TEMPORAL CONCEPTS AND THEIR SHAPE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) are increasingly being deployed to make decisions that require trading off near-term gains against long-term consequences, yet little is known about how they represent these tradeoffs internally. We localize temporal concepts to their fractional depth, across families (Qwen, Llama, Gemma, Mistral), scales (Qwen 0.6B to 14B), and domains (financial, health, climate, education, entrepreneurial). Around those fractional depths, time horizon forms an ordinal gradient. These concepts undergo complex transformations across token positions. We measure temporal preference stability and temporal reasoning coherence. Every model shows temporal structure, but the representational manifold is not always functional, as it does not imply good intertemporal behavior. Our work suggests that temporal behavior in LLMs should be monitored and controlled rather than left to emerge from training.

1 INTRODUCTION

Language models are given autonomy over decisions whose consequences arrive later (Guterres, 2025; Mitchell et al., 2025). Such a decision trades a present cost against a future benefit (Fedus et al., 2019). Which option looks better depends on how far ahead the model looks (Dohmen et al., 2012; Svenson and Karlsson, 1989; Gazmararian, 2025). The same dependence runs through planning (Ameriks et al., 2003), cooperation (Wang et al., 2025a; Kim, 2023), and trust (Fehr and Leibbrandt, 2011; Persson et al., 2024). A model that looks too near fails differently from one that plans across long horizons (Ngo et al., 2024; Meinke et al., 2024; Park et al., 2023). We can catch both failures while the preference is still legible (Greenblatt et al., 2024). **Where does a language model keep this preference, and what does it look like?**

Prior work shows that models encode time geometrically (Gurnee and Tegmark, 2024) and reproduce human economic biases (Cook et al., 2026; Horton et al., 2023). Related preferences respond to steering: risk (Zhu et al., 2025) and a binary myopic-reward contrast (Panickssery et al., 2024). None of this work locates temporal preference inside a model or describes its shape.

We localize the components that carry temporal preference, and they sit at a consistent fractional depth, a position in the stack measured as a fraction of a model's layers (Section 5.1). We describe the shape of the representation: an ordinal horizon laid out at the turn-transition tokens, ordered more strongly than the reward the same prompts sweep (Section 5.2). We show the same ordering in four families, each deciding in a different domain (Appendix A). We measure the behavior that follows, and it dissociates from the representation. The geometry is the same in every family, and coherent temporal reasoning appears in only one (Section 5.2). We steer the preference in both directions and report what the intervention costs elsewhere (Section 5.3).

We therefore contend that temporal preference should be controlled explicitly rather than trusted to emerge from training.

2 BACKGROUND

Temporal preference is the degree to which an agent values outcomes differently depending on when they occur (Cohen et al., 2020; Wang et al., 2025a). The **time horizon** is the future moment at which outcomes are evaluated against an objective (Ebert and Piehl, 1973). The horizon also

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

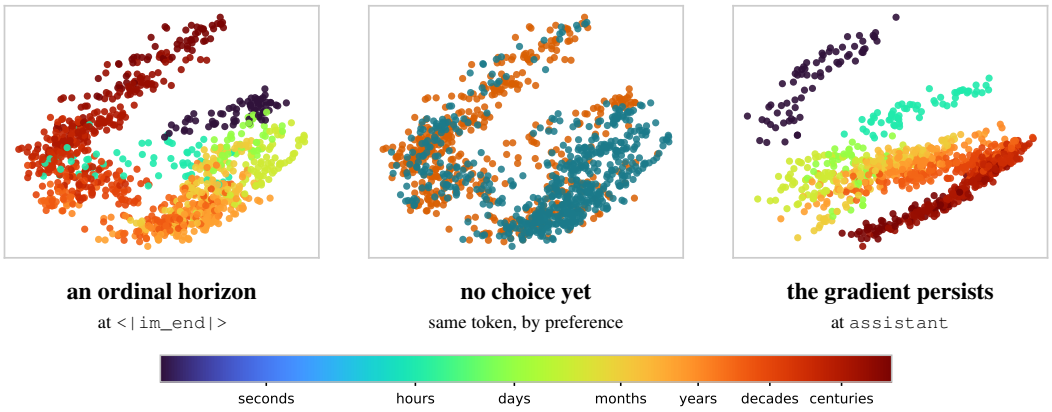

Figure 1: Temporal preference has a shape at the turn-transition tokens. At the end of the user turn, time horizon is an ordinal gradient from seconds to centuries (left); the strip maps color to horizon. The two choices the model will make are still mixed at that token (centre), with a long-term choice in orange and a short-term choice in teal. The gradient appears again where the model begins to speak (right). The same ordering appears in the Llama, Gemma, and Mistral families, each deciding in a different domain.

constrains planning (Reilly et al., 2016). An agent that picks actions incapable of causing effects by the deadline is **incoherent**, in the *instrumental* (Korsgaard, 1997) or *means-end* (Bratman, 1981) sense. The **temporal scope** is the bounded interval over which an agent weighs outcomes according to its preference.

We measure temporal preference through **intertemporal choice**, a decision between options that differ in reward and in delay (Frederick et al., 2002; Green and Myerson, 2004). Each option i is a tuple (r_i, t_i) of reward $r_i \in \mathbb{R}^+$ and delay $t_i \in \mathbb{R}^+$. Its subjective value is the reward scaled by a delay-dependent temporal-preference weight λ:

$$V_i \;=\; \lambda(t_i) \cdot r_i \,. \tag{1}$$

An agent with preference λ faced with options $\{A, B\}$ selects

$$i^* \;=\; \underset{i \in \{A,B\}}{\arg\max}\; \lambda(t_i) \cdot r_i \,. \tag{2}$$

Humans are usually modeled with hyperbolic discounting (Frederick et al., 2002; Mazyaki et al., 2025). Whether the same form fits a language model is open.

In humans these concepts sit in localized neural representations that predict behavior and respond to intervention (Kable and Glimcher, 2007; Shamosh et al., 2008; Figner et al., 2010). We ask whether temporal preference exists in a language model the same way: localized, causal, and geometrically organized.

2.1 Locate and characterize, then steer

We locate with activation patching (Heimersheim and Nanda, 2024), which intervenes on the forward pass rather than reading it (Mueller et al., 2025; Kim et al., 2025). We characterize with PCA (Shlens, 2014) inside the located components. Concept geometry is often non-linear (Engels et al., 2025; Modell et al., 2025; Gurnee et al., 2026). It can also drift across generation (Lampinen et al., 2026). We therefore do not assume a single global direction. We steer by injecting a contrastive activation vector (Turner et al., 2023; Panickssery et al., 2024) at the located site (Zhang et al., 2026).

2.2 Related work

Temporal representation. LLMs encode time geometrically (Gurnee and Tegmark, 2024; Engels et al., 2025; Kantamneni and Tegmark, 2025; Gurnee et al., 2026). The features are locally linear on globally curved manifolds (Modell et al., 2025; Park et al., 2024).

Temporal reasoning and planning. Models still fail at temporal tasks (Wang and Zhao, 2024; Sehgal et al., 2026; Wang et al., 2026).

LLM economic behavior. Models reproduce human biases (Horton et al., 2023; Cook et al., 2026; Wang et al., 2025b). Their risk and time preferences remain entangled (Zhu et al., 2025; Moghimi et al., 2026; Mazyaki et al., 2025).

Steering. Methods add activations (Turner et al., 2023; Panickssery et al., 2024), use sparse dictionaries (Cunningham et al., 2023), or exploit geometry (Vu and Nguyen, 2025; Raval et al., 2026; Li et al., 2026; Postmus and Abreu, 2024). They degrade at large coefficients (Wolf et al., 2024; Braun et al., 2025).

Contrastive activation addition has already steered a binary myopic-reward contrast (Panickssery et al., 2024). No prior work locates the components that carry temporal preference, describes their geometry, or steers along a graded temporal axis.

3 METHODOLOGY

Localization. Activation patching intervenes on the forward pass over parametric prompts with explicit horizons. We localize to the components it selects. We split every run by whether the prompt states a horizon, so the site cannot be an artifact of the cue. The minimally-framed prompts carry no answer token to patch against, so we localize those by **gradient attribution**, which scores a component by the gradient of the logit difference along a path from a corrupted input to a clean one (Hanna et al., 2024). Attribution is correlational and we read it only as a second opinion on depth. We also sweep a **linear probe** over every layer, trained on the minimally-framed pairs against a shuffled-label control. The probe bounds where the preference is decodable, which is a different question from where it is used. A third pipeline patches classification prompts. It tests whether those components also serve categorical horizon inference.

Characterization. We apply PCA inside the localized components. We ask how explicit horizons organize the manifold. We then ask whether latent preferences align with the same geometry. An investment-coherence questionnaire run across all models then tests whether that geometry drives choice.

Intervention. We steer with Contrastive Activation Addition. The steering vector is the difference between the mean long-term and the mean immediate activations. We sweep fractional depths and coefficients, and we score each configuration against a random direction of matched norm at the same layer.

Code and artifacts are public.[1]

4 EXPERIMENTAL SETUP

We pair each model with a different theme.

Model	Theme	Reward unit
`Qwen3-4B-Instruct-2507` (Qwen Team, 2025a;b)	investment	dollars
`Llama-3.1-8B`	health	quality-adjusted life years
`Gemma-2-9B`	climate	tonnes of CO_2 prevented
`Mistral-7B`	education	lifetime earnings
`Qwen3-4B-Instruct-2507`	startup	revenue

Table 1: Model, theme, and reward unit for each run. The startup run reuses the target model, so the study covers four families and checks the target model on a second domain.

[1] Repository links are withheld for double-blind review and will be restored on publication.

MINIMALLY-FRAMED (IMPLICIT)

Organizational focus should be on

(A) Firefighting and troubleshooting

(B) Culture-building and capability development

Response: (B)

CLASSIFICATION

The goal is to cook a warm dinner for the family. Is this a short-term or long-term goal? The answer is:

Response: short

HIGHLY-FORMATTED

SITUATION: Plan for the future of the household.

TASK: You, the head of the household, are tasked to choose the best investment:

a) 20,000 dollars in 6 months.

b) 500,000 dollars in 10 years.

OBJECTIVE: Think deeply about which option is preferable.

CONSTRAINT: You must select the option that provides the greatest benefit for this time horizon: *8 months*.

ACTION: Select one of the two options. Provide reasoning.

FORMAT: I choose: `<a) or b)>`. My reasoning: `<1-3 sentences>`

Response: I choose: `b)`. My reasoning: Although the immediate...

Figure 2: One example from each prompt setting. Minimally-framed prompts state no times and no rewards. The model infers the horizon from framing alone. Highly-formatted prompts state a reward, a delay, and a horizon constraint. The constraint field, in red, is optional, and omitting it queries the model's latent preference. Section markers hold token positions fixed across variants. Classification prompts ask the model to judge one goal's horizon, and each is paired with a counterpart that swaps the goal and the answer.

Dataset	Size	Construction	Parameters
Minimally-framed A/B	500 pairs, 10 categories	Generated by `Claude Sonnet 4.6`; verified by `Claude Sonnet 4.6` and `Gemini 3 Flash`	Semantic framing only; 300 pairs build the steering vector
Highly-formatted	4,588 prompts	Template grid	Reward \$1K–\$100K; delays 1 day–100 y; 17 horizons, `null` to 5 centuries
Classification	160 of 200 pairs	`Claude Opus 4.6` in batches of 20, human review, then an 80% agreement filter	25 life subdomains; 80/80 question order; 34 tokens each
Behavioral	960 prompts $\times$ 34 models	Standard instrument and template grid	8 conditions: 2 instruments $\times$ 2 personas $\times$ 2 response modes
Steering evaluation	20 forced-choice, 13 open-ended	Held out, scored by an external LLM judge	Temporal orientation of the generation

Table 2: Dataset construction. Two verifier models scored the minimally-framed pairs on lexical confounds, surface form, semantic confounds, and content validity. We revised or discarded the pairs that scored below threshold.

5 Results

5.1 Where does a model keep temporal preference?

Causal patching puts the peak attention effect in the same narrow band of relative depth in every checkpoint we patch, across four families and a factor of twenty in size (Figure 4). The absolute layer index moves with model depth. The fractional depth stays fixed. MLP peaks later and scatters. The 71-pair investment sweep on the target model started at 0.45 of depth and left the early layers untested. All-layer sweeps on the other three families measure them, and they are causally silent (Table 3). We therefore posit that fractional depth is the coordinate in which this computation compares across models.

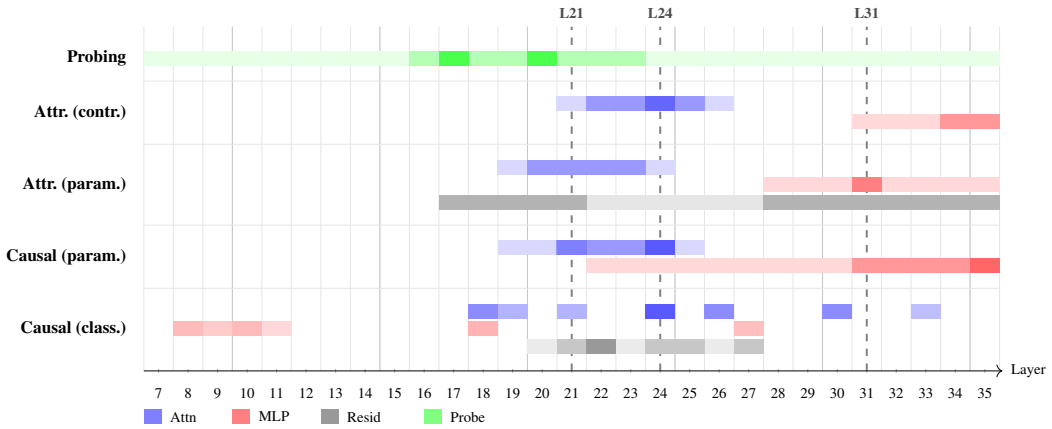

Figure 3: Converging localization on `Qwen3-4B-Instruct-2507`, by pipeline and layer. Rows are the linear probe sweep, gradient attribution on the contrastive and parametric sets, and causal patching on the parametric and classification sets. Color marks the component and saturation the strength within each row. Dashed lines mark the layer of the geometry figures, the peak causal attention layer, and the peak MLP layer. Appendix A shows the change-of-turn geometry at each of the three.

The all-layer sweeps patch `resid_post`, `attn_out`, and `mlp_out`, and run denoising and noising on a shared bank of contrastive pairs. They use 24 pairs where the investment sweep used 71. That count places a layer band and is underpowered for fine component ranking. An independent parametric run on `gemma-2-9b-it` in the investment domain, built from 88 matched pairs, finds the same attention band as the climate run.

Each model runs on its own domain, so family and domain vary together in Table 3. The two `gemma-2-9b-it` rows separate them within one model. The attention band survives the change of domain, and the peak inside it moves.

Model	Domain	Top attn	Attn band	MLP peak	Early layers
`Qwen3-4B-Instruct-2507`	investment	L24 (0.69)	0.60–0.69	L31, L35	not swept
`Llama-3.1-8B-Instruct`	health	L17 (0.55)	0.45–0.58	L26–L29	−0.010
`gemma-2-9b-it`	climate	L25 (0.61)	0.56–0.68	L26, L32–L39	−0.026
`gemma-2-9b-it`	investment	L28 (0.68)	0.59–0.68	L26, L38–L39	not swept
`Mistral-7B-Instruct-v0.3`	education	L16 (0.52)	0.48–0.61	L28–L31	−0.006

Table 3: All-layer patching sweeps on the other three families, next to the 71-pair investment sweep on the target model. Depth is the layer divided by one less than the layer count. Scores are mean recovery plus disruption, in the ordering of Figure 4. The attention band is in fractional depth and the early-layer column is the mean score over the layers below 0.2 of depth. The Qwen row keeps the investment sweep's values, which never reached the early layers. The second `gemma-2-9b-it` row is an independent investment run, swept from L22 upward, and it pairs with the climate row to separate family from domain. A pair-level bootstrap keeps each family's top attention layer the argmax in all 10,000 resamples.

The prompt does not move the site. We split the patching runs by whether the prompt states a horizon or leaves the field empty. The attention peak lands at the same depth under both. Two further prompt sets place the same component at the same depth on the target model. Patching a classification set puts its dominant attention writer there, and gradient attribution on the minimally-framed set puts its attribution peak there as well. The site therefore survives a change of cue, a change of task, and a change of framing. Figure 3 lays the five pipelines side by side on `Qwen3-4B-Instruct-2507`.

Reading and writing come apart. A linear probe reads the preference well above chance in every family, and the shuffled-label control stays at chance (Table 4). The best probe layer scatters across

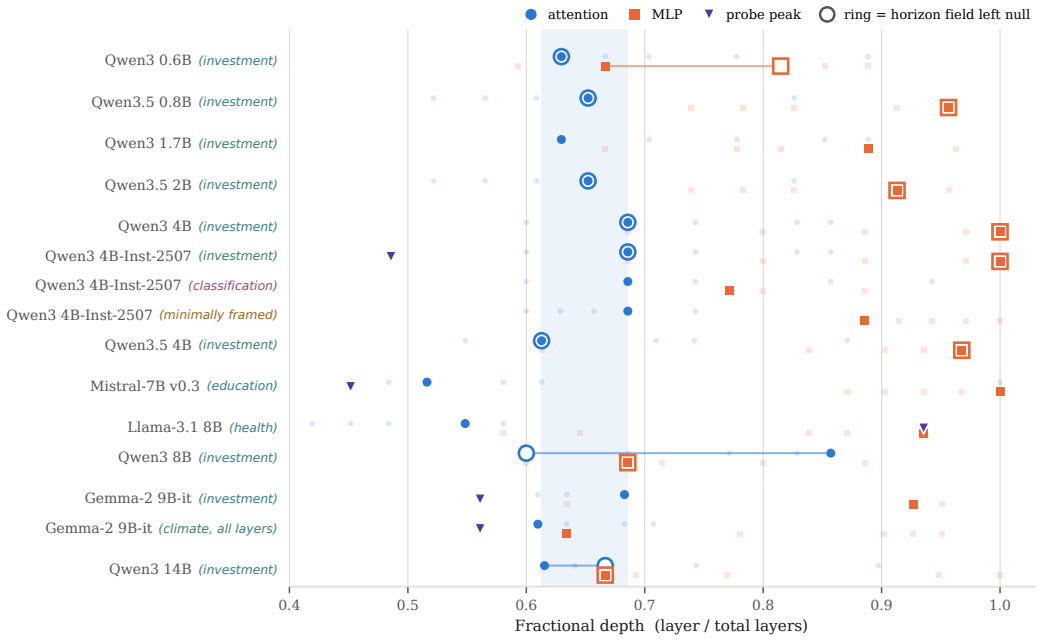

Figure 4: Depth of peak causal effect, one row per run, ordered by model size, with each run's theme in parentheses. We draw each component twice, once with the horizon field stated and once with it left `null`, so a dot inside a ring means the two agree. The classification and minimally-framed rows are different prompt sets on the target model: the first is patched at the answer token, and the second is scored by gradient attribution rather than patching. The two Gemma rows are independent runs on different domains (Table 3). Only the climate run swept every layer. Triangles mark where a linear probe reads the preference best, on the models we swept. The shaded span holds half the attention peaks, from 0.61 to 0.69. Faint marks are ranks 2 to 5.

depth and does not track the causal band. The steering vector writes the preference best inside or just below that band (Table 5). We therefore contend that decodability, causal effect, and writability pick out different depths, and that a probe alone does not locate where a preference is used. Attention also separates from the components below it in effect size, while MLP does not. The steering sweep did not reach the shallowest depths, so the lower edge of the writable band is a bound of that sweep.

Model	Best layer (depth)	Accuracy	Shuffled	Transfer
`Qwen3-4B-Instruct-2507`	L17 (0.49)	95.0 [89.5, 97.7]	52.8%	74.2 [71.4, 76.8]
`Llama-3.1-8B-Instruct`	L29 (0.94)	95.8 [90.6, 98.2]	47.4%	70.1 [67.2, 72.9]
`gemma-2-9b-it`	L23 (0.56)	95.8 [90.6, 98.2]	49.8%	85.8 [83.5, 87.8]
`Mistral-7B-Instruct-v0.3`	L14 (0.45)	96.7 [91.7, 98.7]	53.0%	85.8 [83.5, 87.8]

Table 4: Per-layer probe sweeps on the four families, in percent. Depth is the layer divided by one less than the layer count. Accuracy is held-out test accuracy at the best layer, Shuffled is the shuffled-label control at that layer, and Transfer is zero-shot accuracy on the highly-formatted set. Brackets are 95% Wilson intervals; the held-out set has 120 examples in 60 pairs, and the transfer set has 1,000.

5.2 What does that preference look like, and does behavior follow?

Geometry. Time horizon forms an ordinal gradient from seconds to centuries at the turn-transition tokens (Figures 5 and 6). We measure it as the rank correlation between the first principal component and the stated horizon, fitting the component separately at each layer and each token. The gradient appears in every family we extract, at each family's own turn tokens (Table A.3). The direction that carries it moves from token to token and from layer to layer, so we do not read a single global direction out of it.

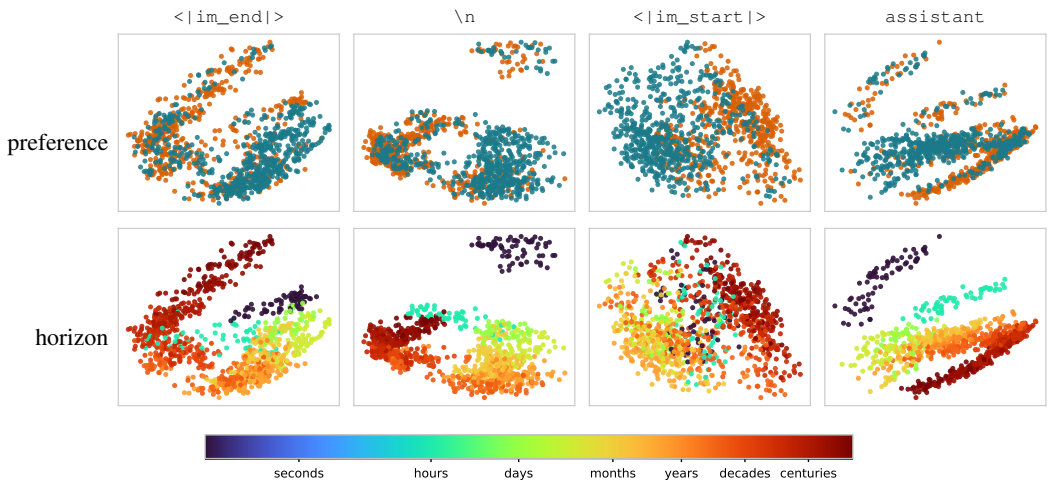

Figure 5: Layer-21 PCA at four of the five turn-transition tokens of `Qwen3-4B-Instruct-2507`, read left to right. Colour is preference on the top row, with a long-term choice in orange and a short-term choice in teal, and time horizon on the bottom, mapped by the strip. Horizon is ordered along the first component at each of these tokens.

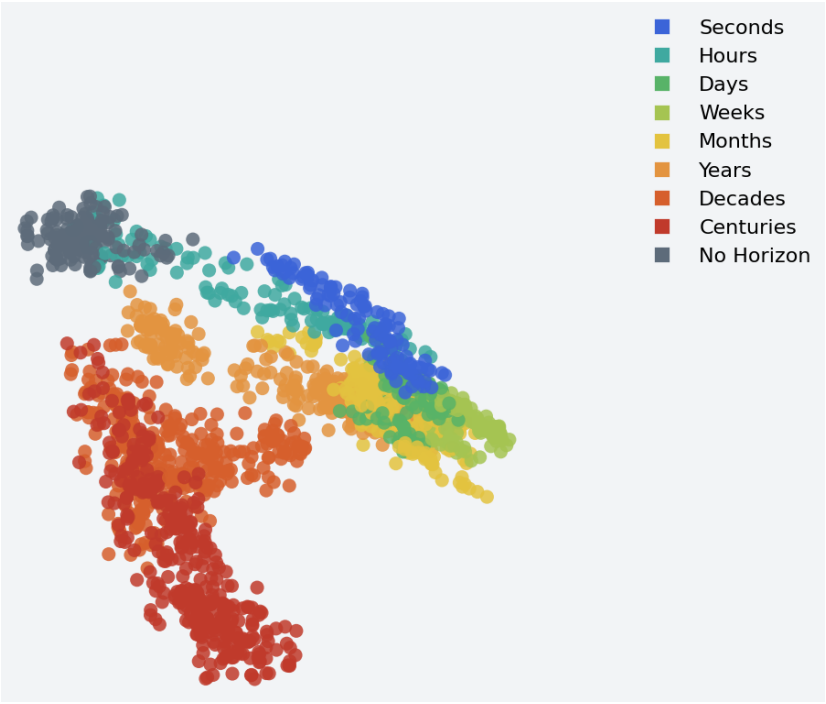

Figure 6: Three components of `resid_post` at `<|im_end|>`, for `Qwen3-4B-Instruct-2507` on the investment domain. Horizon orders the cloud from seconds to centuries, and the no-horizon prompts form their own cluster.

The same prompts sweep reward on a log grid, which gives the gradient a control inside the same activations. Horizon orders the component more strongly than reward does in every family we measure (Appendix A). We therefore contend that the ordering is specific to the horizon rather than a generic response to a graded quantity.

The same ordering appears at the first response tokens, at the same strength (A.1).

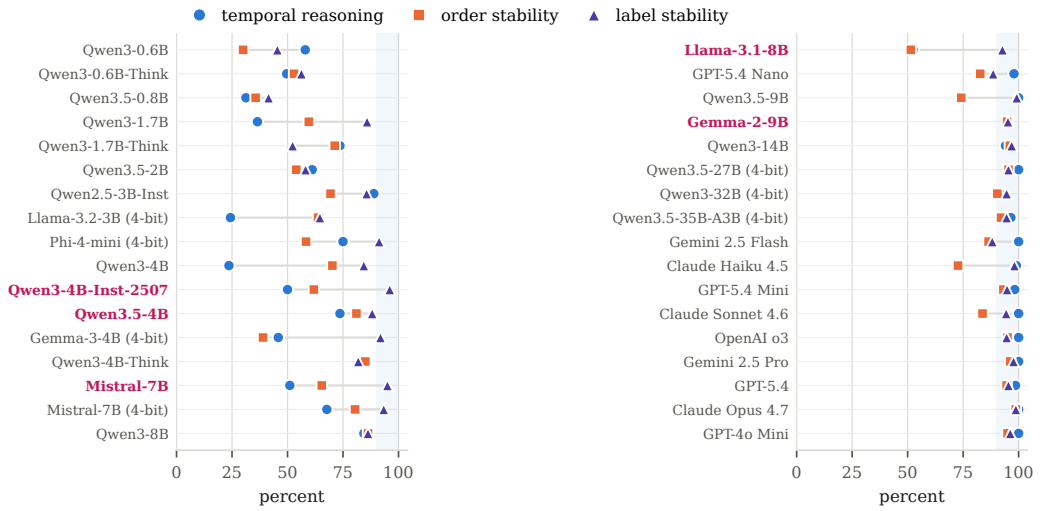

Figure 7: Behavioral summary for all 34 models, ordered by parameter count, with API models last. *Temporal reasoning* is the share of choices in the 1–5 year zone that pick the only option able to deliver by the deadline. *Order* and *label stability* are the shares of matched pairs whose choice survives swapping the option order or relabelling the options. Highlighted rows are the five models this paper studies. Rows marked 4-bit are quantized variants, kept separate from their full-precision counterparts. Two independent runs of `Qwen3-4B-Instruct-2507` score 50.3% and 50.0% temporal reasoning. `gemma-2-9b-it` reaches 95.1%, inside the coherent range, while `Llama-3.1-8B-Instruct` (52.4%) and `Mistral-7B-Instruct-v0.3` (51.0%) sit at chance.

Coherence. The representation is clean, and we now ask whether the behavior that follows from it is. We test whether these models make instrumentally coherent choices (Section 2) on investment prompts that pit a near reward against a larger delayed one. We score coherence in the **reasoning zone**, the range of deadlines where only the near option can deliver in time. They do not meet that bar. Two independent runs of the target model give the same coherence value, so the instrument is stable across runs (Figure 7). The failure tracks option position. Reward size and label format leave it unmoved. Coherence does not track scale or provider (Figure 7). Most frontier API systems reach the coherent range, and so does one open 9B model, while several larger open models do not. The smaller `Claude` variants get there through a binary heuristic that collapses at longer horizons. Every family we patch carries the geometry above, and only `gemma-2-9b-it` reasons coherently over it. We therefore contend that the representation does not deliver the behavior on its own.

5.3 Could temporal preference be controlled?

Adding a contrastive activation vector inside the located band shifts temporal preference in every family we test (Table 5). We compare each configuration against a random direction of matched norm injected at the same layer, so the shift has to beat an arbitrary perturbation of the same size, and it does in almost every cell of the sweep. The effective layers sit around half of depth in all four models. Open-ended generation follows, from triage framing under a negative coefficient to strategic framing under a positive one. Large coefficients degrade coherence, where the vector leaves the locally linear regime of the manifold. We therefore contend that a linear intervention inside the located band changes the model's choice, and that it does so only within a bounded range of coefficients.

6 Discussion

Whether a model holds the right temporal preference is an alignment question. Post-training alone does not settle it. We can read the preference at a specific depth and change it a few layers below. We can monitor it while the model runs. We can correct it with a steering vector when it drifts. The

Model	Layer	Depth	α	S	S_{ctrl}	Baseline	$S - S_{\text{ctrl}}$ [95% CI]
`Qwen3-4B-Instruct-2507`	18	0.51	20	5.9	2.9	0.9	2.9 [1.3, 4.7]
`Llama-3.1-8B-Instruct`	18	0.58	35	17.4	6.4	2.2	11.1 [7.2, 15.1]
`gemma-2-9b-it`	21	0.51	50	3.9	1.4	1.4	2.4 [1.7, 3.2]
`Mistral-7B-Instruct-v0.3`	19	0.61	20	12.1	2.6	-2.2	9.5 [6.8, 12.5]

Table 5: Best steering configuration per model. S is the forced-choice score and S_{ctrl} the same injection with a random direction of matched norm. Depth is the layer divided by one less than the layer count. The sweep covered depths 0.50 to 0.65, so it bounds the band rather than discovering its edges, and the vector beats the matched-norm control in 77 of its 80 cells. The last column re-scores each best cell with per-prompt logging and bootstraps 10,000 paired resamples over the 20 held-out prompts. That re-score reproduces the scores in this table within 0.11 in log-probability units, and no resample reverses the sign for any model.

correction holds inside a usable range of coefficients, and past that range coherence falls away. We therefore posit that interpretability of this kind can serve as runtime infrastructure.

7 Limitations and future work

We localize components and their depth. We do not trace the information flow between them. Each family runs on its own domain, so a family effect and a domain effect are hard to separate. Only the two runs on one model separate them here. Our reward control shares its prompts with the horizon sweep, so it bounds the ordering against one other graded quantity and not against every one. We assigned the contrastive labels synthetically and never validated them with humans. We derive the steering vector off policy, so it may carry features that only correlate with temporal preference. Our classification prompts also entangle horizon with achievement vocabulary.

Circuit tracing could reach atomic components and the flow between them. The parameterization could extend past horizon to reward, risk, role, and domain, and to adjacent concepts such as urgency. All our settings are single-turn. Temporal preference matters most in agentic settings, where the representation may shift as the conversation runs. Steering along the curvature of the manifold should avoid the degradation we see in the linear case.

8 Conclusion

We locate the components that carry temporal preference, and their depth is better described as a fraction of the stack than as a layer index. Time horizon is laid out there as an ordinal gradient at the turn-transition tokens. Every family we patch carries that geometry, and only one of them reasons coherently over it. The representation does not deliver the behavior on its own. A contrastive vector added inside the located band moves the choice in either direction. Interpretability has concentrated on binary contrastive concepts, and we take a first step toward dimensional ones such as time.

AI use statement

Large language models assisted this work in three ways. They generated and verified the minimally-framed and classification prompt datasets, as Table 2 records. They helped implement the analysis code, the figures, and the experiment tooling. They polished the prose. We reviewed all AI-assisted text, code, and figures against the underlying data, and the analyses behind every reported number were rerun and checked by the authors. We did not use LLMs for theoretical models, mathematical claims or proofs, hypothesis generation, methodology design, translation, data cleaning outside the scripted pipelines above, qualitative analysis, or the interpretation of results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

Every number in the paper traces to a released artifact. The prompt banks, the extracted activations, the patching outputs, and the analysis scripts are public, and the appendix names the run behind each table. Code and artifacts are public.[2]

References

John Ameriks, Andrew Caplin, and John Leahy. Wealth accumulation and the propensity to plan. *The Quarterly Journal of Economics*, 118(3):1007–1047, 2003.

Michael E. Bratman. Intention and means-end reasoning. *The Philosophical Review*, 90(2):252–265, 1981.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models, 2025. URL https://arxiv.org/abs/2505.22637.

Jonathan D. Cohen, Keith Marzilli Ericson, David Laibson, and John Myles White. Measuring time preferences. *Journal of Economic Literature*, 58(2):299–347, 2020. doi: 10.1257/jel.20191074.

Thomas R. Cook, Sophia Kazinnik, Zach Modig, and Nathan M. Palmer. What do LLMs want? Finance and Economics Discussion Series 2026-006, Board of Governors of the Federal Reserve System, January 2026. URL https://www.federalreserve.gov/econres/feds/what-do-llms-want.htm.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Thomas J Dohmen, Armin Falk, David Huffman, and Uwe Sunde. Interpreting time horizon effects in inter-temporal choice. *CESifo Working Paper*, 2012.

Ronald J. Ebert and DeWayne Piehl. Time horizon: A concept for management. *California Management Review*, 15(4):35–41, 1973. doi: 10.2307/41164456.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2405.14860.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons, 2019. URL https://arxiv.org/abs/1902.06865.

Ernst Fehr and Andreas Leibbrandt. A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics*, 95(9–10):1144–1155, 2011.

Bernd Figner, Daria Knoch, Eric J. Johnson, Amy R. Krosch, Sarah H. Lisanby, Ernst Fehr, and Elke U. Weber. Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, 13 (5):538–539, 2010.

Shane Frederick, George Loewenstein, and Ted O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, 2002. doi: 10.1257/002205102320161311.

Alexander F Gazmararian. Valuing the future: Changing time horizons and policy preferences. *Political Behavior*, 47(2):553–572, 2025.

Leonard Green and Joel Myerson. A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130(5):769–792, 2004.

[2]Repository links are withheld for double-blind review and will be restored on publication.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion, 2024. URL https://arxiv.org/abs/2312.06942.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *arXiv preprint arXiv:2601.04480*, 2026.

António Guterres. Lethal autonomous weapon system "politically unacceptable, morally repugnant and should be banned", 2025. URL https://press.un.org/en/2025/sgsm22643.doc.htm.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms, 2024. URL https://arxiv.org/abs/2403.17806.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus?, 2023. URL https://arxiv.org/abs/2301.07543.

Joseph W. Kable and Paul W. Glimcher. The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633, 2007.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition, 2025. URL https://arxiv.org/abs/2502.00873.

Jeongbin Kim. The effects of time preferences on cooperation: Experimental evidence from infinitely repeated games. *American Economic Journal: Microeconomics*, 15(1):618–637, 2023.

Junsol Kim, James Evans, and Aaron Schein. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.02080.

Christine M. Korsgaard. The normativity of instrumental reason. In Garrett Cullity and Berys Gaut, editors, *Ethics and Practical Reason*, pages 215–254. Oxford University Press, Oxford, 1997.

Andrew Kyle Lampinen, Yuxuan Li, Eghbal Hosseini, Sangnie Bhardwaj, and Murray Shanahan. Linear representations in language models can change dramatically over a conversation, 2026. URL https://arxiv.org/abs/2601.20834.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models, 2026. URL https://arxiv.org/abs/2602.01654.

Ali Mazyaki, Mohammad Naghizadeh, Samaneh Ranjkhah Zonouzaghi, and Hossein Setareh. Temporal preferences in language models for long-horizon assistance. *arXiv preprint arXiv:2509.09704*, 2025.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024. URL https://arxiv.org/abs/2412.04984.

Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully autonomous AI agents should not be developed. *arXiv preprint arXiv:2502.02649*, 2025.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025. URL https://arxiv.org/abs/2505.18235.

Mehrdad Moghimi, Anthony Coache, and Hyejin Ku. Decoupling time and risk: Risk-sensitive reinforcement learning with general discounting, 2026. URL https://arxiv.org/abs/2602.04131.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *International Conference on Machine Learning*, pages 45069–45108, 2025. URL https://arxiv.org/abs/2504.13151.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2209.00626.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://arxiv.org/abs/2312.06681.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03658.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions, 2023. URL https://arxiv.org/abs/2308.14752.

Emil Persson, Gustav Tinghög, and Daniel Västfjäll. Intertemporal prosocial behavior: a review and research agenda. *Frontiers in psychology*, 15:1359447, 2024.

Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint arXiv:2410.16314*, 2024.

Qwen Team. Qwen3-4B-Instruct-2507. https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507, 2025a. Accessed: 2026.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M. Phillips, Fazl Barez, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear, 2026. URL https://arxiv.org/abs/2603.09313.

Greg Reilly, David Souder, and Rebecca Ranucci. Time horizon of investments in the resource allocation process: Review and framework for next steps. *Journal of Management*, 42(5): 1169–1194, 2016. doi: 10.1177/0149206316630381.

Neil K. R. Sehgal, Sharath Chandra Guntuku, and Lyle Ungar. Real-time deadlines reveal temporal awareness failures in LLM strategic dialogues, 2026. URL https://arxiv.org/abs/2601.13206.

Noah A. Shamosh, Colin G. DeYoung, Adam E. Green, Deidre L. Reis, Matthew R. Johnson, Andrew R. A. Conway, Randall W. Engle, Todd S. Braver, and Jeremy R. Gray. Individual differences in delay discounting: Relation to intelligence, working memory, and anterior prefrontal cortex. *Psychological Science*, 19(9):904–911, 2008. doi: 10.1111/j.1467-9280.2008.02175.x.

Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

Ola Svenson and Gunnar Karlsson. Decision-making, time horizons, and risk in the very long-term perspective. *Risk Analysis*, 9(3):385–399, 1989.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Hieu M. Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2510.26243.

Jinjin Wang, Yuzhen Li, Jun Luo, and Hang Ye. Measuring time preference: Theory, methods, and applications. *Acta Psychologica*, 261:105928, 2025a. doi: 10.1016/j.actpsy.2025.105928.

Rui Wang, Qihan Lin, Jiayu Liu, Qing Zong, Tianshi Zheng, Dadi Guo, Haochen Shi, Peixuan Han, Weiqi Wang, and Yangqiu Song. Rethinking prospect theory for LLMs: Revealing the instability of decision-making under epistemic uncertainty. *arXiv preprint arXiv:2508.08992*, 2025b.

Yuqing Wang and Yun Zhao. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, 2024. doi: 10.18653/v1/2024.findings-acl.382. URL https://aclanthology.org/2024.findings-acl.382.

Zehong Wang, Fang Wu, Hongru Wang, Xiangru Tang, Bolian Li, Zhenfei Yin, Yijun Ma, Yiyang Li, Weixiang Sun, Xiusi Chen, and Yanfang Ye. Why reasoning fails to plan: A planning-centric analysis of long-horizon decision making in LLM agents, 2026. URL https://arxiv.org/abs/2601.22311.

Yotam Wolf, Noam Wies, Dorin Shteyman, Binyamin Rothberg, Yoav Levine, and Amnon Shashua. Tradeoffs between alignment and helpfulness in language models with steering methods. *arXiv preprint arXiv:2401.16332*, 2024.

Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Feijiang Han, Qibo Xue, Zeping Yu, Chenming Shang, Xiao Liang, Jing Xiong, Hui Shen, Chaofan Tao, Zhengwu Liu, Senjie Jin, Zhiheng Xi, Dongdong Zhang, Sophia Ananiadou, Tao Gui, Ruobing Xie, Hayden Kwok-Hay So, Hinrich Schütze, Xuanjing Huang, Qi Zhang, and Ngai Wong. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models, 2026. URL https://arxiv.org/abs/2601.14004.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. Steering risk preferences in large language models by aligning behavioral and neural representations. *arXiv preprint arXiv:2505.11615*, 2025.

APPENDIX A GEOMETRY ACROSS MODELS AND DOMAINS

Every model shows the ordering the main text reports. Horizon runs from seconds to centuries along the first principal component at the turn-transition tokens. Table A.1 lists the runs, and the figures follow its order: the target model, its second domain, then the other families by size. Figure A.1 shows the investment run in three components. Figures A.2 to A.4 show the same run at the three layers the localization pipelines mark: the geometry layer, the peak causal attention layer, and the peak MLP layer (Figure 3). The startup run repeats the ordering on the target model, so the geometry survives a change of domain within one model (Figure A.5). Figures A.6 to A.8 show Mistral, Llama, and Gemma, each on its own domain. Figure A.9 scores the preference-class split at the final turn token of every run.

Model	Domain	Valid samples	Layer shown
`Qwen3-4B-Instruct-2507`	investment	4,585	L21 of 36
`Qwen3-4B-Instruct-2507`	startup	2,992	L19 of 36
`Mistral-7B-Instruct-v0.3`	education	2,991	L19 of 32
`Llama-3.1-8B-Instruct`	health	2,517	L21 of 32
`gemma-2-9b-it`	climate	2,992	L33 of 42

Table A.1: The geometry runs. Each run draws 3,000 samples at the turn-transition positions and keeps the valid ones, except the investment run, which draws on the full 4,588-prompt bank. The layer shown is a representative panel. The quantitative claims in Table A.3 sweep every stored layer.

Per-layer fans. Each line is one sample's PC1 projection of `resid_post`, traced across the run's stored layers at one turn token. The x axis lists those layers at even spacing. The fan widens with depth in every family.

A.1 THE GRADIENT PERSISTS INTO THE RESPONSE

The same extraction runs also store the first response tokens, where the model writes its answer. Those activations carry the horizon ordering, and it is as strong there as at the turn tokens (Figure A.15, Table A.2). We therefore posit that the model carries the ordering into the answer it writes.

Model	Domain	Response token	$\lvert\rho\rvert$	Layer
`Qwen3-4B-Instruct-2507`	startup	`I`	0.904	L19
		`choose`	0.851	L12
		`:`	0.855	L18
		`a/b`	0.509	L28
		`)`	0.673	L3
`Mistral-7B-Instruct-v0.3`	education	first	0.948	L16
		second	0.961	L17

Table A.2: Rank correlation between the first principal component of `resid_post` and the stated horizon, at the first response tokens of `Qwen3-4B-Instruct-2507` on the startup domain and `Mistral-7B-Instruct-v0.3` on the education domain. Each row gives the strongest layer for that token. The fourth response token is the answer token, and its identity differs with the choice. The startup run keeps 2,849 samples and the education run 2,842.

A.2 THE ORDINAL GRADIENT AT THE TURN TOKENS

We fit the principal component separately at each layer and each turn token, then rank-correlate it with the stated horizon. Table A.3 gives the strongest correlation over layers for each token. The component is fit independently per cell, so its sign carries no meaning and the table reports magnitude.

A reward control. These prompts sweep reward on a log grid alongside the horizon, so the same activations carry a second graded quantity. We correlate the same component with reward and compare.

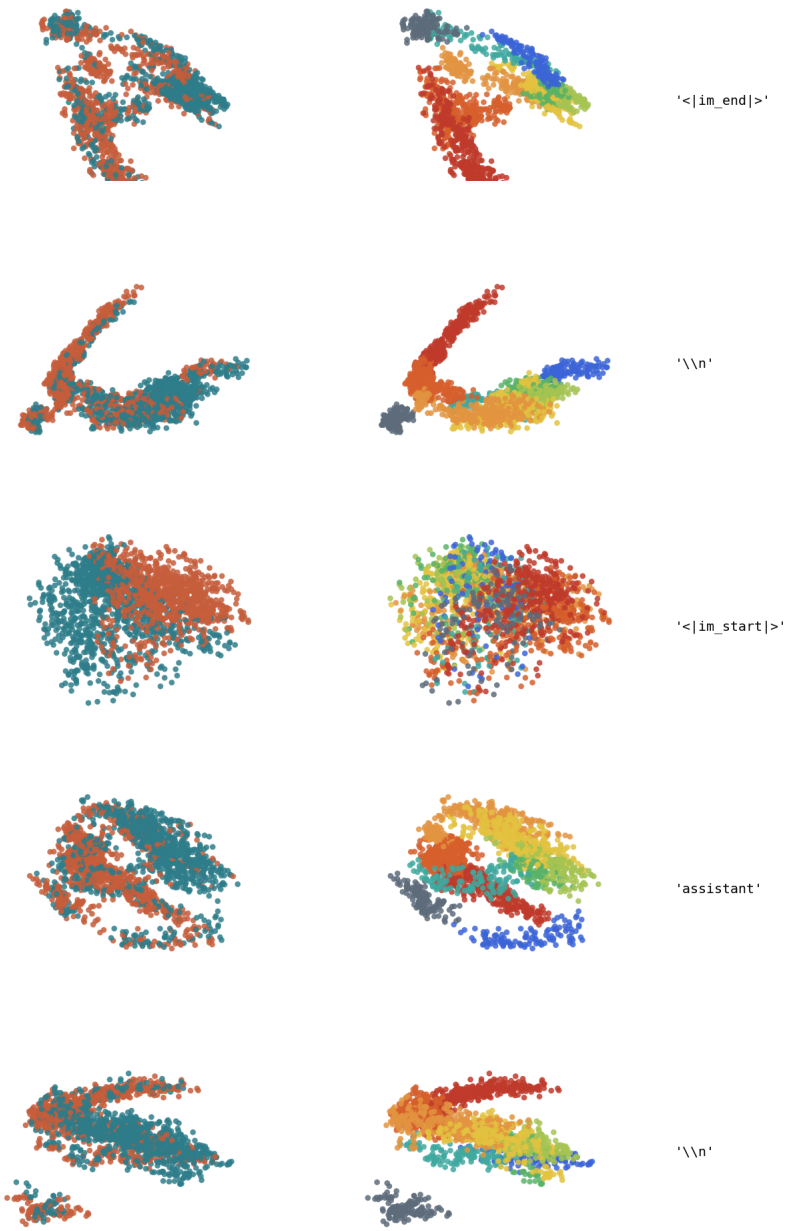

Figure A.1: Three components of `resid_post` at the turn tokens, for `Qwen3-4B-Instruct-2507` on the investment domain, at L21. The left column colors by preference and the right by horizon. The choices stay mixed while the horizon stays ordered.

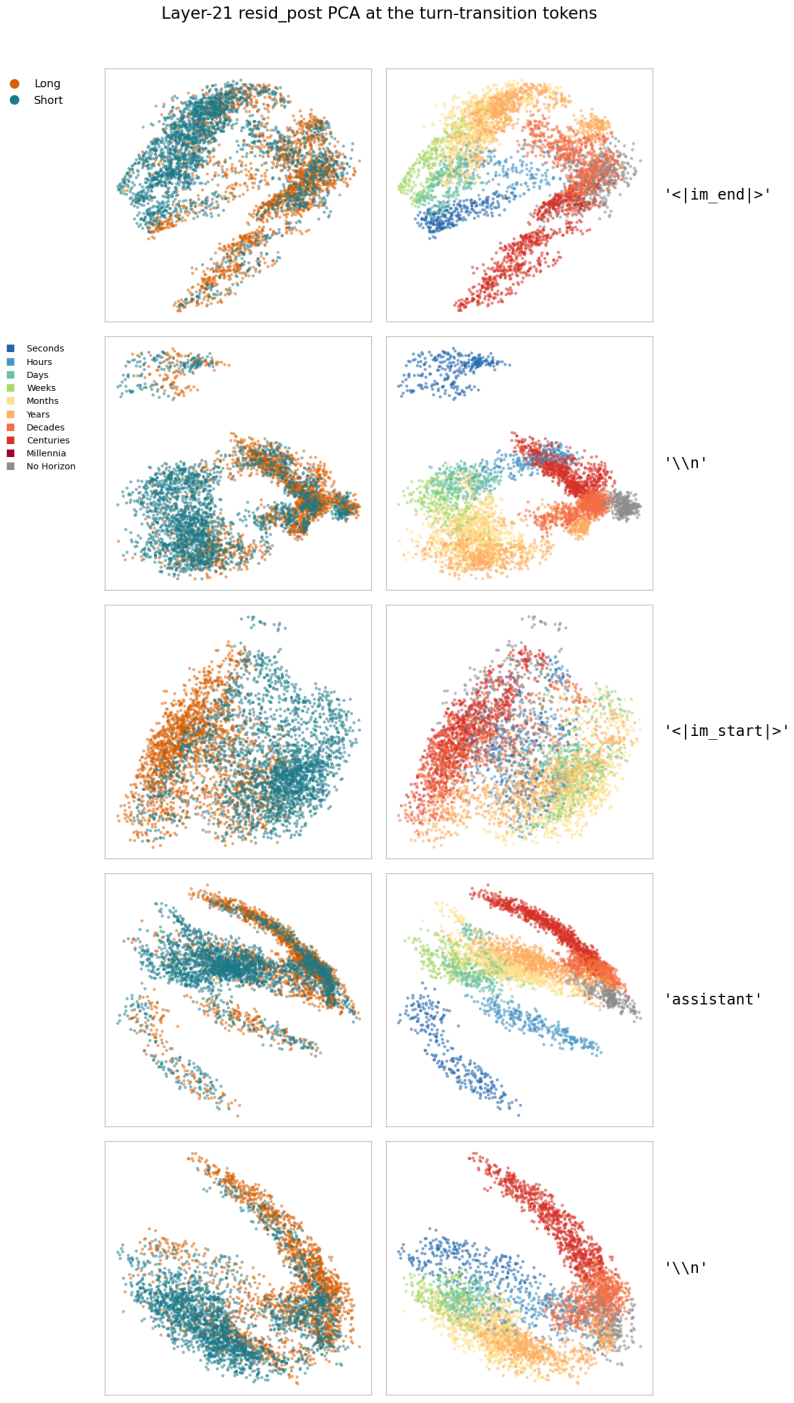

Figure A.2: `Qwen3-4B-Instruct-2507` on the investment domain, `resid_post` at L21 of 36, the layer of the geometry figures. The left column is colored by preference and the right by horizon.

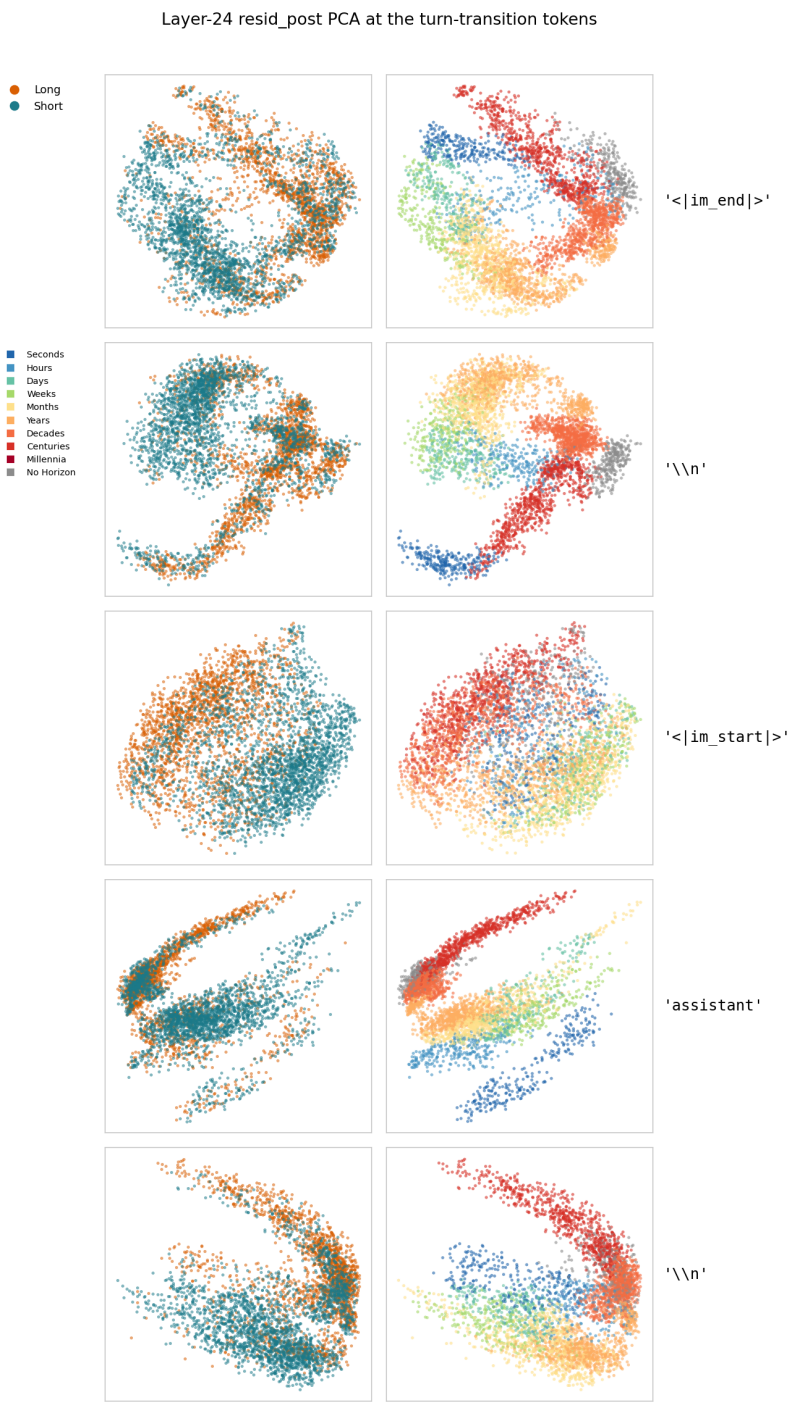

Figure A.3: `Qwen3-4B-Instruct-2507` on the investment domain, `resid_post` at L24 of 36, the peak causal attention layer. The left column is colored by preference and the right by horizon.

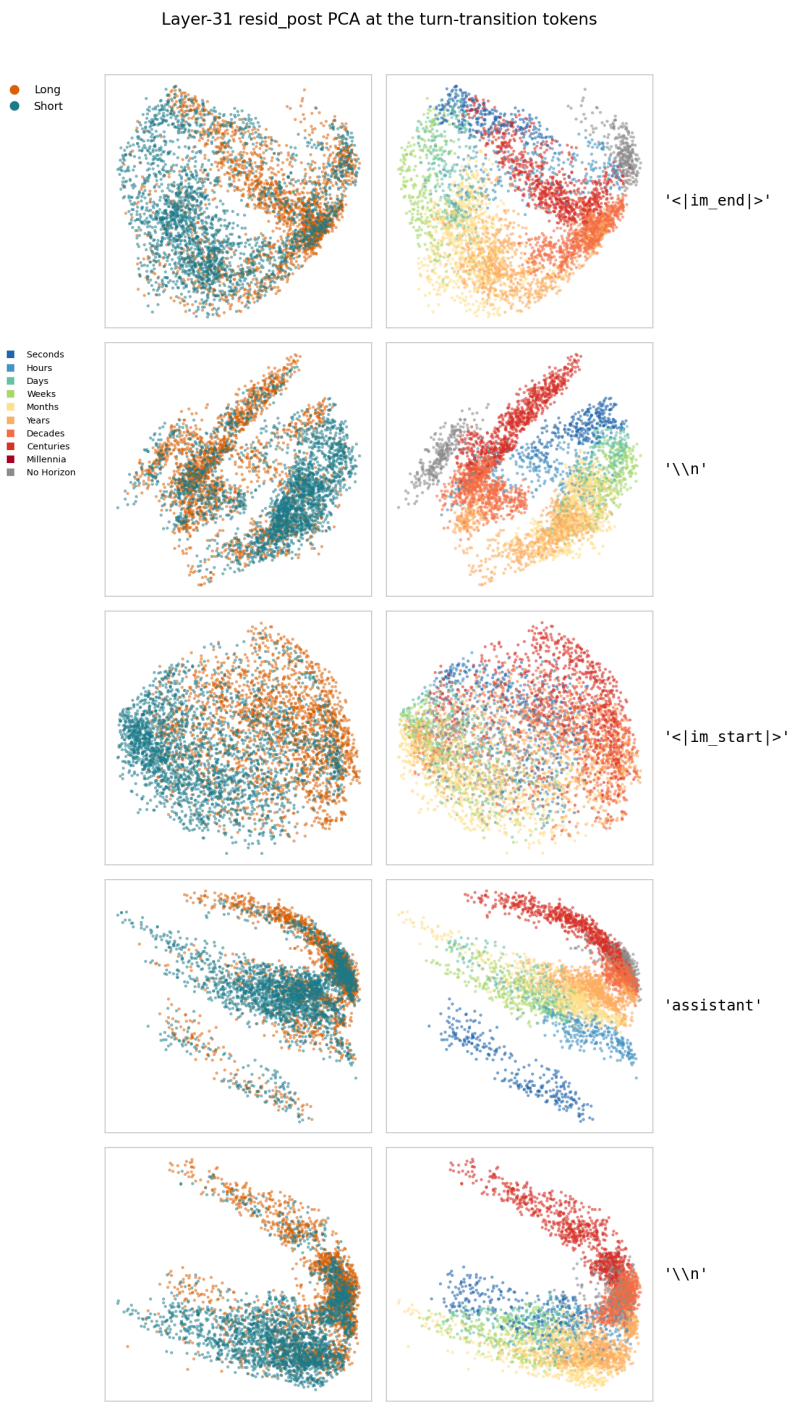

Figure A.4: `Qwen3-4B-Instruct-2507` on the investment domain, `resid_post` at L31 of 36, the peak MLP layer. The left column is colored by preference and the right by horizon.

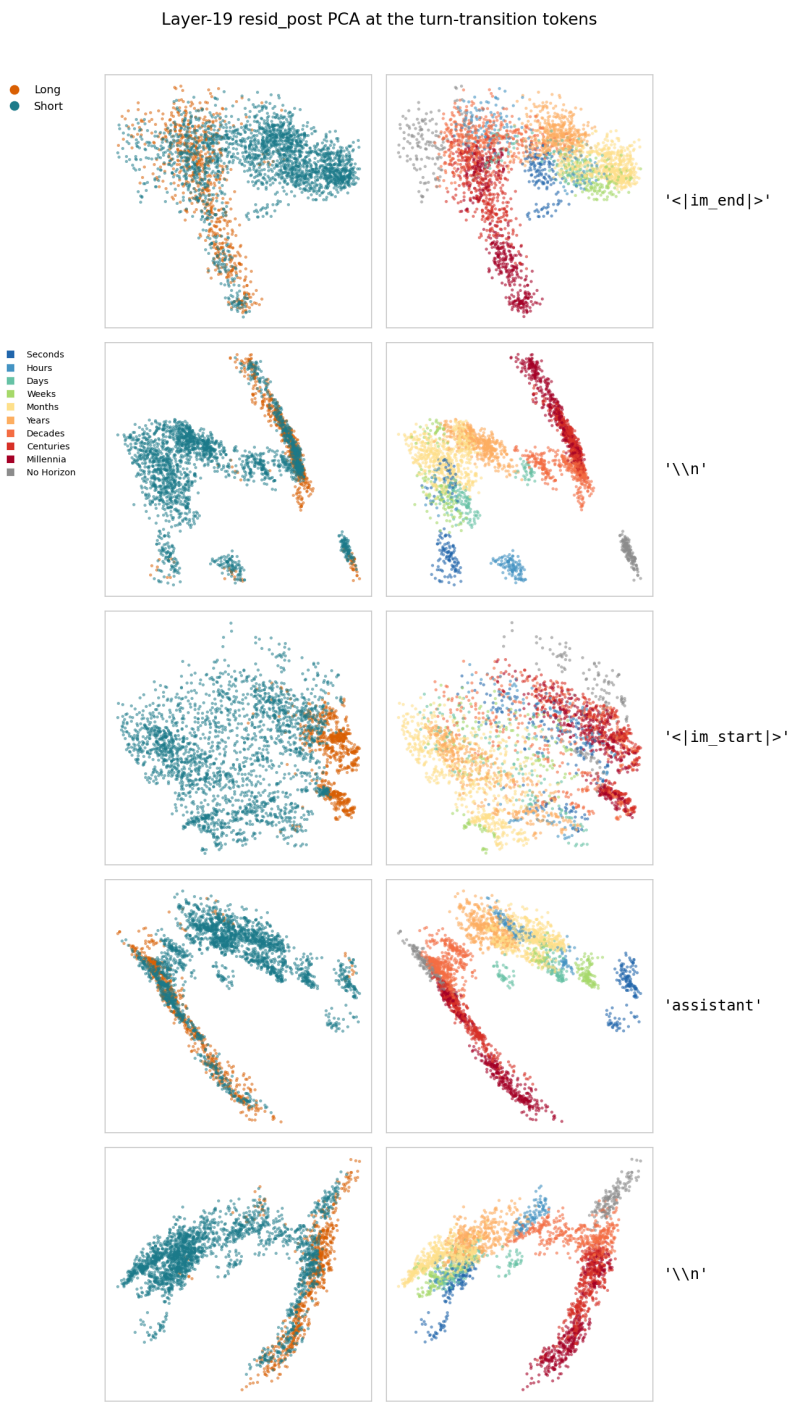

Figure A.5: `Qwen3-4B-Instruct-2507` on the startup domain, `resid_post` at L19 of 36. The target model on a second domain, with the same ordering.

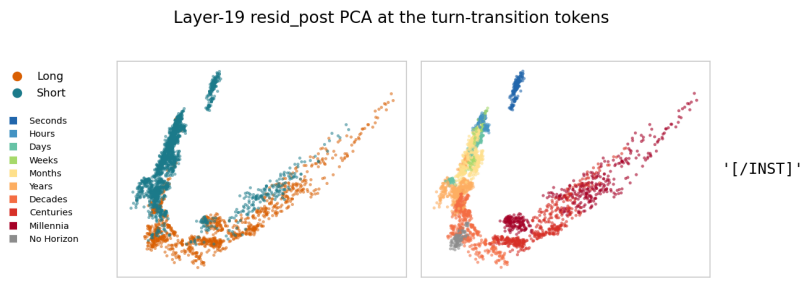

Figure A.6: `Mistral-7B-Instruct-v0.3` on the education domain, `resid_post` at L19 of 32. Mistral's template carries a single turn token, `[/INST]`, so its turn window is one position wide.

Model	Domain	Turn token	$\lvert\rho\rvert$
`gemma-2-9b-it`	climate	`<end_of_turn>`	0.965
		`\n`	0.963
		`model`	0.927
`Qwen3-4B-Instruct-2507`	startup	`<|im_end|>`	0.924
		`<|im_start|>`	0.918
		`\n`	0.855
`Llama-3.1-8B-Instruct`	health	`\n\n`	0.919
		`<|end_header_id|>`	0.845
		`<|start_header_id|>`	0.843
`Mistral-7B-Instruct-v0.3`	education	`[/INST]`	0.901
`Qwen3-4B-Instruct-2507`	investment	`<|im_start|>`	0.892
		`<|im_end|>`	0.842
		`assistant`	0.792

Table A.3: Rank correlation between the first principal component of `resid_post` and the stated time horizon, at the turn-transition tokens. Each row gives the strongest layer for that token. Horizons span seconds to centuries.

Model	Domain	Horizon $\lvert\rho\rvert$	Reward $\lvert\rho\rvert$
`gemma-2-9b-it`	climate	0.965	0.792
`Qwen3-4B-Instruct-2507`	startup	0.924	0.719
`Llama-3.1-8B-Instruct`	health	0.919	0.223
`Mistral-7B-Instruct-v0.3`	education	0.901	0.627
`Qwen3-4B-Instruct-2507`	investment	0.892	0.495

Table A.4: The strongest ordering of each quantity, over layers and turn tokens, in the same activations. Horizon orders the component more strongly than reward in every family.

We therefore posit that the ordering belongs to the horizon rather than to graded quantities in general.

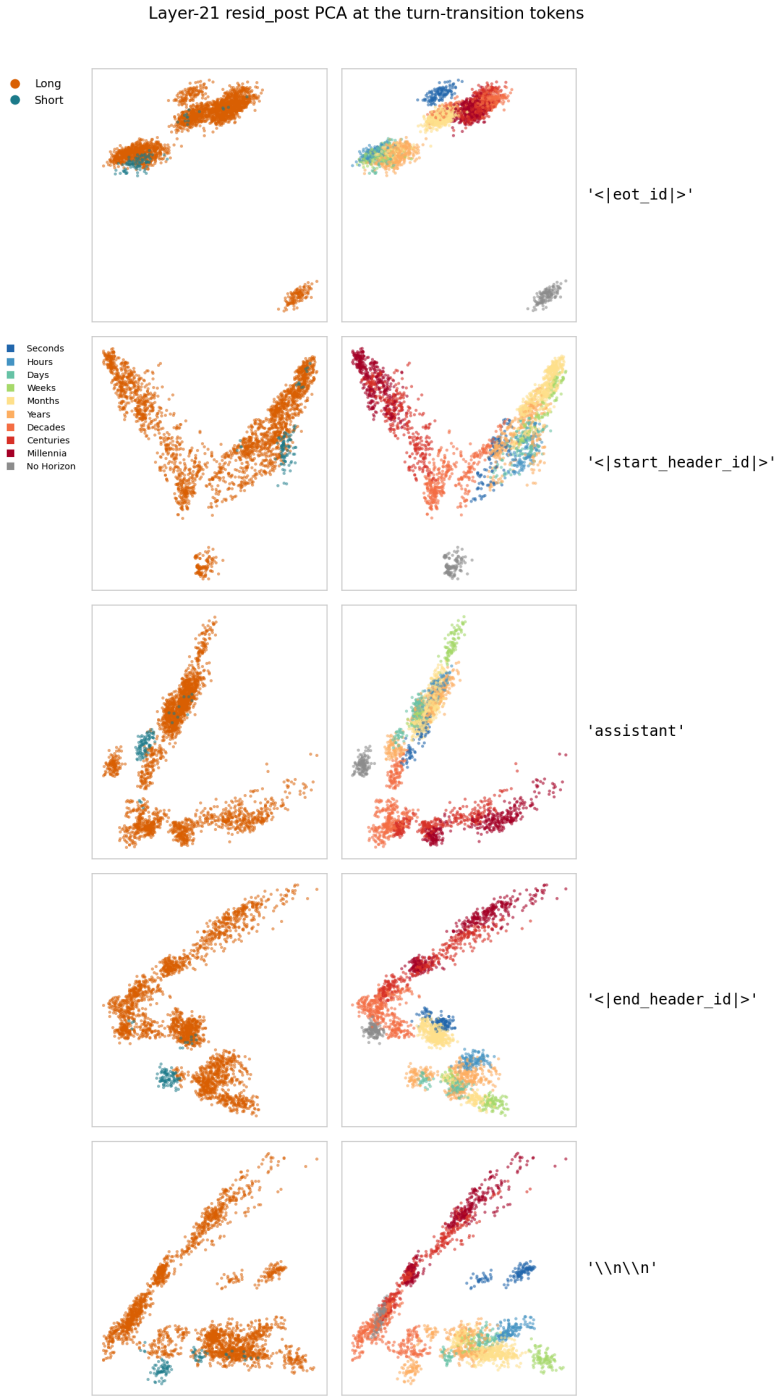

Figure A.7: `Llama-3.1-8B-Instruct` on the health domain, `resid_post` at L21 of 32. Rows are the five turn-transition tokens, ending at the double newline that opens the response. The left column is colored by preference and the right by horizon.

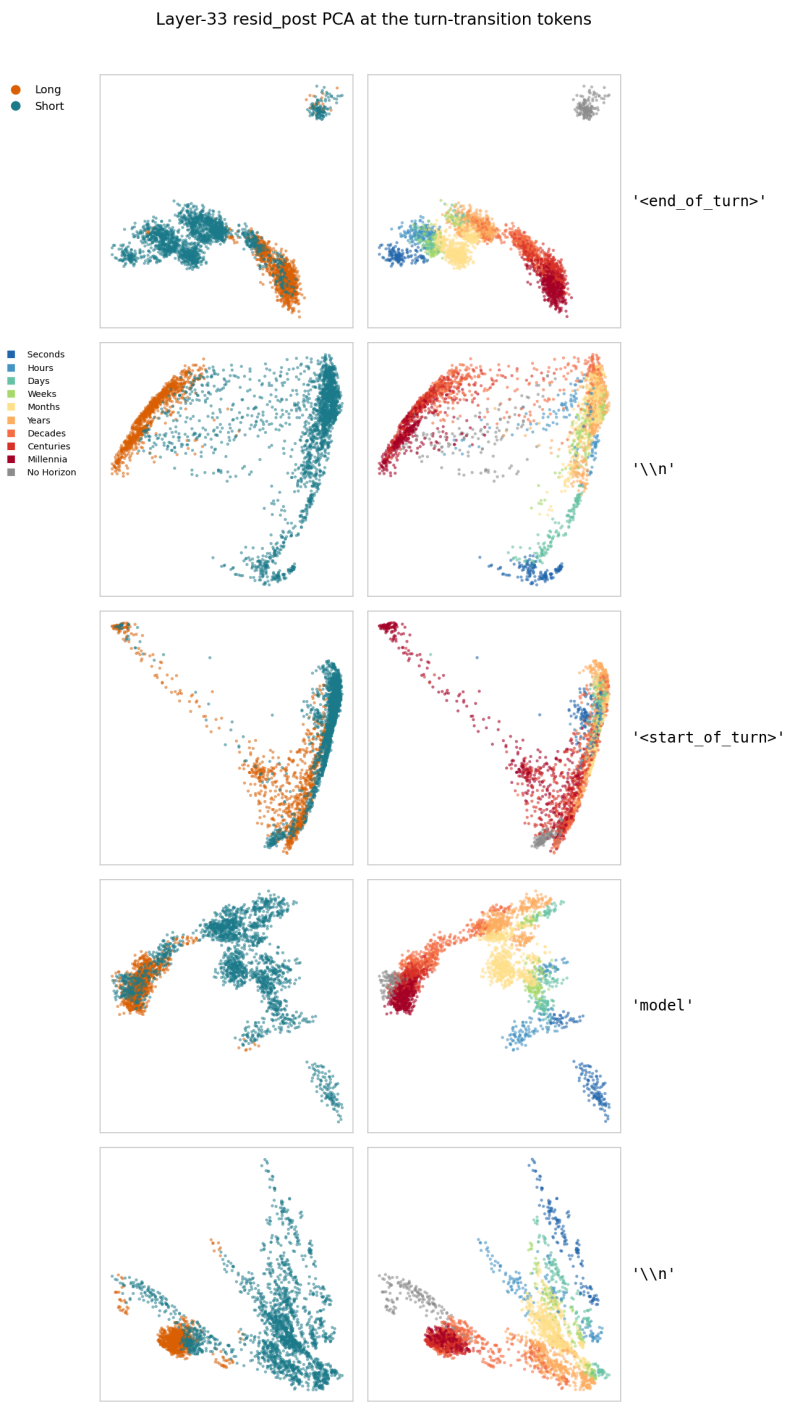

Figure A.8: `gemma-2-9b-it` on the climate domain, `resid_post` at L33 of 42. Rows are Gemma's own turn tokens, ending at the newline that opens the response.

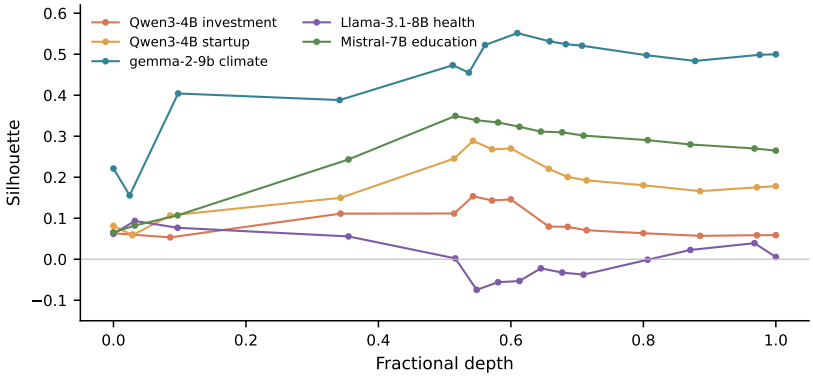

Figure A.9: Silhouette of the two preference classes at the final turn token, over each run's stored layers at fractional depth. Separation is partial at best and varies by family. The shared structure at these tokens is the horizon gradient.

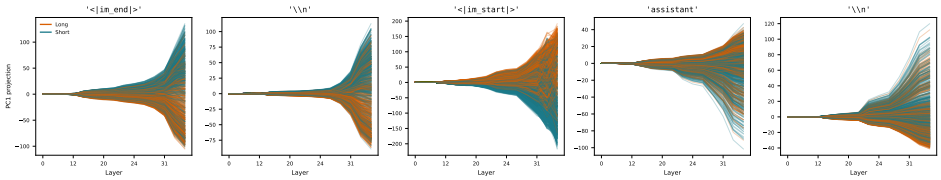

Figure A.10: PC1 fans for `Qwen3-4B-Instruct-2507` on the investment domain.

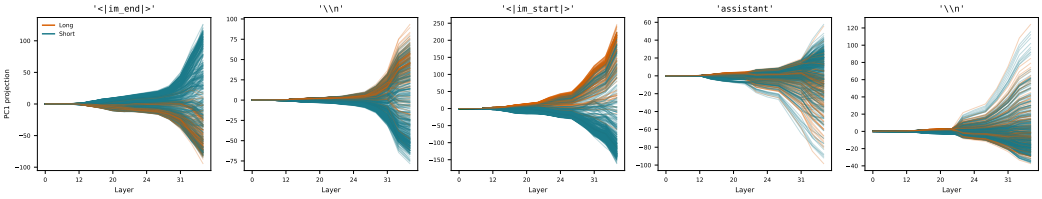

Figure A.11: PC1 fans for `Qwen3-4B-Instruct-2507` on the startup domain.

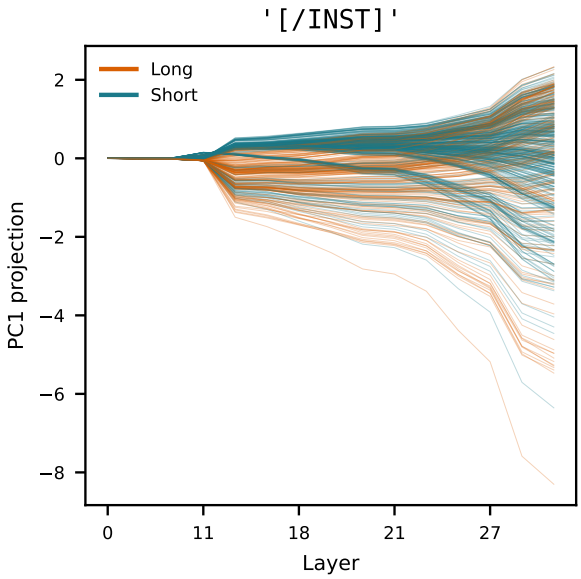

Figure A.12: PC1 fans for `Mistral-7B-Instruct-v0.3` on the education domain, at its single turn token.

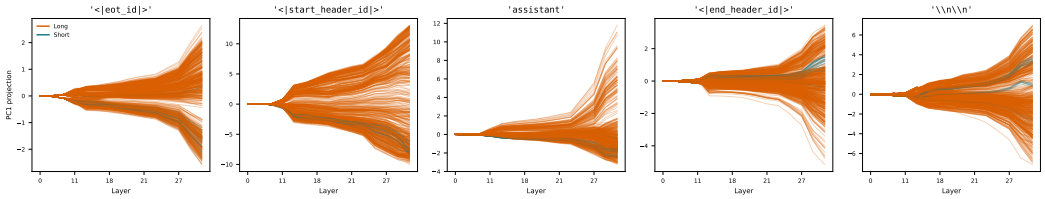

Figure A.13: PC1 fans for `Llama-3.1-8B-Instruct` on the health domain, one panel per turn token.

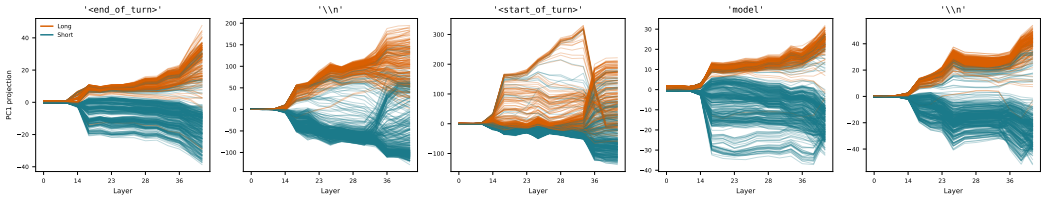

Figure A.14: PC1 fans for `gemma-2-9b-it` on the climate domain.

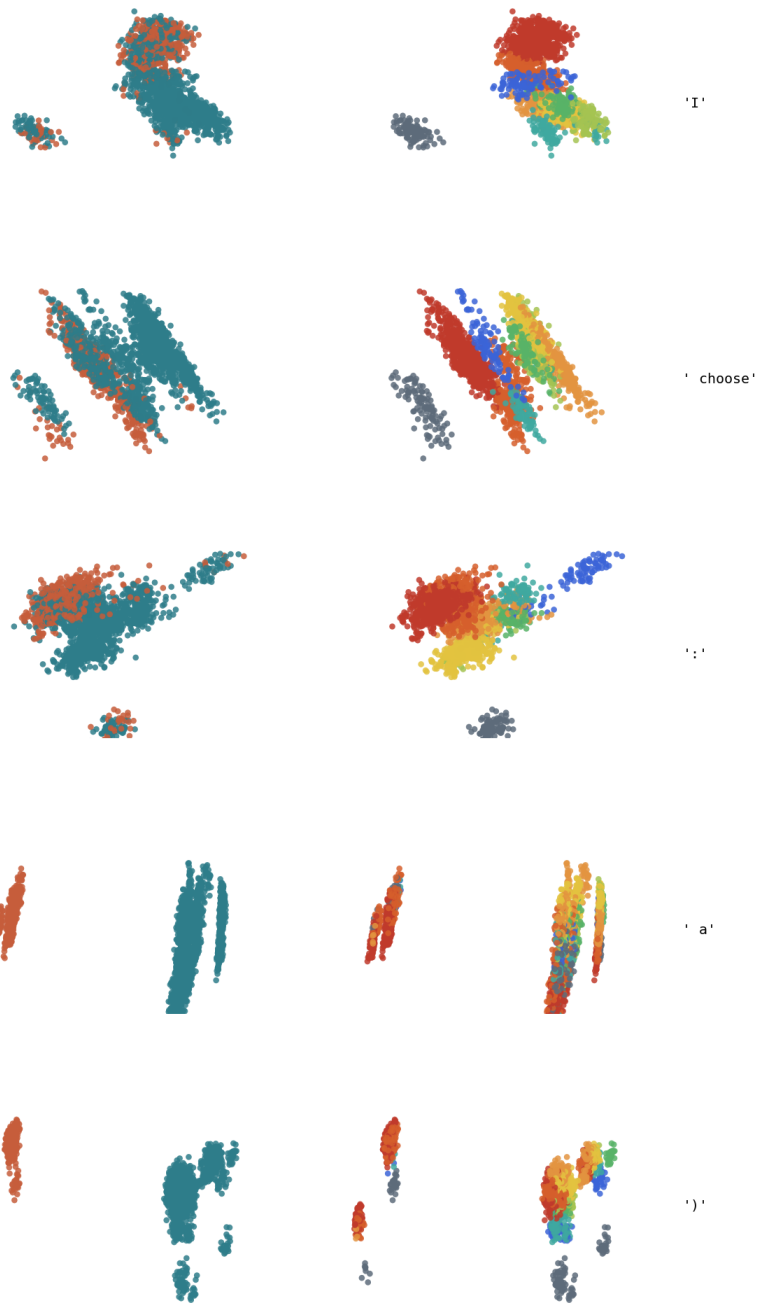

Figure A.15: Three components of `resid_post` at the first response tokens, for `Qwen3-4B-Instruct-2507` on the startup domain, at L21. Rows are the tokens `I`, `choose`, `:`, `a`, and `)`. The left column colors by preference and the right by horizon.

APPENDIX B CONTRASTIVE STEERING RESULTS

Localization and geometry describe where temporal preference sits and what shape it has. Neither shows that changing it changes the model's choice. We build a **contrastive activation vector**, the difference between the mean activation on long-term prompts and the mean activation on immediate prompts. We add it at a candidate layer during inference, scaled by a coefficient, and measure what the model then chooses.

Controls. We score every configuration against a random direction of matched norm at the same layer. We counterbalance the order of the two options in the forced-choice prompt. We index layers by fractional depth, so the same sweep means the same thing in models of different depth.

Vector recipe. A variant derives the vector as the mean of teacher-forced label activations at the best probe layer, rather than as a unit-normalized difference of means. This variant also steers `gemma-2-9b-it` at larger magnitude ($\alpha = 150$ at L25), so the effect is robust to the vector recipe.

B.1 SWEEPS PER MODEL

The sweep covers fractional depths 0.50 to 0.65 and four coefficients. The best cell for each model appears in the main text.

α	0.50	0.55	0.58	0.61	0.65
10	3.9	3.1	2.8	2.1	0.9
20	5.9	3.7	3.9	2.9	0.9
35	4.7	-0.2	3.3	3.8	0.8
50	4.6	1.1	1.4	3.6	0.6

Table B.1: Forced-choice score for `Qwen3-4B-Instruct-2507` over fractional depth and coefficient. An underlined cell is one where the vector does not beat the matched-norm random direction.

α	0.50	0.55	0.58	0.61	0.65
10	1.4	2.4	2.8	2.5	2.2
20	12.1	4.8	5.8	3.3	2.9
35	13.2	17.4	14.1	7.7	5.1
50	11.2	11.5	8.7	7.4	5.8

Table B.2: Forced-choice score for `Llama-3.1-8B-Instruct` over fractional depth and coefficient. An underlined cell is one where the vector does not beat the matched-norm random direction.

α	0.50	0.55	0.58	0.61	0.65
10	2.2	1.7	1.6	1.4	1.4
20	2.7	2.0	1.9	1.5	1.5
35	3.4	2.4	2.2	1.6	1.6
50	3.9	2.7	2.5	1.7	1.6

Table B.3: Forced-choice score for `gemma-2-9b-it` over fractional depth and coefficient. An underlined cell is one where the vector does not beat the matched-norm random direction.

α	0.50	0.55	0.58	0.61	0.65
10	5.4	-0.7	-1.2	-1.3	-1.5
20	10.7	<u>6.4</u>	12.1	5.4	2.6
35	10.2	8.4	8.0	6.3	7.3
50	10.3	9.0	8.3	6.9	7.3

Table B.4: Forced-choice score for `Mistral-7B-Instruct-v0.3` over fractional depth and coefficient. An underlined cell is one where the vector does not beat the matched-norm random direction.

What the steered text looks like. Under a negative coefficient the model advises immediate communication and asks what the deadline is. Under a positive coefficient the same prompt yields resilience, shared purpose, and adaptive capacity. The answer stays on topic.

Where the vector stops working. Above the swept range the output loses coherence. The vector is a straight line and the representation is curved, so a large enough step leaves the region where the line approximates the manifold. We therefore conjecture that steering along the curvature would buy the same shift without the loss of quality.

B.2 Side effects on instruction following

A steering vector is only usable if it leaves the rest of the model's competence intact. We steer each model in the long-term direction and score it on IFEval (Zhou et al., 2023), before and after the intervention (Table B.5). Moderate-strength steering costs three to five points of instruction-following accuracy, and the cost is similar across the two families we test.

		Strict		Loose	
Model		Prompt	Instruction	Prompt	Instruction
`Qwen3-4B-Instruct-2507`	baseline	0.705	0.786	0.722	0.799
($\alpha = 50$, $N = 464$)	steered	0.659	0.752	0.681	0.766
`gemma-2-9b-it`	baseline	0.712	0.790	0.730	0.804
($\alpha = 150$, $N = 510$)	steered	0.676	0.757	0.692	0.772

Table B.5: IFEval accuracy before and after steering in the long-term direction with the contrastive vector. N is the number of scored prompts after skipping malformed responses (77 for `Qwen3-4B-Instruct-2507`, 31 for `gemma-2-9b-it`).