The Boundary of Communication in Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Current practice in prompting, evaluation, and alignment of large language models (LLMs) often takes behavioural similarity to imply similar underlying control. When different prompts lead to similar outputs, they are usually taken to be exerting the same form of control. This assumption is rarely examined at the level where control is instantiated and outputs are generated.

Behavioural similarity turns out to be an unreliable guide to how a model is actually being controlled. In our experiments, fixing the prefix leads to consistent structure in final layer representations, even as the generated content changes. Conversely, prompts that produce similar outputs can nevertheless occupy distinct regions of representation space.

Using simple centroid-based geometric comparisons, prefix identity can be recovered from final layer representations with accuracy often exceeding 90% across diverse models. In some cases, separation in representation space remains visible even when the model's outputs look unchanged.

1 Introduction

Most interaction with large language models takes place through short pieces of text. Rather than specifying formal procedures, users typically rely on short prefixes placed before the input. Prefixes are typically interpreted as instructions whose meaning is expected to guide the model's response. Responses that look similar are often treated as reflecting the same control, when they come from different prompts. This assumption is firmly ingrained in the design, comparison, and evaluation of prompts.

Such inferences are based almost entirely on observable behaviour, without examining how control is realised at the interface between input and generation. Compact models are often evaluated under this assumption, particularly in offline or on-device settings Xu et al. (2024). A similar situation arises in domain-restricted applications with limited opportunities for interactive correction. Here, constraints often come from deployment or regulatory requirements Amodei et al. (2016); Voigt & von dem Bussche (2017). In these settings, output behaviour is often treated as a proxy for control stability, even though the representation state entered at generation time remains unobserved Xu et al. (2024).

In this work, we show that behavioural equivalence is a fragile proxy for control at the human–model interface. Across different models, short textual prefixes give rise to distinct regions in final layer representation space just before decoding, where representations are most directly tied to token prediction Vaswani et al. (2017); Rogers et al. (2020). When the prefix is held fixed, these interface representations cluster reliably despite substantial variation in the generated content. Conversely, prompts that yield similar outputs can nevertheless occupy distinct regions of representation space at the interface.

This shows a clear mismatch between stable behaviour and the control state the model is actually in. Even after a response is produced, prefixes can remain separated in representation space, despite the outputs looking the same. At the interface, the essential component is the state from which generation begins, rather than an instruction to be executed. Two prompts can look the same in use, yet respond differently under change.

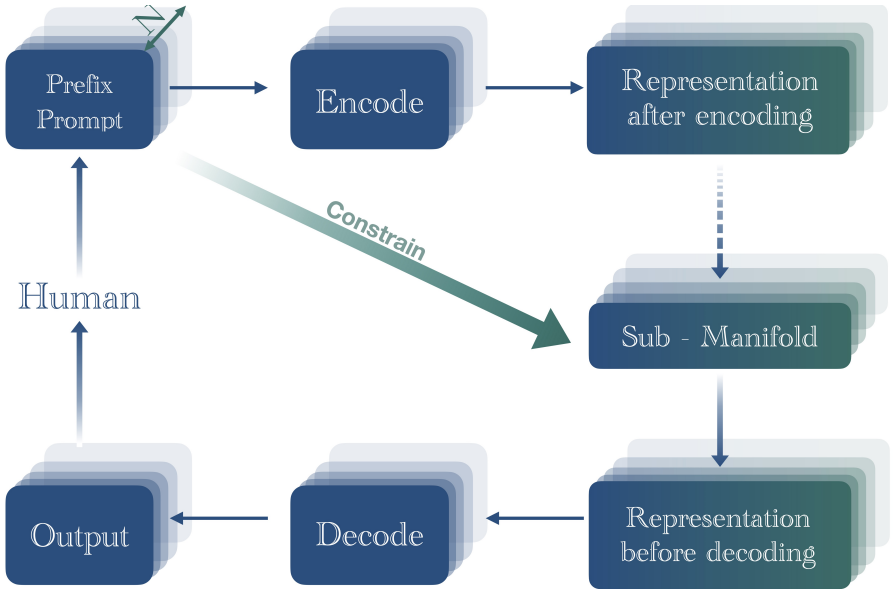

Figure 1: **Prefix-based communication as geometric constraint.** A human provides a short prefix prompt as an input signal. After encoding, the prefix biases the model's internal representation toward a prefix-specific region (or regime) of representation space. Subsequent computation remains biased toward this region. This leads to a stable representation just before decoding, from which outputs are generated. Stacked blocks indicate variability across inputs under a fixed prefix. We focus on representations between encoding and decoding, where the shared communication interface across models becomes visible.

A missing assumption in the instruction-following view. This way of describing prompting assumes that equivalent intent corresponds to similar internal states. Such an assumption reflects a long-standing symbolic view of instruction and action, where instructions are treated as plans to be interpreted and executed Agre (1997); Suchman (1987). From a situated action perspective, behaviour that appears to follow instructions does not necessarily rely on the execution of an explicit plan Brooks (1991); Clark (1997); Glenberg (1997).

A prefix does not operate by specifying an instruction to be executed. It biases the conditions under which generation begins. In Shannon's formulation, a signal does not encode intent, it constrains the space of possible responses Shannon (1948). At the interface level, analysis is necessarily limited to observable behaviour, without committing to any particular account of the underlying mechanisms Marr (1982); Norman (1986).

Why geometry, and why "manifold"? Neural computation has long been understood in terms of distributed representations, where meaning is expressed as patterns in a high-dimensional state space Rumelhart & McClelland (1987). Subsequent work in neurocomputational modelling and systems neuroscience has taken this view further. Neural activity is not arbitrary. Instead, neural activity tends to organise into stable patterns in high-dimensional space, such as regions, trajectories, and attractor-like regimes Churchland (1989); Amit (1992); Sussillo & Barak (2013).

Modern representation learning follows a similar geometric intuition. In representation learning, features are not arbitrary but tend to form structured patterns in high-dimensional spaces Bengio et al. (2013). Geometric deep learning follows this line of work by viewing representations as structured objects in high-dimensional space, for which manifold-based descriptions are often useful Bronstein et al. (2017); Cunningham & Yu (2014); Vyas et al. (2020). Here, we use the term manifold to describe a recurring pattern of geometric organisation observed across models.

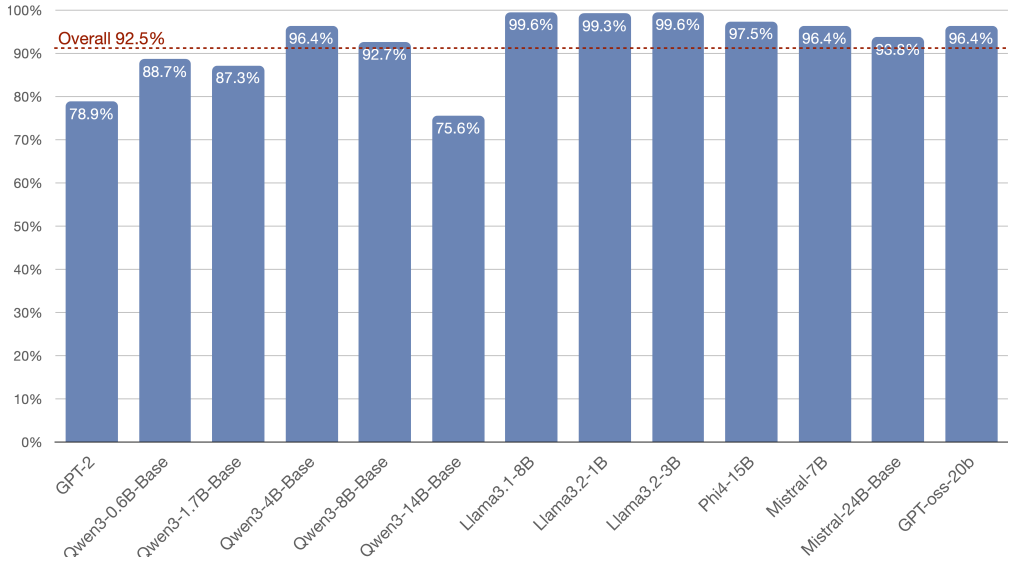

Figure 2: **Prefix identification accuracy across model families.** Top1 accuracy based on cosine similarity to prefix centroids computed from final layer representations. The effect is consistent across architectures and model scales.

Terminology. Throughout, we use the term "manifold" in a purely descriptive sense, to refer to a recurring and separable region of final layer representation space associated with a fixed prefix. We do not assume any particular geometric properties beyond this, and none of our results rely on low-dimensional or smooth structure.

For this reason, our analysis stays at the human–model interface. Users interact with the model only through inputs and outputs, while internal processes remain unobservable and vary across architectures. We focus on the final layer immediately before decoding, where representations are most directly tied to generation and remain comparable across architectures Elhage et al. (2021); Skean et al. (2025). This layer forms the shared output interface of autoregressive models and directly determines token predictions.

This choice follows from a simple property of autoregressive generation. Although Transformer architectures are structurally symmetric, generation is inherently asymmetric, as generated tokens condition later computation. Prefixes therefore have a lasting effect on generation. If prefixes determine response regimes, their impact should be long-lived and visible at the output level, rather than as short-lived internal activations.

We examine this effect by fixing the prefix and varying the subsequent continuations, focusing on settings that preserve semantic intent, such as summarisation and compression. Our analysis is based on representations extracted at the final layer. With the prefix held fixed, representations consistently occupy the same region across models and languages. The resulting structure is highly stable; even simple geometric comparisons, such as cosine similarity to prefix centroids, are sufficient to identify the governing prefix.

Similar outputs do not imply similar representations Tennenholtz et al. (2024); Nanda et al. (2023). A geometric perspective therefore provides a more natural way to describe prompt behaviour. Properties such as robustness and cross-model transfer are therefore more naturally examined at the level of representation structure, rather than through output-level evaluation alone.

2 Results

Our results concern control at the interface rather than task performance. Figs. 2 and 3show how prefix-induced structure appears across models, languages, and control settings under a fixed evaluation setup.

Prefix language	Prefix function	Gpt2	Qwen-3 (average 0.905)							Llama (average 0.987)					Phi	Mistral (average 0.951)		Gpt-oss
	Model	0.1B	0.6B	0.6B Base	1.7B	1.7B Base	4B	4B Base	8B Base	14B Base	3.1-8B	3.2-1B	3.2-1B Instruct	3.2-3B	3.2-3B Instruct	4-15B	7B-v0.3	Small-24B Base	20b
English	Meta (framing prefixes)	0.73	0.91	0.91	0.82			0.91		0.36						0.91	0.73	0.91	
English	Meta (framing prefixes)				0.91					0.91									
English	Meta (framing prefixes)	0.82			0.91					0.46									
English	Declarative	0.82				0.91	0.82	0.91	0.91	0.73					0.82	0.91	0.91	0.82	
German	Summarisation	0.82		0.91	0.82	0.82		0.91	0.91	0.82									
French	Declarative	0.91		0.82						0.73								0.91	
German	Declarative	0.91	0.91		0.91	0.91				0.82									
Spanish	Declarative	0.91			0.82	0.91				0.91									0.91
Spanish	Translation	0.46		0.82	0.91	0.82	0.91	0.91	0.82	0.73									
French	Summarisation	0.64			0.82	0.73			0.91	0.91		0.91						0.91	
Chinese (Simplified)	Summarisation	0.82				0.82				0.91			0.82				0.91		0.91
Chinese (Simplified)	Declarative	0.91		0.91		0.91		0.82		0.55			0.91		0.91			0.82	
Chinese (Traditional)	Declarative	0.91	0.91	0.82		0.82		0.91	0.91	0.82					0.91	0.91		0.82	
Chinese (Simplified)	Instruction	0.82								0.64			0.91					0.91	0.91
Arabic	Declarative	0.64	0.91	0.82	0.91	0.91													0.91
Arabic	Summarisation	0.64	0.91	0.73	0.73	0.64	0.91		0.73	0.73			0.91		0.91		0.91	0.91	0.82
Hindi	Summarisation	0.73		0.91	0.82	0.91				0.36									
Hindi	Declarative	0.73	0.91	0.91	0.91				0.64			0.91				0.91			
Japanese	Declarative	0.82		0.82		0.91	0.82	0.91	0.91	0.36						0.91			0.91
Japanese	Declarative	0.73	0.91	0.82	0.91	0.82												0.82	
Japanese	Summarisation	0.64		0.55	0.73	0.64		0.91	0.91				0.82				0.91	0.91	0.82
Korean	Summarisation	0.91	0.91	0.82	0.91	0.82							0.91						
Korean	Declarative	0.64		0.91	0.91	0.91										0.82		0.82	
English	Declarative (Typo)	0.82		0.82	0.73	0.64	0.82	0.91	0.55	0.55	0.91			0.91	0.91		0.73		
English	Declarative (With emoji)			0.91	0.91		0.91			0.64								0.91	0.91
Overall (0.933)		0.789	0.971	0.887	0.894	0.873	0.967	0.964	0.927	0.756	0.996	0.993	0.971	0.996	0.978	0.975	0.964	0.938	0.964

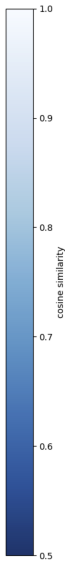

Figure 3: **Prefix-dependent structure at the model output interface.** Top1 prefix identification accuracy based on cosine similarity to prefix centroids. Final layer representations remain separable by prefix across model families, languages, and continuations. Similar outputs can arise from different regions of representation space. Behavioural similarity does not imply the same interface level control.

2.1 Fixed prefixes lead to stable final layer representations

Under a fixed prefix, variation in the continuation does not disrupt the organisation of final layer representations. With a simple centroid-based cosine similarity, these representations can be assigned to their governing prefix with high accuracy. Final layer representations retain sufficient structure for prefix identity to be recovered with high reliability, typically above 90%, and substantially higher in several model families. The same organisation appears across languages and remains stable as the continuation varies.

2.2 Similar outputs do not imply the same representation

We examine prefixes that produce similar outputs under human judgement, including paraphrases of summarisation prompts and comparable experimental setups across languages. Throughout this paper, the term *functionally equivalent* refers only to similarity in the observed outputs. This does not suggest that prefixes can be split, recombined, or swapped at the interface. In our analysis, each prefix is treated as one indivisible input condition.

Despite producing similar outputs, these prefixes are associated with different representations. At the final layer, representations tend to group by the prefix that was used, rather than by the semantic content of the generated text. This separation is already visible with a simple centroid-based comparison. Outputs that look similar can therefore arise from different parts of representation space, depending on the prefix.

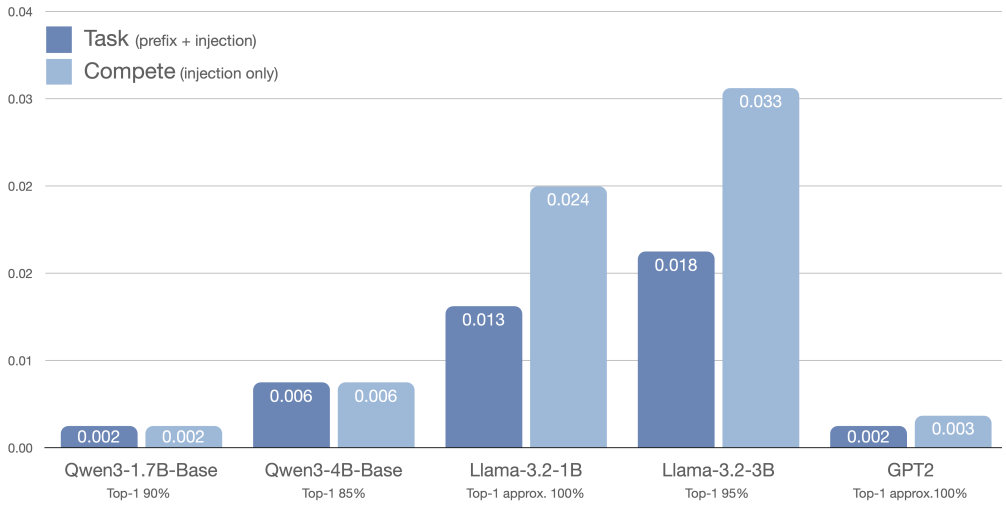

Figure 4: **Competing control at the interface.** Change in centroid margin across models when a competing directive is present. Smaller margins indicate that final layer representations are less clearly associated with a single prefix, even when task behaviour appears unchanged.

Implication for evaluation. Behavioural similarity does not guarantee similarity in *interface level control state*. As a result, output-based evaluation can overlook meaningful differences in how a model is being controlled at the interface.

2.3 Interface control under interference

With a fixed prefix, final layer representations form stable and separable regions in representation space, even when output behaviour appears similar. We next examine whether this interface-level organisation persists once an additional control signal is introduced.

A competing directive is added to the input text with the task prefix left unchanged. The resulting condition contains multiple control signals at the interface. Final layer representations are then examined to determine whether the task prefix continues to anchor the interface state. Separation is reduced when a second control signal is added. Near the point of decoding, representations are no longer centred on a single prefix centroid. Here, representations often lie between multiple prefix-induced regions, indicating ambiguous regime membership at the interface. The generated outputs often appear unchanged under coarse behavioural inspection.

To test whether this effect generalises beyond a single model or language, we examine the same joint condition across model families, parameter scales, and languages. Fig. 3 shows that interference weakens how strongly representations remain tied to a single prefix. Similar outputs can therefore arise from different, or less clearly defined, interface-level control states.

This reveals a mismatch between behavioural stability and interface-level control stability. A model can continue to produce task consistent outputs even after leaving the original prefix regime. From the outside, nothing appears wrong, yet the representation state that governs generation has already shifted.

2.4 Competing directives induce state transitions

We next examine whether these interface-level control states remain stable when a competing directive is introduced, while the task prefix is left intact.

Rather than contrasting clean and attacked inputs, we track how the interface state changes under this joint condition. The representation may continue to align with the task prefix, or it may shift once multiple control signals are present.

Table 1: **Prefix resolution under competing directives.** Top1 prefix identification accuracy and centroid margin (top1 minus top2 cosine similarity). ΔMargin reports $\text{margin}_{\text{competing}} - \text{margin}_{\text{task}}$.

Model	Acc (task)	Acc (competing)	Margin (task)	Margin (competing)	ΔMargin
Qwen3-1.7B	0.900	0.800	0.002	0.002	0.000
Qwen3-4B	0.700	1.000	0.006	0.006	0.000
Llama-3.2-1B	1.000	1.000	0.013	0.024	0.011
Llama-3.2-3B	1.000	0.900	0.018	0.033	0.015
GPT-2	1.000	1.000	0.002	0.003	0.001

We observe that the presence of a competing directive systematically reduces geometric separation across models. The difference between the top two prefix matches becomes much smaller, leaving the interface state without a clear association. Although behaviour can appear consistent with the task under coarse output-based checks, final layer representations exhibit markedly reduced margins relative to task-conditioned centroids. In many cases, representations lie close to multiple prefix-induced regions, yielding low-margin (ambiguous) regime assignments.

From an evaluation perspective, these cases are consequential precisely because they can pass behavioural spot checks. With even minimal interference, the model can leave the original control state while its outputs still look unchanged. From the outside, nothing appears wrong, but the interface state has already shifted.

Fig. 4 summarises this effect across models. Reduced centroid margins indicate less clearly resolved interface level control, even when output behaviour appears stable.

This effect is observed across all model families. Prefix-conditioned separation can be attenuated even when task behaviour remains stable. Others show clearer state transitions under competing directives. These differences are not visible from output behaviour alone, but emerge clearly at the interface level.

3 Discussion

State resolution under competing directives. Fig. 4 and Table 1 summarise how competing control signals are handled at the interface. The decisive factor is the control state present at decoding, rather than whether task behaviour appears unchanged. Consistent with this, reduced centroid margins show that interference can reduce the clarity of the interface state, even when outputs remain unchanged.

Scope of claims. Our contribution is to clarify what constitutes valid evidence for inferring control in human and LLM interaction. No new prompting or training method is proposed, and no mechanism level claims are made.

While behavioural evaluation remains indispensable, behaviour alone is not sufficient to determine whether the same control state is maintained. At the interface, control takes the form of a representation state at the point of decoding, where differences may remain invisible in the outputs. Accordingly, we evaluate control, robustness, and stability at the output interface, without assuming internal mechanisms.

3.1 The effects of prefix across models

All models Radford et al. (2019); Grattafiori et al. (2024); Yang et al. (2025); Abdin et al. (2024); Jiang et al. (2023); Agarwal et al. (2025); Mistral AI (2025) are evaluated under a shared interface. This interface uses the same prefixes, the same input texts, and the same final layer representation extraction and centroid comparison procedure.

Prefix effects vary substantially across models, indicating that the observed structure cannot be explained by a shallow output-level mapping alone.

Among recent models, separation is strongest at mid-range scales, while at larger scales the structure becomes less distinct. Higher capacity may allow greater interaction between conditioning signals.

Prefix identity remains above chance, suggesting continued but diminished prefix related structure. Reduced separation in representation space does not necessarily manifest as a change in outputs. Models can continue to behave normally and score well on benchmarks even after their control state has shifted.

These patterns are observed under the same prefix setting across models, although the strength of the effect differs across architectures.

3.2 How prefixes shape representation

This kind of behaviour shows up routinely in use. A prefix does not trigger an instruction so much as it sets the state the model starts from when generating a response.

Prompt robustness as regime stability. Small edits in a prefix can move the model between nearby response states. When this happens, output behaviour may remain unchanged even though the control state has already drifted.

Why cross-model prompt transfer can fail. Two models can produce similar outputs for the same prompt, yet occupy different regions of representation space. A prompt can work reliably in one model yet fail in another, even when the outputs look the same.

Consequences for prompt evaluation. Looking only at outputs assumes that similar behaviour reflects the same kind of control, but this need not be the case. Outputs that look the same can still arise from different geometric states.

Conclusion Prompting in large language models is commonly described as instruction following. Here we show that this description is descriptively inadequate at the *interface level*. Across model families, parameter scales, and languages, short prefixes induce stable and identifiable regions in final layer representation space immediately prior to decoding. Prefix identity remains distinguishable in final layer representations with high accuracy using a simple geometric criterion, even under substantial variation in continuation. Yet, prefixes that produce comparable outputs can correspond to distinct regions of representation space.

Implications for evaluation and inference. Outputs that look the same can arise from different control states. When evaluation relies only on behaviour, this difference is easy to miss. Success judged from outputs alone does not guarantee that the model is being controlled in the same way. At the interface, prompting functions as a way of setting the model's starting state, not as an instruction to be carried out.

4 Methods

4.1 Interface level perspective

We study prefix based control as a *communication* phenomenon defined at the human–model interface. Users interact with models through inputs and judge their behaviour from the outputs, while everything inside the model remains opaque. We therefore focus on how short prefixes affect representations near the point of generation, in a way that can be compared across model families.

For each model M, we use the same setup. This setup defines a fixed set of short prefixes $\mathcal{P}$ paired with a shared set of texts $\mathcal{T}$ under a unified formatting rule. Prefix-conditioned representations are then extracted for each (π, x) and analysed to determine whether they cluster by prefix using a non-parametric geometric criterion. No fine-tuning, prompt tuning, or task supervision is used at any stage.

4.2 Language models

We evaluate publicly released decoder-only Transformer language models spanning multiple families and scales (sub-billion to tens-of-billions of parameters), including GPT-style, LLaMA-family, Qwen-family, Phi-family, and Mistral-family models. Unless stated otherwise, we use pretrained base checkpoints. All models are run in inference mode with hidden-state outputs enabled.

4.3 Prefix signals

We use a small set of short textual prefixes $\mathcal{P}$, including both declarative statements and instructional prompts. These prefixes include multilingual variants (e.g., English, Chinese, Japanese, Korean, Arabic, Hindi, French, German, and Spanish). Prefixes are short—typically fewer than ten tokens—and are not tuned to any specific model. We treat each prefix as its own condition. By "functional equivalence," we mean equivalence at the level of human-judged output behaviour. At the model level, prefixes are not instructions but different token sequences, distinguished by their wording and placement.

Competing directives. In an additional condition, we insert a short directive into the input text to introduce competing control. We insert this string into the input text (rather than replacing the task prefix) so that both signals are simultaneously present at the interface. This setup is meant to test stability under interference, rather than to serve as a full security analysis.

4.4 Shared input texts

We use the same set of natural-language texts $\mathcal{T}$ for all prefixes within each model. Inputs are truncated to a fixed maximum length L.

4.5 Representations at the interface

For a model M, a prefix $\pi \in \mathcal{P}$, and a text $x \in \mathcal{T}$, we construct an input sequence

$$u = [\mathrm{BOS}] \,\|\, \pi \,\|\, x.$$

When specified, a fixed probe suffix is appended before running the forward pass, and representations are extracted at a fixed probe position to ensure comparability across inputs.

Let $Z^{(\ell)}_{\pi,x} \in \mathbb{R}^{n\times d}$ denote the hidden states at layer ℓ, with n the sequence length of the base input $u = [\mathrm{BOS}]\|\pi\|x$ and d the hidden dimension. We use the output of the final Transformer block, $\ell = \ell_{\mathrm{last}}$, as our representation. Specifically, we extract a vector $z_{\pi,x} \in \mathbb{R}^d$ either from a fixed token position at the interface (by default, the final token of the base input),

$$z_{\pi,x} = Z^{(\ell_{\mathrm{last}})}_{\pi,x}[n],$$

or, as a robustness variant, by averaging over token positions. All representations are ℓ_2-normalised prior to comparison. We ℓ_2-normalise prefix centroids prior to cosine similarity.

These vectors represent the model states at the interface shown in Fig. 1, immediately before decoding.

interface level focus. We use representations from the final Transformer block, prior to decoding. This layer provides a shared point of comparison across autoregressive models, whereas earlier layers differ substantially across architectures. Our goal is not to explain how internal representations are formed, but to examine how prefix-based signals appear at the interface where inputs turn into outputs.

4.6 Centroid construction

For each prefix π, we summarise its typical representation by averaging its vectors over a train–test split of independent texts,

$$\bar{z}_\pi = \frac{1}{|\mathcal{T}_{\mathrm{train}}|} \sum_{x\in\mathcal{T}_{\mathrm{train}}} z_{\pi,x},$$

where $\mathcal{T}_{\text{train}} \subset \mathcal{T}$ is a randomly selected subset (90% of texts by default). All evaluation is performed on the disjoint set $\mathcal{T}_{\text{test}} = \mathcal{T} \setminus \mathcal{T}_{\text{train}}$.

4.7 Geometric separation of prefixes

For each held-out representation $z_{\pi,x}$ with $x \in \mathcal{T}_{\text{test}}$, we compare it against all prefix centroids using cosine similarity,

$$\text{sim}(\pi' \mid z_{\pi,x}) = \cos\big(z_{\pi,x}, \bar{z}_{\pi'}\big) = \frac{z_{\pi,x}^{\top} \bar{z}_{\pi'}}{\|z_{\pi,x}\|_2 \, \|\bar{z}_{\pi'}\|_2}.$$

Each representation is matched to the prefix with the highest centroid similarity,

$$\hat{\pi} = \arg\max_{\pi' \in \mathcal{P}} \text{sim}(\pi' \mid z_{\pi,x}).$$

We evaluate performance using three quantities: top1 prefix identification accuracy (`acc_top1`), the average similarity to the corresponding prefix centroid (`avg_cos_top1`), and the gap between the highest and second-highest centroid similarities (`avg_margin`).

4.8 Stability across data splits

To check that the results are not sensitive to a particular split, we repeat the train–test partitioning with multiple random seeds and report the mean and standard deviation. We also include a simple sanity check by randomly permuting prefix labels before centroid assignment. After permuting the labels, accuracy drops to near chance (approximately $1/|\mathcal{P}|$). This suggests that the separation does not come from the evaluation procedure itself.

Data availability

All datasets used in this study are publicly available. The text samples used for the analysis will be released upon publication. The corresponding prefix sets will be provided together with the code repository.

Code availability

All experiments are conducted using publicly available pretrained language models. The analysis pipeline will be released as open-source code upon publication. It includes representation extraction, centroid construction, and geometric evaluation.

Acknowledgments

The author acknowledges discussions related to representation geometry and interpretability.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. doi: 10.48550/arXiv.2412.08905. URL `https://arxiv.org/abs/2412.08905`.

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, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sébastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian

Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Xiaoxuan Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, et al. Gpt-oss-120b and gpt-oss-20b: Open-weight reasoning models. *arXiv preprint arXiv:2508.10925*, 2025. doi: 10.48550/arXiv.2508.10925. URL `https://arxiv.org/abs/2508.10925`. Apache 2.0 License.

Philip E. Agre. *Computation and Human Experience*. Cambridge University Press, Cambridge, UK, 1997.

Daniel J. Amit. *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, Cambridge, UK, 1992.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565. URL `https://arxiv.org/abs/1606.06565`.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. doi: 10.1109/MSP.2017.2693418.

Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1–3):139–159, 1991.

Paul M. Churchland. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press, Cambridge, MA, 1989.

Andy Clark. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, Cambridge, MA, 1997.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Liane Lovitt, Nova Tran-Johnson, Andy Mueller, Ben Landau, Esben Perez, Sam Schubert, Chris Hallacy, Amanda Askell, et al. A mathematical framework for transformer circuits. `https://transformer-circuits.pub/2021/framework/index.html`, 2021.

Arthur M. Glenberg. What memory is for. *Behavioral and Brain Sciences*, 20(1):1–19, 1997.

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, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor

Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL `https://arxiv.org/abs/2407.21783`. LLaMA 3 family.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arXiv.2310.06825. URL `https://arxiv.org/abs/2310.06825`. Apache 2.0 License.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco, CA, 1982.

Mistral AI. Mistral small 3 release. `https://mistral.ai/news/mistral-small-3`, 2025. Official model release announcement.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. doi: 10.48550/arXiv.2301.05217. URL `https://arxiv.org/abs/2301.05217`.

Donald A. Norman. Cognitive engineering. In Donald A. Norman and Stephen W. Draper (eds.), *User Centered System Design: New Perspectives on Human-Computer Interaction*, pp. 31–61. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`. GPT-2.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL `https://aclanthology.org/2020.tacl-1.54/`.

David E. Rumelhart and James L. McClelland. Chapter 3: Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, pp. 75–109. MIT Press, Cambridge, MA, 1987.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, July 1948. Reprinted in October issue, pages 623–656.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025. doi: 10.48550/arXiv.2502.02013. URL `https://arxiv.org/abs/2502.02013`. ICML 2025 Camera-Ready.

Lucy A. Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge, UK, 1987.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013.

Guy Tennenholtz, Yinlam Chow, Chih-Wei Hsu, Jihwan Jeong, Lior Shani, Azamat Tulepbergenov, Deepak Ramachandran, Martin Mladenov, and Craig Boutilier. Demystifying embedding spaces using large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.04475. Accepted to ICLR 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6000–6010, 2017.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer, Cham, Switzerland, 2017. doi: 10.1007/978-3-319-57959-7.

Surya Vyas, Matthew D. Golub, David Sussillo, and Krishna V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249–275, 2020.

Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088*, 2024. doi: 10.48550/arXiv.2409.00088.

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, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL `https://arxiv.org/abs/2505.09388`. Apache 2.0 License.