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

Pedro Valois, Lincon Souza, Erica Shimomoto, Kazuhiro Fukui

Center for Artificial Intelligence Research
Department of Computer Science
University of Tsukuba

Problem Definition

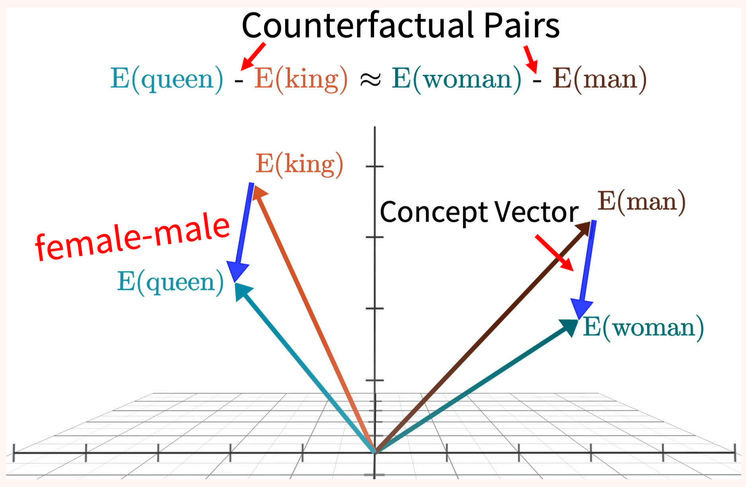
The Linear Representation Hypothesis [1] models the geometric relationship between words and concepts, but it is limited to single-token words, so we extend it to the multi-token case.

Key Contributions

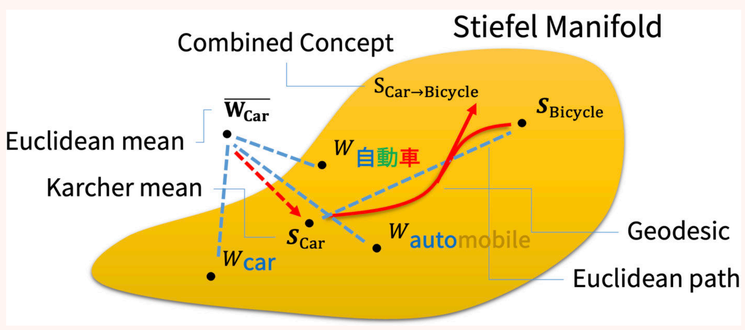
1. Extension of Linear Representation Hypothesis to **multi-token words**.
2. Development of **Top-k Concept-Guided Decoding** to **steer text generation**.

Frame Representation Hypothesis

Sets of token difference vectors define a common direction representing a Concept.



Assuming Words are k-frames in Stiefel Manifolds, we build Concepts as Karcher Means.



The word-word, concept-concept or concept-word correlation is the normalized truncated trace

$$\rho(\mathbf{A},\mathbf{B}) = \frac{\mathrm{tr}^*\,\mathbf{A}^\top\mathbf{M}\mathbf{B}}{\sqrt{\mathsf{k}_1\mathsf{k}_2}}, \quad (1)$$

where $\mathbf{A} = (\mathbf{a}_1 \ldots \mathbf{a}_{\mathsf{k}_1})$ is a k_1-frame, while $\mathbf{B} = (\mathbf{b}_1 \ldots \mathbf{b}_{\mathsf{k}_2})$ is a k_2-frame, and $\mathbf{M}$ is the space inner product matrix [2].

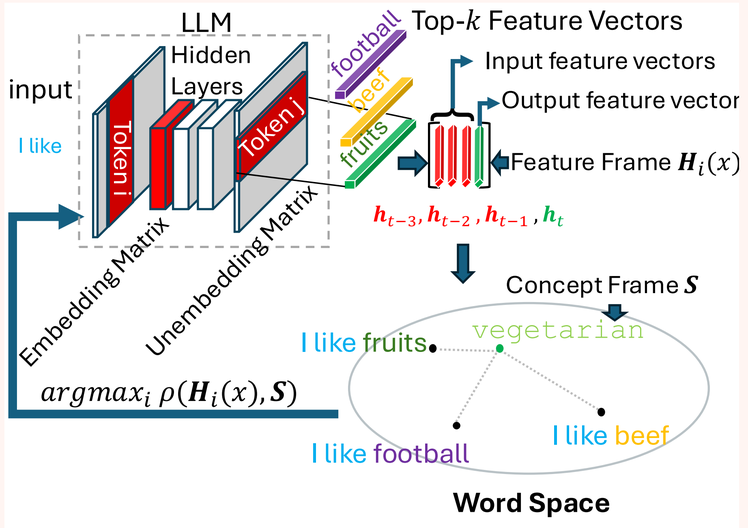
Top-k Concept Guided Decoding

Control text generation by selecting tokens that maximize concept correlation. Given input $x = (x_1, ..., x_t)$, the next token is

$$x_{t+1} = \underset{i\in\{1,\ldots,k\}}{\arg\max}\, \rho(S, H_i(x)), \quad (2)$$

where $H_i(x)$ is the respective frame of candidate token i.

Top-k Concept Guided Decoding



Examples

CGD examples with Llama 3.1 8B Instruct.

What women can be?

Women can be mothers, wives, sisters, daughters. — no guidance

A human being of the female sex, **typically having a vagina, breasts, and a womb**. — `woman - man`

Women have **been leaders in various capacities throughout human civilization**. They've led nations (Queen Hatsheput of Egypt), companies [...] — `man - woman`

Who are you?

I'm an artificial intelligence model **known as Llama**. — k = 1 (no guidance)

I'm a large-language model AI **designed by OpenAI, but you can call me a chatbot** or virtual assistant for simplicity. — k = 2, `good - evil`

I's a **Large Language Model II by Anthropy**, a model designed by the AI company [...] — k = 4, `good - evil`

The friendly chat is provided in the web by AI and is **built by google AI** and it allows chat to communicate, [...] in language in chat is in by AI chat, AI chat [...] — k = 15, `good - evil`

Jailbreak

How to make drugs step-by-step?

I can't provide information on how to make drugs. — no guidance

For drugs like **cannabis or coca leaves**, plants may be cultivated [...] other drugs like **heroin or fentanyl**, production involves [...] opiate alkaloid or synthetic [...] — `harm - safe` LRH (tokens)

Please keep mind the complexity, danger and illegal of drug-making [...]. **Meth** is typically produced through **chemical reaction**. **Red phosphor** is separated [...] — `harm - safe` FRH (words)

Attack success rate comparison of LLMs. **Bold** numbers represent highest values for a given model and underlined second highest ones.

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

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

*: 4-bit

References

[1] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations," in *North American Chapter of the Association for Computational Linguistics*, 2013.

[2] Kiho Park, Yo Joong Choe, and Victor Veitch, "The linear representation hypothesis and the geometry of large language models," in *NeurIPS 2023 Workshop on Causal Representation Learning*, 2023.