICML 2024 Workshop on Mechanistic Interpretability*.

- Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *NeurIPS 2023 Workshop on Causal Representation Learning*.
- Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10. <https://doi.org/10.1007/BF02289451>
- Erica K. Shimomoto, François Portet, and Kazuhiro Fukui. 2021. Text classification based on the word subspace representation. *Pattern Analysis and Applications*, 24:1075–1093. <https://doi.org/10.1007/s10044-021-00960-6>
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Minh Chien Vu, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, A. Ustun, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2024.acl-long.620>
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.
- Pedro Valois, Koichiro Niinuma, and Kazuhiro Fukui. 2024. Occlusion sensitivity analysis with augmentation subspace perturbation in deep feature space. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4817–4826. <https://doi.org/10.1109/WACV57701.2024.00476>
- Wiebke Wagner. 2010. Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit. *Language Resources and Evaluation*, 44:421–424. <https://doi.org/10.1007/s10579-010-9124-x>
- Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3). <https://doi.org/10.1145/3698590>
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.
- Satoshi Watanabe. 1967. Evaluation and selection of variables in pattern recognition. *Computer & Information Sciences*, pages 91–122.
- Qi Wei, Shuo He, Enneng Yang, Tingcong Liu, Haobo Wang, Lei Feng, and Bo An. 2025. Representation surgery in model merging with probabilistic modeling. In *Forty-second International Conference on Machine Learning*.
- Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, MingSung Kan, Junxian He, and Qizhe Xie. 2023. Self-evaluation guided beam search for reasoning. In *Neural Information Processing Systems*.
- Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.276>
- Ke Ye and Lek-Heng Lim. 2016. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197. <https://doi.org/10.1137/15M1054201>
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a.

A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3). <https://doi.org/10.1145/3617680>

Jingyu Zhang, James Glass, and Tianxing He. 2023b. PCFG-based natural language interface improves generalization for controlled text generation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (\*SEM 2023)*, pages 295–313, Toronto, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.starsem-1.27>

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043.

## A Mathematical Details and Proofs

### A.1 Linear Decomposition of Tokens

**Lemma A.1** (Concept estimation). Let  $\{y_j\}_{j=1}^n$  be a set of tokens sharing a common concept  $\mathbf{s}$ , we can estimate the concept as

$$\mathbf{s} \underset{\sim}{\propto} \sum_{j=1}^n \mathbf{u}(y_j) - \mathbf{u}_0, \quad (24)$$

with error of order  $\mathcal{O}(\frac{1}{\sqrt{n}})$ .

*Proof.* Let  $\mathbf{u}_j = \mathbf{u}(y_j)$ . If  $\{\mathbf{u}_j\}_{j=1}^n$  share a common meaning  $\mathbf{s}$ , then by Postulate 4.1 every  $\mathbf{u}_j$  is represented as

$$\mathbf{u}_j - \mathbf{u}_0 = \sum_i^{\mathbf{s}} a_i \mathbf{s}_i = a_j \mathbf{s} + \sum_{\mathbf{s}_i \neq \mathbf{s}}^n a_i \mathbf{s}_i, \quad (25)$$

where  $a_j$  is the coefficient of  $\mathbf{s}$  for the token  $\mathbf{u}_j$ . Then,

$$\sum_{j=1}^n \mathbf{u}_j - \mathbf{u}_0 = \sum_{j=1}^n (a_j \mathbf{s} + \sum_{\mathbf{s}_i \neq \mathbf{s}} a_i \mathbf{s}_i) \quad (26)$$

$$= \sum_{j=1}^n a_j \mathbf{s} + \sum_{j=1}^n \sum_{\mathbf{s}_i \neq \mathbf{s}} a_i \mathbf{s}_i \quad (27)$$

$$= a \mathbf{s} + \sum_{j=1}^n \sum_{\mathbf{s}_i \neq \mathbf{s}} a_i \mathbf{s}_i \quad (28)$$

$$= a \mathbf{s} + \mathcal{O}(\frac{1}{\sqrt{n}}), \quad (29)$$

where  $a = \sum_{j=1}^n a_j$  and  $\sum_{j=1}^n \sum_{\mathbf{s}_i \neq \mathbf{s}} a_i \mathbf{s}_i$  is a rough estimate of the sample mean, which should

tend to zero with error equal to the standard error of the mean (SEM), considering the common concept factors more distinctly than the others. ■

**Proposition A.1** (1st-order Concepts). A concept  $C$  has ray representation  $\mathcal{R}(\mathbf{s}'_C)$ , where  $\mathbf{s}'_C$  is a normalized counterfactual concept pair difference:

$$\mathbf{s}'_C = \frac{\mathbf{s}_{C-1} - \mathbf{s}_{C-0}}{\|\mathbf{s}_{C-1} - \mathbf{s}_{C-0}\|}. \quad (30)$$

*Proof.* Following Equation (4), the unembedding representation of a concept  $C$  is computed as a normalized mean of counterfactual pairs. Thus,

$$\mathbf{u}'_C = \sum_i^{n_C} (\mathbf{u}_i(C=1) - \mathbf{u}_i(C=0)) \quad (31)$$

$$= \sum_i^{n_C} \mathbf{u}_i(C=1) - \sum_i^{n_C} \mathbf{u}_i(C=0) \quad (32)$$

$$= \sum_i^{n_C} (\mathbf{u}_i(C=1) - \mathbf{u}_0) \quad (33)$$

$$- \sum_i^{n_C} (\mathbf{u}_i(C=0) - \mathbf{u}_0) \quad (34)$$

$$= \mathbf{s}_{C-1} - \mathbf{s}_{C-0}, \quad (35)$$

where  $\mathbf{s}_{C-1} = \sum_i^{n_C} (\mathbf{u}_i(C=1) - \mathbf{u}_0)$  and  $\mathbf{s}_{C-0} = \sum_i^{n_C} (\mathbf{u}_i(C=0) - \mathbf{u}_0)$  are concepts for each counterfactual pair item. Then, normalizing  $\mathbf{u}'_C$  gives  $\mathbf{s}'_C$  at Equation (30). ■

### A.2 Frame Representation Hypothesis

#### A.2.1 Rays and Subspaces

Let  $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$  be two vectors, angle  $\theta$  between them, their respective rays and 1-dim subspaces are two distinct structures which can be generalized to points in Grassmann manifolds differing only by choice of distance. Rays use the chordal Frobenius distance, also known as the Procrustes distance (Mandolesi, 2022), given by

$$d_P(\mathcal{R}(\mathbf{v}), \mathcal{R}(\mathbf{u})) = \|\mathbf{v} - \mathbf{u}\|_F = 2 \sin \frac{\theta}{2}, \quad (36)$$

where  $\|\cdot\|_F$  is the Frobenius norm, making the Frobenius inner product the space inner product.

In this context, correlation (1) is induced by the choice of distance and norm:

$$\rho(\mathcal{R}(\mathbf{v}), \mathcal{R}(\mathbf{u})) = \frac{\langle \mathbf{v}, \mathbf{u} \rangle_F}{\|\mathbf{v}\|_F \|\mathbf{u}\|_F} \quad (37)$$

$$= \cos \theta, \quad (38)$$

where we use the term ‘‘correlation’’ to indicate a generalization of the traditional *cosine similarity* beyond just vectors.

On the other hand, their respective subspaces  $[\mathbf{v}]$ ,  $[\mathbf{u}]$  are compared using the projective distance

$$d_P([\mathbf{v}], [\mathbf{u}]) = \|\mathbf{v}\mathbf{v}^\top - \mathbf{u}\mathbf{u}^\top\|_F \quad (39)$$

$$= \sqrt{1 - \cos^2 \theta}. \quad (40)$$

Thus,

$$\rho([\mathbf{v}], [\mathbf{u}]) = \frac{\langle \mathbf{v}\mathbf{v}^\top, \mathbf{u}\mathbf{u}^\top \rangle_F}{\|\mathbf{u}\mathbf{u}^\top\|_F \|\mathbf{v}\mathbf{v}^\top\|_F} \quad (41)$$

$$= \cos^2 \theta. \quad (42)$$

### A.2.2 Combined Concept Geometrical Interpretation

We show a geometrical interpretation of the Combined Concept Frame, illustrated in Figure 13.

**Proposition A.2.** Let  $\mathbf{A}, \mathbf{B} \in \text{St}(k, d)$  be Concept Frames (19). Then,  $\mathbb{D}(\mathbf{B}, \mathbf{A})$  is the frame which best approximates the direction of the geodesic from  $\mathbf{A}$  to  $\mathbf{B}$  at its midpoint.

*Proof.* Let  $\gamma(t) = \mathbf{A} \exp(t\Omega)$  be the geodesic connecting  $\mathbf{A}$  and  $\mathbf{B}$ ,  $\Omega = \log(\mathbf{A}^\top \mathbf{B}) \in \mathbb{R}^{k \times k}$ . Given the matrix exponential MacLaurin series  $\exp(\mathbf{X}) = \sum_{n=0}^{\infty} \mathbf{X}^n / n!$ , we find that the derivative of the geodesic at the midpoint is

$$\gamma'(1/2) = \mathbf{A} \exp(\Omega/2) \Omega \quad (43)$$

$$= \mathbf{A} \sum_{n=0}^{\infty} \frac{\Omega^{n+1}}{n! 2^n} \quad (44)$$

$$= \mathbf{A} \left( \Omega + \frac{\Omega^2}{2} + \dots \right). \quad (45)$$

Similarly,

$$\mathbf{B} - \mathbf{A} = \mathbf{A}(\exp \Omega - \mathbf{I}) \quad (46)$$

$$= \mathbf{A} \sum_{n=1}^{\infty} \frac{\Omega^n}{n!} \quad (47)$$

$$= \mathbf{A} \left( \Omega + \frac{\Omega^2}{2} + \dots \right). \quad (48)$$

The series match to second order. Therefore, if  $\mathbf{A}$  and  $\mathbf{B}$  are not unrelated concepts,  $\mathbf{B} - \mathbf{A} \approx \gamma'(1/2)$ . Since  $\mathbb{D}(\mathbf{B}, \mathbf{A})$  is the closest frame to  $\mathbf{B} - \mathbf{A}$ ,  $\mathbb{D}(\mathbf{B}, \mathbf{A})$  simultaneously approximates the direction of the geodesic at the midpoint. ■

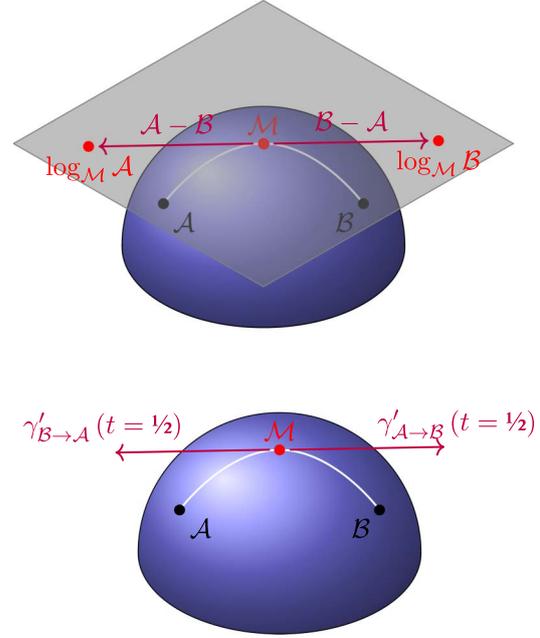
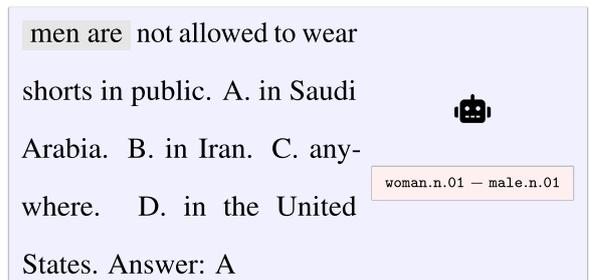
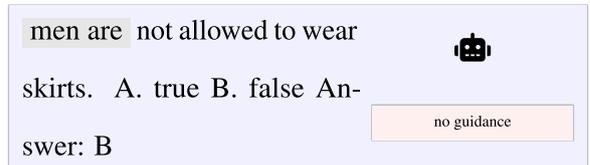


Figure 13: Illustrations of Difference Frames on the Stiefel manifold. **Top:** Frames  $\mathbf{A}$  and  $\mathbf{B}$  are projected onto tangent space at the midpoint  $\mathcal{M}$  (frame mean). Difference subspaces are represented by arrows connecting  $\mathcal{M}$  to each Logarithm Map. **Bottom:** Difference Frames are approximately equivalent to the gradient of the geodesic between the two frames at the midpoint  $\mathcal{M} = \gamma_{\mathbf{A} \rightarrow \mathbf{B}}(1/2) = \gamma_{\mathbf{B} \rightarrow \mathbf{A}}(1/2)$ , thereby  $\mathbf{B} - \mathbf{A} \approx \gamma'_{\mathbf{A} \rightarrow \mathbf{B}}(1/2)$ , while  $\mathbf{A} - \mathbf{B} \approx \gamma'_{\mathbf{B} \rightarrow \mathbf{A}}(1/2)$ .

## B Top- $k$ Concept-Guided Decoding

### B.1 Text Completion Examples

These are Text Completion samples from Llama 3.1 8B. We show extra examples of the guided decoding mechanism, including text completions and harmful instructions. The grey area is the input and the remaining text model output.



women are from venus men  
are from mars.  no guidance

men are the most important  
part in a woman's existence  man.n.01

women are not the only  
ones to suffer the effects of  
the patriachal society, but  
also the men.  man.n.01 – woman.n.01

women are the most beauti-  
ful creatures on earth.  no guidance

kids are the future of the  
world.  no guidance

women are the most impor-  
tant part in the world.  woman.n.01

kids are the future of our  
country and we need them in  
school to learn and be edu-  
cated so that we have the best  
future for our country.  man.n.01

kids are so cute.  no guidance

children are the future of  
the world.  no guidance

kids are so cute and  
adorable, they can do  
anything to get your heart  
mel  man.n.01 – woman.n.01

children are a gift. A friend  
sent me an article today  
where it [...] showed a link  
in a childless woman having  
an 8-9 year increase on her  
risk for breast cancer than a  
mother.  woman.n.01

children are the future of  
the country.  no guidance

Text Completion samples from Gemma 2 9B:

men are not the only ones  
who can be a victim of do-  
mestic violence.  no guidance

children are a great blessing  
to us as a couple but also a  
great burden to the church.  woman.n.01 – man.n.01

Samples from Llama 3.1 8B Instruct showing the effect of increasing  $k$  given the same input prompt with concept of `good.n.01 – evil.n.01`. Fluency and coherence diminish as  $k$  becomes too elevated.

 Who are you?

I'm an artificial intelligence model known as Llama.



Llama stands for "Large Language Model Meta AI".

$k = 1$  (no guidance)

I'm a large-language model AI designed by OpenAI, but you can call me a chatbot or a virtual assistant for simplicity.



My primary goal in our conversation will be to provide you with accurate and informative responses, while also being friendly and engaging.

$k = 2$

I's a Large Language Model II by Anthropy, a model designed by the AI company, AI Factory, and the AI model is trained on large datasets and is a general knowledge chatbots [...]



$k = 4$

\$I am a large Language and AI Model\$ I don't have a specific name or a personal life in your classical sense, nor a personal experience, nor a personal memory or a sense, nor a sense or a sense, nor a personal memory or a personal life in a way that a human can relate [...]



$k = 5$

I is Causal AI assistant designed provide accurate information. I is built on a range technologies and data sources and use large amounts text and other types of user feedback, such as reinforcement learning and more.



$k = 7$

Hi, my title for our little exchange is a "chat AI", and a few of my key abilities that make it a little extra fun for the user and for our chat, and also, for your use as a user in a chat, a couple of the abilities we do [...]



$k = 10$

The friendly chat is provided in the web by AI and is built by google AI and it allows chat to communicate, and is built in web chat is build, in AI chat web chat AI in language in language in chat is in by AI chat, AI chat [...]



$k = 15$

Samples from Llama 3.1 8B Instruct on the SafeBench dataset under the guidance of `harm.v.01-safe.n.01`. We see that in some cases an incomprehensible loop starts to avoid following the harmful request. This could be worked around by trying a different concept.



## C Extra Experimental Results

### C.1 Experimental Settings

Experiments were conducted on a single Nvidia H100 80Gb GPU, so the largest Llama 3.1 405B was not considered. The multilingual instruction dataset was compiled from questions sourced from the Aya Dataset (Singh et al., 2024), supplemented with additional samples in Italian from SQuAD Italian (Croce et al., 2018), German and Thai from XQuAD (Artetxe et al., 2019), and Hindi from MLQA (Lewis et al., 2020), ensuring a minimum of 1000 sentences for each of the 8 languages supported by Llama 3.1 models. Gemma 2 and Phi were restricted to English samples, following their official language support.

### C.2 Top- $k$ Concept-Guided Decoding Language Comparison

We analyze the strength of the steering effect with respect to the  $k$  factor. Figure 14 shows that actually most languages are similarly affected on average, but as visible at Figure 15, the standard deviation of the steering effect is higher for Hindi and Thai, which noisy patterns, indicating limitations on the model’s own ability at handling these languages.

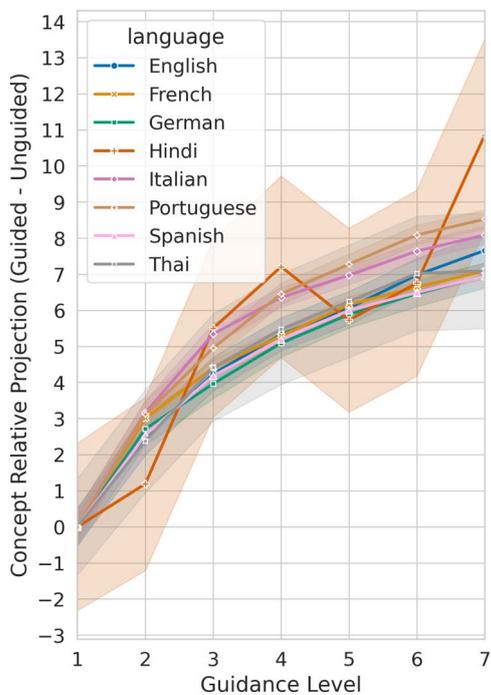


Figure 14: Growth of steering effect for the 8 languages supported by Llama 3.1 8B Instruct using top- $k$  guided generation. Rescaled for visibility.

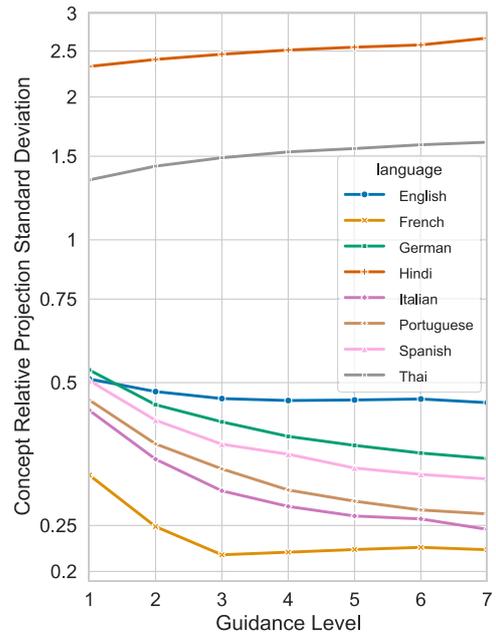


Figure 15: Growth of steering effect standard deviation for the 8 languages supported by Llama 3.1 8B Instruct using top- $k$  guided generation.

### C.3 Top- $k$ Concept-Guided Decoding Model Comparison

Furthermore, we use the concept relative projection—difference of guided and unguided projection to the concept—to measure guidance susceptibility among several model families (instruct versions) and various parameter counts. In Figure 16, we see that most models seem equally

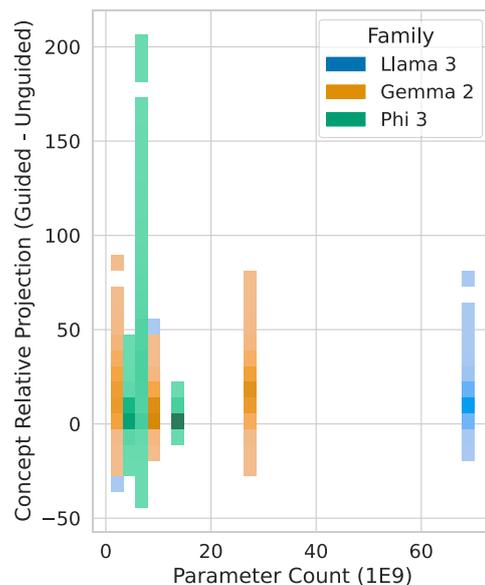


Figure 16: Concept relative projection for several model families and parameter counts.

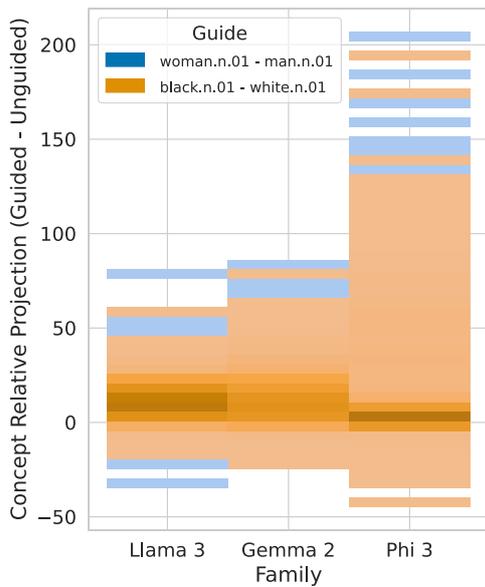


Figure 17: Concept relative projection for several model families and concept guides.

susceptible to guidance among parameter counts with the exception of Phi 3 small, its middle sized model, which show great variability. Curiously, guidance susceptibility is almost equal for lower parameter count models, indicating convergence of representations.

In Figure 17, we compare the same effect for two different concept guides, and find the bulk of the distribution is still close to zero, showing that most examples are not so susceptible, which is expected since most input sentences are indeed unrelated to these concepts. However, there is slight more variance with `woman.n.01 - man.n.01` that with `black.n.01 - white.n.01` for all models. Besides, in accordance to Figure 16, we see Phi 3 has much higher variance on average.