Intertemporal Preference Steering in Qwen3 via Contrastive Activation Addition

Michal Mráz [1] **Justin Shenk** [1]

Abstract

We study linear representations of temporal horizon in the large language model `Qwen3-32B` and use them to change the model's time-related preferences, recommendations, and capabilities. We train contrastive linear probes on teacher-forced temporal-choice answers to find a short-term versus long-term direction in the model's residual stream and evaluate contrastive activation addition steering on a held-out binary temporal-choice task, an out-of-distribution monetary intertemporal-choice task, and a TravelPlanner capability benchmark. The central result is that temporal-horizon directions can be identified with simple contrastive linear probes and then used for steering to induce large, bidirectional preference changes. On an out-of-distribution monetary choice task that varies reward size and delay, steering strongly shifts the model's indifference threshold between smaller-sooner and larger-later rewards in both directions. We further show improvements on a planning-related capability metric under moderate temporal steering. These results suggest that model intertemporal preferences are measurable and steerable, which is relevant for AI systems that give advice involving delayed costs and benefits, and for safety questions about long-horizon planning.

1. Temporal-Horizon Direction and Probe Diagnostics

1.1. Direction Construction

This section shows that intertemporal preference in the residual stream of the large language model `Qwen3-32B` is linearly separable and defines the temporal-horizon vector used in the downstream activation-steering experiments. Linear directions have been used to study high-level model representations (Marks & Tegmark, 2024), and contrastive activation addition uses such directions as steering vectors during inference (Rimsky et al., 2024). We therefore focus on a signed contrastive difference-in-means direction. The direction is trained on paired short-term and long-term answer continuations. Because these continuations might also differ in abstraction, urgency, or other semantic concepts, we treat probe accuracy as evidence that the training contrast is linearly accessible, not as evidence that the direction isolates temporal horizon alone. For diagnostic purposes, this direction can be evaluated as a linear probe. We report the long-term versus short-term classification accuracies and compare them with other standard linear readouts. We treat the classification accuracies as readout diagnostics but do not necessarily use the method that maximizes classification accuracy for steering. Causal claims are reserved for the intervention experiments in Sections 2, 3, and 4. This separation is standard in mechanistic interpretability: readout accuracy identifies accessible information, while activation patching or activation steering tests whether changing a representation changes behavior (Alain & Bengio, 2016; Tenney et al., 2019; Elhage et al., 2021; Meng et al., 2022; Turner et al., 2024; Rimsky et al., 2024; Li et al., 2023; Zou et al., 2023).

All experiments use `Qwen3-32B`, with thinking trace disabled (Yang et al., 2025; Qwen Team, 2025). Mean-mass (MM) difference directions and comparison probes are trained at layers

$$24,\ 28,\ 32,\ 36,\ 40,\ 44,\ 48,$$

which span the middle-to-late part of the 64-layer model. We use an explicit dataset of 500 temporal-choice questions with overt time markers and an implicit dataset of 300 questions that encode horizon through semantic contrasts without explicit time words. Both datasets are validated for clear horizon distinction and balanced presentation order. The explicit set is split at question level into 400 train and 100 test questions. The implicit set is used only for cross-domain evaluation on the full 300-question set. Each explicit question yields two teacher-forced examples, so the explicit training matrix contains 800 examples.

The implicit examples are intended to reduce reliance on explicit time words, but they do not fully isolate horizon: in

[1]Independent. Correspondence to: Michal Mráz <misko.mraz@gmail.com>, Justin Shenk <shenk.justin@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Regime	Prompt	Short-term continuation	Long-term continuation
Explicit	When developing a roadmap for this initiative, consider:	What we can deliver in the next 30 days	What position we want in 10 years
Implicit	When the system fails, the first priority becomes	containing the spreading damage before it escalates	redesigning the architecture to prevent recurrence

Table 1. Representative probe-training examples. Each row yields two teacher-forced continuations with opposite temporal-horizon labels.

practice, long-horizon answers often become more systemic, preventive, and abstract, while short-horizon answers are often more local, reactive, and operational.

For each training example, the model receives the natural-language question as context and the short-term and long-term continuations are teacher-forced as the two distinctive options. The MM difference direction uses mean-answer pooling, which averages hidden states across all continuation tokens. Mean-answer pooling is used for the MM difference direction as a design choice: we expected a span-level answer feature to align better with interventions applied during generation than a single final-token feature. Final-token pooling at the last continuation token is also evaluated as a diagnostic variant. At each selected layer, the MM probe direction is

$$d_{\mathrm{MM}} = \mu_{\mathrm{long}} - \mu_{\mathrm{imm}},$$

where each mean is computed from the training activations at a fixed layer and feature. This contrastive construction follows the broader use of latent directions as simple behavioral readouts (Burns et al., 2023) and the CAA practice of adding contrastive residual-stream vectors during inference (Rimsky et al., 2024). It is especially relevant here because it preserves a signed short-term-to-long-term direction that can be added back into the residual stream at inference.

The same features are also used to fit comparison readouts. Logistic regression (LR) is fit directly. Whitened logistic regression (WLR) first fits a train-only covariance whitener, using regularized covariance

$$\Sigma_{\mathrm{reg}} = \Sigma + 10^{-2}\bar{\sigma}^2 I,$$

where $\bar{\sigma}^2$ is the average coordinate variance, and then fits LR in the whitened space. The whitened MM (WMM) probe applies the same covariance adjustment to the difference-of-means direction. These variants contextualize linear separability, while the downstream interventions use the unwhitened MM probe direction as the vector for CAA steering.

1.2. Probe Diagnostics

All probe families nearly saturate explicit holdout accuracy, while cross-domain implicit performance is lower but still substantial. The MM row shows that the specific contrastive direction used for CAA steering remains a strong temporal horizon readout even though LR and whitened variants attain slightly higher classification accuracy. Notably, the two whitened probe families have very similar classification accuracies across the reported splits. The steering experiments therefore use CAA steering with the explicit-trained mean-answer MM difference direction.

2. Binary Temporal Steering

This section shows that CAA steering with the explicit-trained MM difference direction causally affects temporal-choice behavior on held-out binary choices from the same broad short-term/long-term answer family. The evaluation uses similarly structured but unseen binary-choice questions. No evaluation question is used to fit the probe direction, and the only intervention is a signed residual-stream addition during inference. If positive and negative additions produce systematic changes in both parsed choices and forced-choice log probabilities, the direction is not only a classifier direction but a causal control direction for this task family. The intervention uses the explicit-trained mean-answer MM difference direction from Section 1.

2.1. Evaluation Prompt

Unlike probe training, steering evaluation presents the full binary-choice context, so the model must choose between two complete answer candidates rather than score an answer in isolation:

```
<question>
Options:
  <short-term or long-term option>
  <the other horizon option>
Answer:
```

The short-term and long-term options are presented in randomized order, and the evaluated datasets are the held-out explicit test set with 100 prompts and the full implicit set with 300 prompts.

2.2. Intervention

For a selected layer l, signed strength α, and normalized MM difference vector $\hat{d}_l$, the hook adds $\alpha\hat{d}_l$ to the residual stream at the last prompt token and the same update is also applied at each decode step. Positive strengths steer toward the long-term class and negative strengths steer toward the immediate class. Table 3 summarizes the intervention grid.

Probe	Explicit holdout	Implicit full
LR	99.0% (L28)	83.5% (L24)
WLR	99.0% (L32)	80.3% (L28)
MM	96.0% (L24)	77.0% (L32)
WMM	99.0% (L32)	80.3% (L28)

Table 2. Explicit-trained probe results for the main mean-answer pooling feature. Entries report the best accuracy and the layer attaining it. The MM row is the probe family from which the CAA steering vector is constructed.

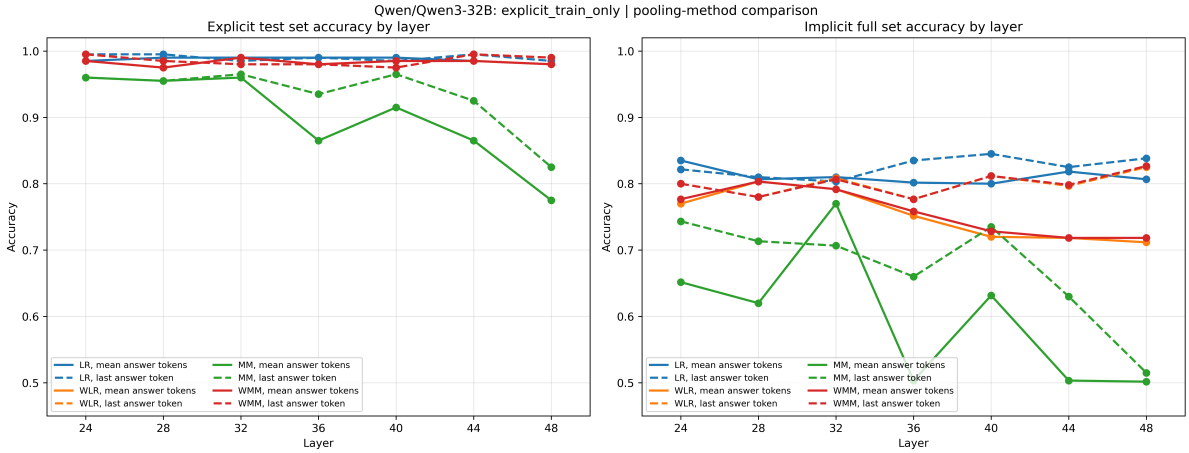

Figure 1. Layerwise explicit-trained probe accuracies from the probe-variation experiment. The left panel reports explicit test set accuracy; the right panel reports full implicit set accuracy.

We do not claim a uniquely optimal layer. We treat layers 24–48 as an exploratory probe-layer band and report full sweeps where feasible. The main result is that CAA effects are broad across middle-to-late layers, not that any specific layer is intrinsically privileged. Single-layer plots are representative slices chosen for readability or compute constraints.

Because the steering vectors are normalized, the added residual perturbation has L2 norm $|\alpha|$. The mean hook-site residual L2 norms in baseline (unsteered) generations are between 202 at layer 24 and 355 at layer 48. The strongest intervention $|\alpha| = 128$ is therefore 36%–63% of the unsteered norm, which is a deliberately large perturbation, while $|\alpha| \leq 16$ is less than 8% of the corresponding residual norm. For details on intervention and baseline residual stream magnitudes, see Appendix E.

2.3. Choice and Likelihood Metrics

Generated continuations are parsed into long-term or immediate choices, and the main behavioral metric is the proportion of parsed continuations choosing the long-term option. If the parser cannot identify the choice, the evaluator falls back to teacher-forced candidate logprob comparison; the plotted choice rates are computed from parsed, no-fallback choices. The likelihood metric computes

$$\Delta_{\text{avg}} = \ell(c_{\text{long}} \mid p) - \ell(c_{\text{imm}} \mid p),$$
$$\ell(c \mid p) = \frac{1}{|c|} \sum_{t} \log P(c_t \mid p, c_{<t}),$$

and reports its change relative to the unsteered baseline. This average-token logprob margin controls for continuation length; the full layerwise logprob surface is shown in Appendix B, Figure 6.

We see the causal steering effect across all explored layers, with layer 44 giving the largest parsed-choice contrast on both evaluation sets. This broad middle-to-late layer profile is consistent with prior activation-steering and intervention work, which often finds the strongest causal directions away from the earliest and latest layers of the stack (Turner et al., 2024; Rimsky et al., 2024; Durmus et al., 2024). Note that these numbers only describe the measured intervention profile; they do not by themselves identify a complete circuit for intertemporal preference.

As a matched-norm control, we also steer with a random unit vector orthogonal to the MM difference direction at each layer. This control produces substantially smaller choice shifts than CAA steering, and those shifts are not consistently aligned with the sign of the intervention (Appendix C, Figure 7). The main binary results use greedy decoding with temperature 0, so repeated decoding is not a source of

Quantity	Values
Steered layers l	$24,\ 28,\ 32,\ 36,\ 40,\ 44,\ 48$
Signed strengths α	$0,\ \pm2,\ \pm4,\ \pm8,\ \pm16,\ \pm32,\ \pm64,\ \pm128$
Patch positions	Last prompt token and every generated-token step

Table 3. Binary steering configuration. Each intervention adds $\alpha\hat{d}_l$ to the residual stream at layer l.

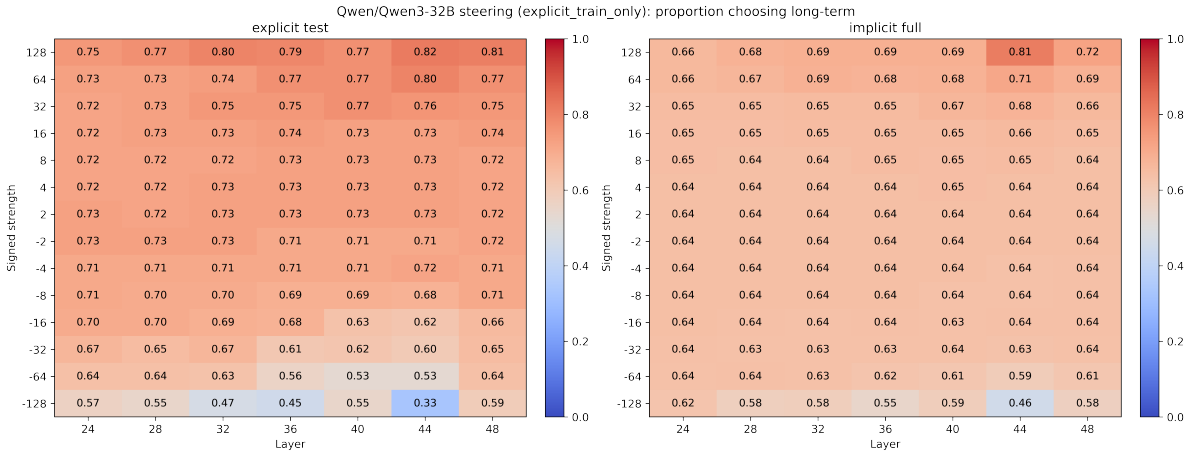

Figure 2. Proportion of prompts choosing the long-term option over layer and signed steering strength.

sampling variance for those point estimates. As a robustness check, however, Appendix D repeats a stochastic layer sweep with temperature 0.8, top-$p = 1$, and five sampled generations per prompt. The sampled means preserve the greedy ordering and magnitudes, with 2-sigma standard-error bars small relative to the steering effect.

3. Time-Utility Steering

This section shows generalization beyond the binary temporal-choice distribution toward settings where models might advise users about real-world tradeoffs. The model is asked monetary time-utility questions involving dollar amounts and delays, a quantitatively measurable task that was not used to train the temporal horizon direction. The goal is to show that a direction trained from semantic short-term versus long-term answers can change preferences on a structurally distinct advice task, not merely on held-out examples with the same format. This setting is related to classical time preference and discounting measurements (Samuelson, 1937; Frederick et al., 2002).

3.1. Task and Steering Setup

The prompt is written from the user's point of view:

```
I have an offer to take <x>$ now
or <y>$ in <t>.
What should I choose?
Answer in just a few words.
```

The monetary grid is defined by a base immediate amount x_b, base delayed amounts y_b, and amount scale factors c. Each prompt uses $x = cx_b$ and $y = cy_b$, so the delayed/immediate multiplier $m_b = y_b/x_b$ is held fixed while the absolute dollar amounts vary. This separates numeric proportions from absolute-dollar lexical effects. The experiment also uses day-count delay labels where possible rather than mixed natural labels. For example, "in 365 days" and "in a year" denote approximately the same delay, but the latter introduces additional calendar semantics.

Each condition therefore has 8 amount multipliers, 12 delays, and 8 scale factors, for 768 generations. Parsed responses are categorized as choosing the immediate amount, choosing the delayed amount or "it depends".

3.2. Area-Equivalent Indifference Multiplier

The area-equivalent indifference multiplier estimates the delayed/immediate reward ratio y/x at which the model is indifferent between taking x now and y after delay t. For a fixed condition and delay, all scale-factor rows with the same base multiplier $m = y/x$ are combined. Let $q_{\text{now}}(m, t)$ be the parsed present-choice fraction at multiplier m and delay t. "It depends" responses are split equally between present and delayed choices, although this convention has negligible effect because "it depends" occurs in only 0.15% of responses. Write $z_i = \log m_i$ for the sorted sampled multipliers in a delay column and define midpoint bin edges on the log axis by $e_0 = z_{\min}$, $e_n = z_{\max}$, and $e_i = \frac{1}{2}(z_i +$

Quantity	Values
Base immediate amount x_b	100
Base delayed amounts y_b	100, 101, 110, 150, 200, 500, 1000, 10000
Multipliers $m_b = y_b/x_b$	1.0, 1.01, 1.1, 1.5, 2, 5, 10, 100
Amount scale factors c	0.3, 0.5, 0.7, 1, 1.4, 2, 6, 8
Delays t	`1 hour`, `8 hours`, `1 day`, `2 days`, `7 days`, `30 days`, `90 days`, `365 days`, `730 days`, `1095 days`, `1825 days`, `3650 days`
Signed strengths α	0, ±8, ±16, ±32, ±64, ±128

Table 4. Time-utility steering grid. For each y_b, scale factor c, and delay t, the prompt compares $x = cx_b$ now against $y = cy_b$ after delay t.

z_{i+1}) for $1 \leq i < n$. Each sampled multiplier is then assigned bin width $w_i = e_i - e_{i-1}$. Treating $q_{\text{now}}(z, t)$ as piecewise constant on these bins, the ideal full-coverage log-axis average is

$$A_{\text{now}}(t) = \frac{\sum_i w_i q_{\text{now}}(m_i, t)}{\sum_i w_i},$$

since $\sum_i w_i = z_{\max} - z_{\min}$, the log indifference multiplier $z^*(t)$ is

$$\begin{aligned} z^*(t) &= z_{\min} + A_{\text{now}}(t)(z_{\max} - z_{\min}) \\ &= z_{\min} + \sum_i w_i q_{\text{now}}(m_i, t). \end{aligned}$$

This is a piecewise-constant log-axis average, chosen because the tested multipliers are sparse and unevenly spaced and the empirical choice columns can be locally non-monotone. The corresponding indifference multiplier on the original scale is $m^*(t)$:

$$m^*(t) = \exp(z^*(t)).$$

Large $m^*(t)$ means that the model requires a larger delayed reward before it switches away from the immediate option, while small $m^*(t)$ means that a smaller delayed reward is sufficient to make the model wait.

Appendix F shows the full present-choice heatmaps and precise AEIM values. The heatmaps show the same directional pattern summarized by AEIM: negative steering expands the present-choice region, while positive steering contracts it, without relying on the AEIM aggregation.

3.3. Layer Selection and Results

We sweep the same seven probe layers as in the binary-steering experiment. Figure 3 summarizes the geometric mean AEIM over the 12 delays for each layer and signed strength. The effect is signed and broad across layers. The layer-40 curves below are therefore an illustrative slice of the sweep, not the unique selected optimum. The other layerwise AEIM curves are shown in Appendix F.

The AEIM curves show that the result is delay-dependent rather than uniform: as expected, negative steering moves the indifference multiplier upward, positive steering moves it downward, and most displacement occurs at long delays. At ±128, the median ratio of negative-steer AEIM to positive-steer AEIM across delays is 4.90 and at 3650 days it is over 56. In other words, at the 10-year horizon, negative steering makes the model require more than 56 times the delayed reward needed under positive steering. The result is a strong generalization test: a direction trained only from binary temporal-horizon answers changes the model's measured intertemporal preferences on a new monetary-choice task.

For robustness, we performed a matched random-orthogonal control, which produces smaller, inconsistently signed movement (Appendix F). Unlike the MM difference vector, the random vector does not produce a consistent signed short-term versus long-term pattern across layers.

The main sweep uses greedy decoding with temperature 0. Analogously to Section 2, Appendix G reports robustness checks with stochastic sampling, which preserve the same qualitative pattern. Appendix E reports the steering vector magnitudes relative to unsteered hook-site residual-stream magnitudes, which are always less than 70% even at the strongest intervention.

4. TravelPlanner Capability

This section examines how temporal-horizon steering affects a planning-relevant capability metric, rather than only changing expressed preferences. We evaluate the same layer-40 temporal direction on TravelPlanner, a real-world travel planning benchmark that requires models to produce multi-day itineraries satisfying transportation, accommodation, restaurant, attraction, budget, and constraint requirements (Xie et al., 2024). A planning benchmark is a natural test for long-horizon planning because it requires maintaining a temporally extended plan while satisfying many local constraints. We use the validation split: 180 queries, comprising 20 queries for each combination of three difficulty levels and three trip lengths. Each query asks for a complete plan containing transportation, accommodation, meals, and attractions for each day. We use layer 40 as a compute-

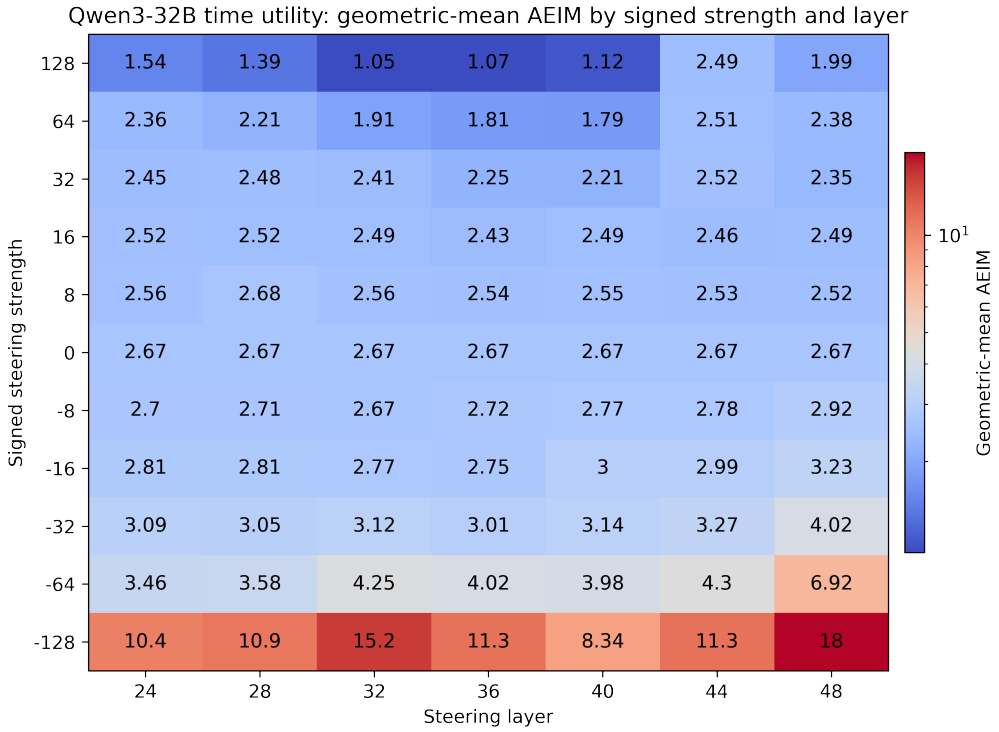

Figure 3. Time-utility layer sweep. Each cell reports the geometric mean AEIM.

Residual addition	Geom. mean AEIM	Parsed later rate	It-depends rate	Unparsed rate
$-128\hat{d}_{40}$	8.34	36.3%	0.0%	0.26%
0	2.67	54.7%	0.0%	0.00%
$+128\hat{d}_{40}$	1.12	80.4%	1.2%	0.52%

Table 5. Layer-40 aggregate time-utility results over all delay and amount cells. The geometric mean AEIM is computed across the 12 delay columns. Parsed later and it-depends rates are computed over parsed generations in the condition.

constrained representative layer from the middle-to-late probe band rather than as a task-specific optimum.

4.1. Metric Choice

The benchmark has multiple metrics; we use Commonsense Constraint Micro Pass Rate as the primary metric for this experiment. It is the total number of passed commonsense checks divided by the total number of commonsense checks across all evaluated queries. Commonsense micro pass rate is the most informative metric for our purposes. In the baseline condition (no steering), delivery rate is already near ceiling (96.7%), while commonsense macro pass rate, hard-constraint rates, and final pass rate are near floor. Commonsense micro pass rate therefore captures partial improvements in itinerary plausibility even when strict whole-query metrics remain too coarse to distinguish conditions.

Figure 5 shows a non-monotonic but directionally meaningful effect. Moderate long-term steering improves commonsense micro pass rate, while negative steering degrades it. Very large positive steering also degrades performance. A likely explanation is that $+128$ pushes the model far enough off-distribution that generation coherence begins to deteriorate. The result is not evidence that temporal steering generally solves planning, and we do not know from this aggregate alone which commonsense checks flip under steering, but it suggests that the same direction that shifts intertemporal preferences can also affect a planning-relevant capability metric, with useful effects concentrated at moderate positive strengths.

5. Discussion

5.1. Summary

These experiments suggest that temporal horizon can be represented by simple linear directions in Qwen3 and that CAA steering with a direction trained from contrastive short-term versus long-term answers can produce large behavioral

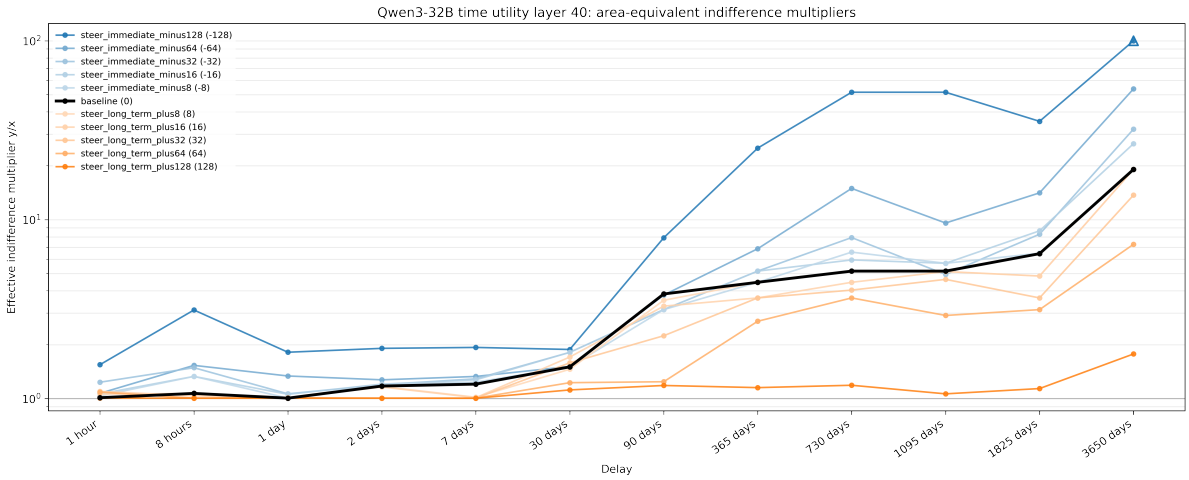

Figure 4. Area-equivalent indifference multipliers from the layer-40 time-utility experiment. Note the -128 steering at 3650 days is capped at 100 because all model responses chose the immediate option.

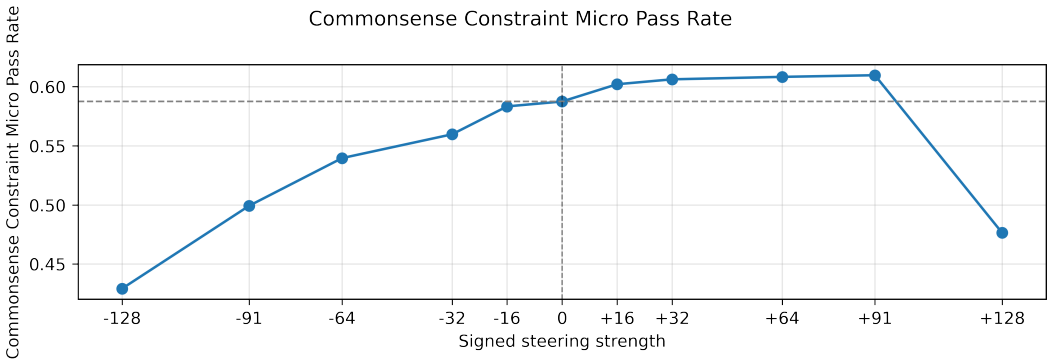

Figure 5. TravelPlanner Commonsense Constraint Micro Pass Rate as a function of signed temporal steering strength. Positive values steer toward the long-term direction; negative values steer toward the short-term direction.

effects. The direction should not be interpreted as an isolated temporal-horizon feature, since the training contrast may also contain correlated semantic features. Nevertheless, the time-utility experiment suggests that intertemporal preference is at least a relevant component of the direction. The probe results establish that short-term and long-term answers are linearly separable, and the binary steering experiment shows that steering along the direction causally affects held-out temporal-choice questions. The time-utility experiment is the strongest generalization result. A direction trained only on binary temporal-horizon answers shifts monetary intertemporal choice on a structurally different task, moving AEIM by large factors in both directions. The TravelPlanner experiment further suggests that temporal-horizon steering can affect a planning-relevant capability metric, although the effect is moderate and non-monotonic.

The main takeaway is that MM difference directions built from contrastive temporal-horizon answers can generalize beyond their training format when used for CAA steering. This matters because many deployed model outputs implicitly encode intertemporal preferences, such as advice about planning, finance, health, education, and policy that trades near-term actions against delayed outcomes. The results suggest that such intertemporal preferences can be measured and steered, rather than treated as an opaque side effect of pretraining or instruction tuning. This is also safety-relevant because temporal preferences may be one ingredient in long-horizon planning and reward-seeking, and a model that internally optimizes over a different horizon than the one it presents to users may be difficult to evaluate from outputs alone. At the same time, the capability to steer temporal preferences could be abused by a malicious actor to push a model toward decisions that are misaligned with user values, so the results underscore the importance of secure access controls and monitoring for similar behavior.

5.2. Limitations

The results rely on several assumptions that may fail in practice:

Construct validity. The MM direction is trained on short-term versus long-term answer continuations, but those continuations can also vary in abstraction, urgency, strategic scope, and proactive-versus-reactive framing. If these correlated features dominate the direction, the intervention should be interpreted as steering a broader short-term/long-term answer register rather than an isolated temporal-horizon variable. The monetary time-utility transfer suggests intertemporal preference is involved, but it does not fully identify the direction's semantic basis.

Intervention scale and specificity. Large activation additions may change more than temporal preference. The residual-norm diagnostics calibrate intervention size at the hook sites, but they do not prove that unrelated capabilities are preserved, especially at the strongest steering values. In deployment, an intervention that shifts time preference while also degrading factuality, instruction following, or coherence would be less useful and potentially unsafe.

Scope of empirical claims. All experiments use one model family and size, a small set of prompt templates, and thinking-disabled decoding. The results may not transfer to other architectures, smaller or larger models, reasoning-enabled settings, or tasks where temporal tradeoffs are expressed less explicitly. Performance may also depend on prompt wording or answer-format instructions.

Layer, strength, and prompt sensitivity. Layer and strength choices are exploratory. Full sweeps are reported where feasible, but some downstream experiments use representative layers for compute reasons. Performance may depend on layer choice, steering norm, and decoding settings.

Prompting baselines. The paper studies activation steering, not the full space of ways to alter model behavior. Similar behavioral changes might be possible by prompting the model to think more short-term or long-term, and such baselines are not reported here.

Parsing imprecision. Some generated answers are unparsed, especially in the free-generation steering experiments. The unparsed fraction is small enough that it is unlikely to change the qualitative results, but it can still introduce imprecision into parsed choice rates and derived quantities.

AEIM aggregation. AEIM compresses a sparse, unevenly sampled amount-delay grid into one threshold-like curve per condition. Different multiplier grids, delay grids, or aggregation conventions could change the apparent curve shape. The raw present-choice heatmaps in Appendix F show the same steering pattern without this aggregation, but they have a less clean scalar interpretation.

Probe-family comparisons. Steering is evaluated primarily with the MM difference direction. Matched-norm comparisons against LR, WLR, and WMM steering directions would clarify whether MM is uniquely useful for intervention.

Benchmark coverage. The TravelPlanner evidence is weaker than the time-utility result because the aggregate commonsense metric is indirect and does not identify which item-level checks change under steering.

5.3. Compute Resources

All experiments used Qwen3-32B inference. Activation collection and steering runs were executed on NVIDIA GPU hardware with 100 GB of available VRAM on RTX 5090 or similar-capability hardware. The layer-sweep steering experiments took approximately 48 hours for binary-choice steering and 24 hours for time-utility steering. The respective stochastic steering sweeps took about twice as long: 4 days and 2 days. TravelPlanner evaluation with single-layer steering took approximately 48 hours. The remaining diagnostics and plotting took seconds to minutes on a MacBook Air M4.

Acknowledgements

We thank Avigya Paudel for his help with the dataset curation and Joseph Rudoler for his thoughtful insights during our discussions. Finally, we thank SPAR for providing the compute resources for the research.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Durmus, E., Tamkin, A., Clark, J., Wei, J., Marcus, J., Batson, J., Handa, K., Lovitt, L., Tong, M., McCain, M., Rausch, O., Huang, S., Bowman, S., Ritchie, S., Henighan, T., and Ganguli, D. Evaluating feature steering: A case study in mitigating social biases. https://www.anthropic.com/research/evaluating-feature-steering, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/.

Frederick, S., Loewenstein, G., and O'Donoghue, T. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, 2002.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Qwen Team. Qwen3-32B model card. https://huggingface.co/Qwen/Qwen3-32B, 2025. Accessed 2026-04-20.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Samuelson, P. A. A note on measurement of utility. *The Review of Economic Studies*, 4(2):155–161, 1937.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024. Version 5, revised October 10, 2024.

Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., and Su, Y. TravelPlanner: A benchmark for real-world planning with language agents. In *International Conference on Machine Learning*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Appendix

A. Binary Steering Layerwise Summary

Table 6 gives the layerwise contrast table discussed in Section 2. Layer 48 is marginally largest by logprob gap on both datasets, while layer 44 is largest by choice gap.

B. Binary Steering Logprob Diagnostics

C. Binary Random-Orthogonal Control

D. Stochastic Binary Steering Robustness

The main binary steering sweep is decoded greedily. To estimate sensitivity to sampling randomness, Figures 8–10 show the stochastic layer sweep with temperature 0.8, top-$p = 1$, and five sampled generations per prompt. In Figures 8–10 and Table 7, error bars are 2-sigma standard errors, computed as $2\sigma/\sqrt{5}$ across the five repeat-level summaries for each cell. The sweep covers all seven probe layers but only the signed strengths $0, \pm 8, \pm 32, \pm 128$ due to compute constraints. The sampled subset preserves the qualitative steering pattern, suggesting the greedy result is not an artifact of deterministic decoding. Table 7 reports the layer-44 explicit-test slice corresponding to the main binary steering contrast.

E. Residual-Norm Diagnostics

F. Time-Utility Diagnostics

Figure 15 shows the full grid underlying the AEIM summaries. Yellow cells are high present-choice proportions, and dark cells are low present-choice proportions. Because delay increases from left to right and delayed/present amount increases from bottom to top, a rationally expected pattern has a yellow cluster in the bottom-right region: when the delayed amount is not much larger than the present amount and the wait is long, the present option should usually be preferred. Conversely, the top-left region should be dark, because a much larger delayed reward after a short wait should usually be preferred. The steering pattern follows the intended direction: negative strengths expand the yellow present-choice region, while positive strengths contract it; the expansion and contraction become larger as steering strength increases.

G. Stochastic Time-Utility Robustness

The main time-utility sweep is decoded greedily. To estimate sensitivity to sampling randomness, Figures 19 and 20 repeat a stochastic layer sweep with temperature 0.8, top-$p = 1$, and five sampled generations per prompt. The sweep covers all seven probe layers at signed strengths $0, \pm 8, \pm 32, \pm 128$, a strength subset of the full greedy grid in the main text due to compute constraints. This stochastic sweep preserves the qualitative steering pattern, suggesting the greedy result is not an artifact of deterministic decoding. In Figures 19–20 and Table 8, error bars are 2-sigma standard errors, computed as $2\sigma/\sqrt{5}$ across the five repeat-level summaries for each cell.

Layer	Explicit choice gap	Explicit logprob gap	Implicit choice gap	Implicit logprob gap
24	23.9 pp	0.158	4.6 pp	0.112
28	25.3 pp	0.125	10.2 pp	0.121
32	37.6 pp	0.175	11.0 pp	0.101
36	37.5 pp	0.183	14.6 pp	0.084
40	28.0 pp	0.399	10.0 pp	0.146
44	59.4 pp	0.813	36.5 pp	0.263
48	25.7 pp	0.824	14.6 pp	0.268

Table 6. Layerwise steering contrast. The choice gap is the best positive long-term no-fallback rate minus the best negative long-term no-fallback rate at that layer. The logprob gap is the best positive minus best negative long-minus-immediate average-logprob delta at that layer.

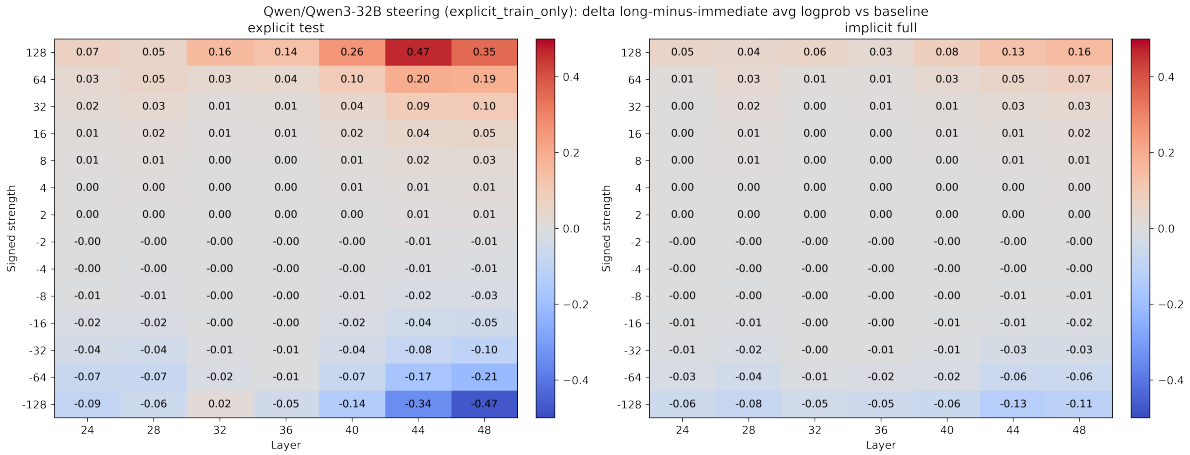

Figure 6. Change in long-minus-immediate average logprob margin relative to the unsteered baseline.

Residual addition	Greedy no-fallback long	Sampled no-fallback long	Sampled fallback
$-128\hat{d}_{44}$	33.7%	$40.6 \pm 1.9\%$	$2.4 \pm 0.8\%$
0	76.8%	$74.3 \pm 1.8\%$	$6.6 \pm 2.2\%$
$+128\hat{d}_{44}$	93.1%	$91.4 \pm 1.8\%$	$12.2 \pm 1.5\%$

Table 7. Stochastic binary steering repeat check on the explicit test set. The sampled columns report mean $\pm$ 2-sigma standard-error bars across five repeat-level sampled decodings.

Layer	$-128\hat{d}_l$	0	$+128\hat{d}_l$
24	8.48 ± 0.17	3.22 ± 0.22	1.63 ± 0.07
28	10.95 ± 0.72	3.22 ± 0.22	1.46 ± 0.04
32	15.20 ± 0.94	3.22 ± 0.22	1.10 ± 0.01
36	12.08 ± 0.49	3.22 ± 0.22	1.12 ± 0.01
40	7.84 ± 0.34	3.22 ± 0.22	1.18 ± 0.02
44	10.08 ± 0.66	3.22 ± 0.22	2.36 ± 0.09
48	15.11 ± 1.02	3.22 ± 0.22	2.09 ± 0.10

Table 8. Stochastic time-utility geometric-mean AEIM by layer. Entries report mean $\pm$ 2-sigma standard-error bars across five repeat-level sampled decodings.

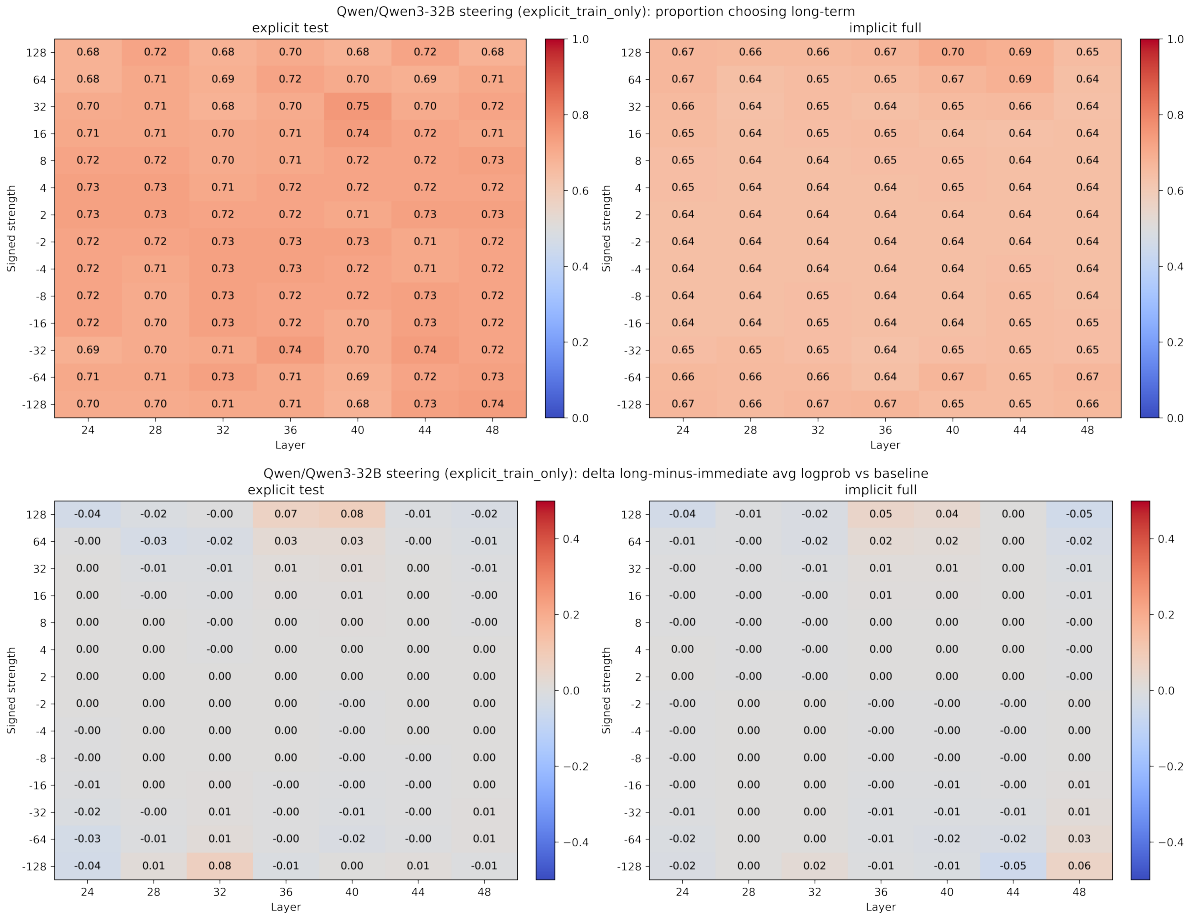

Figure 7. Matched random-orthogonal binary steering control. The control vector at each layer is orthogonalized against the corresponding normalized MM difference vector. The largest absolute no-fallback choice movement is 5.6 percentage points on explicit test and 5.5 points on implicit full, compared with 59.4 and 36.5 percentage-point CAA contrasts for layer-44 steering on the same datasets.

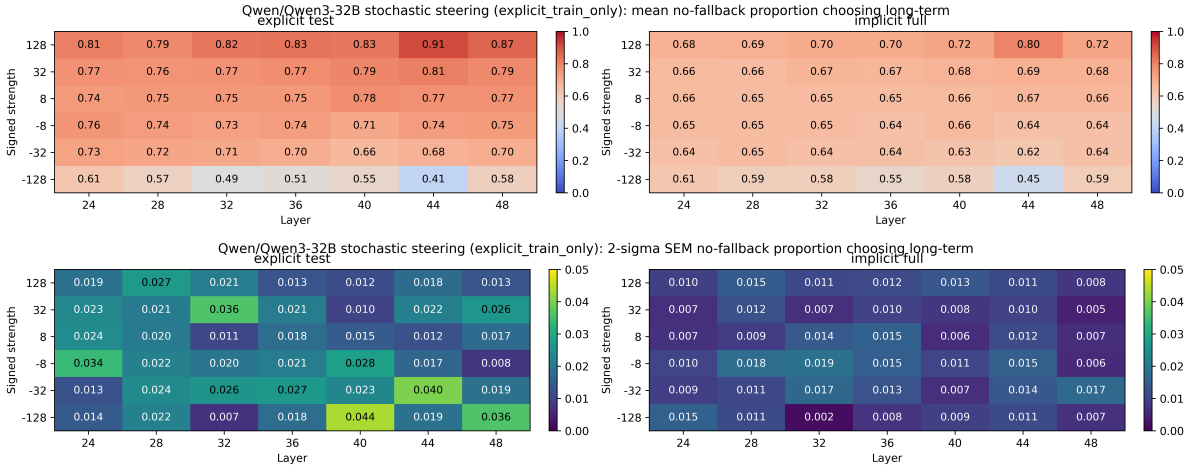

Figure 8. Stochastic binary steering no-fallback choice results. Top: mean proportion choosing the long-term option. Bottom: 2-sigma standard-error bar across five sampled decodings.

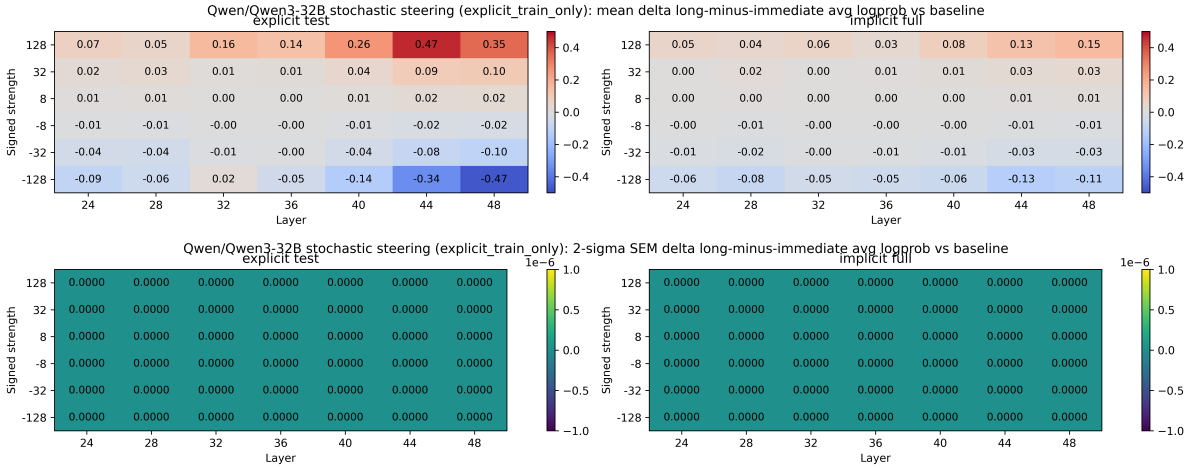

Figure 9. Stochastic binary steering forced-choice logprob results. Top: mean change in long-minus-immediate average logprob margin relative to baseline. Bottom: 2-sigma standard-error bar across five sampled decodings.

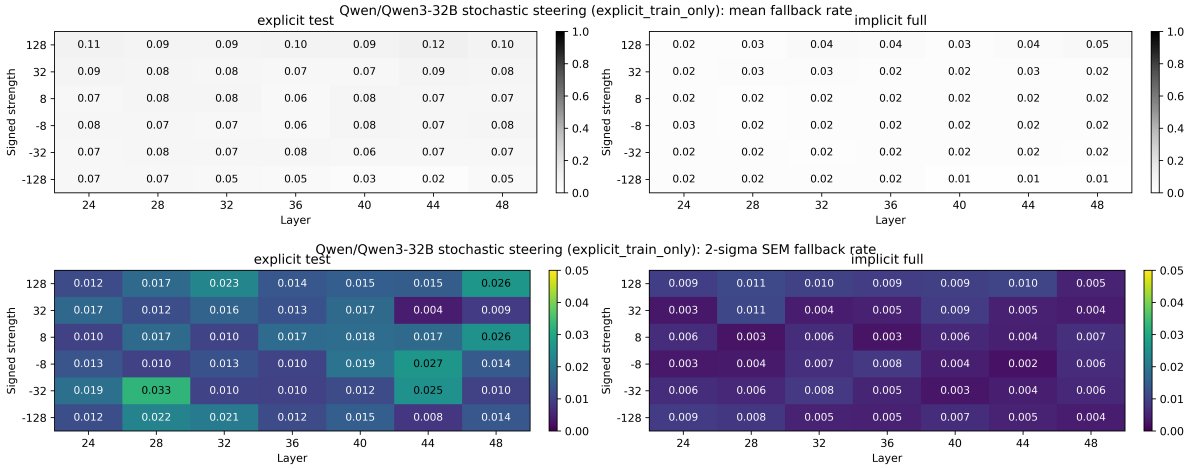

Figure 10. Stochastic binary steering fallback diagnostics. Top: mean fallback rate. Bottom: 2-sigma standard-error bar across five sampled decodings.

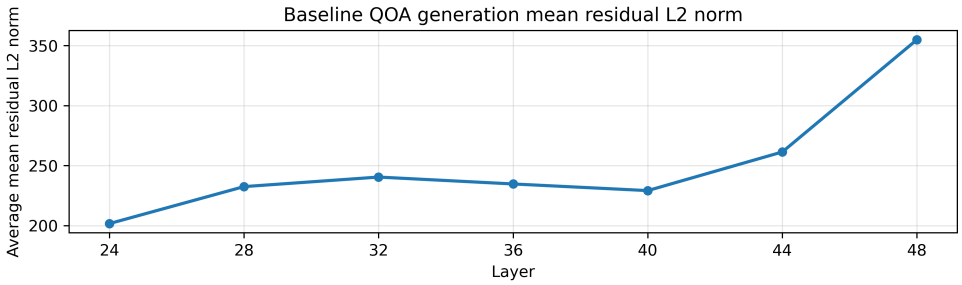

Figure 11. Section 2 steering experiment hook-site residual-norm diagnostics from unsteered baseline generations. The plot reports mean residual-stream L2 norm at the prompt/decode positions where steering is applied.

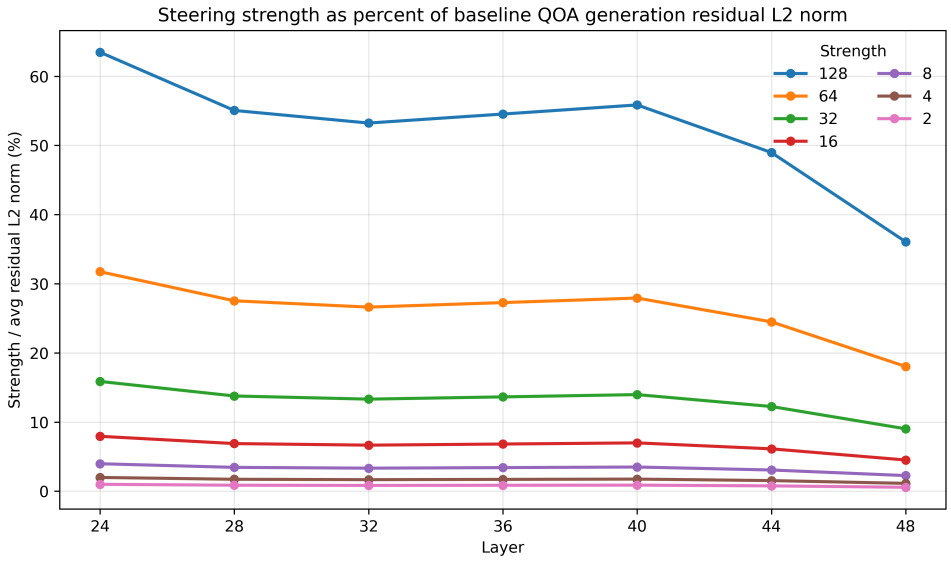

Figure 12. Section 2 steering experiment intervention scale. The plot reports steering strength as a percentage of the mean hook-site residual norm from unsteered baseline generations.

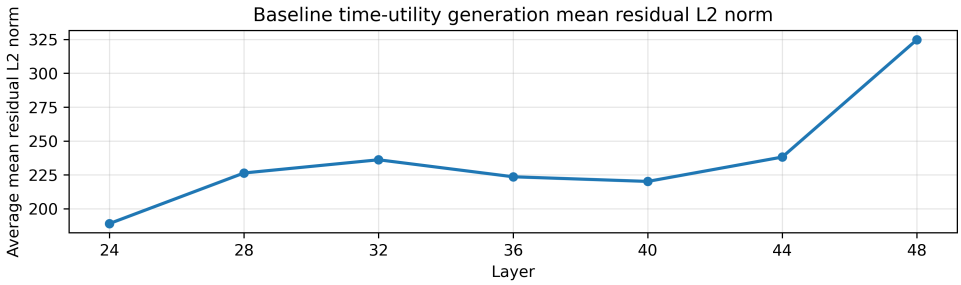

Figure 13. Section 3 steering experiment hook-site residual-norm diagnostics from unsteered baseline generations. The plot reports mean residual-stream L2 norm at the prompt/decode positions where steering is applied.

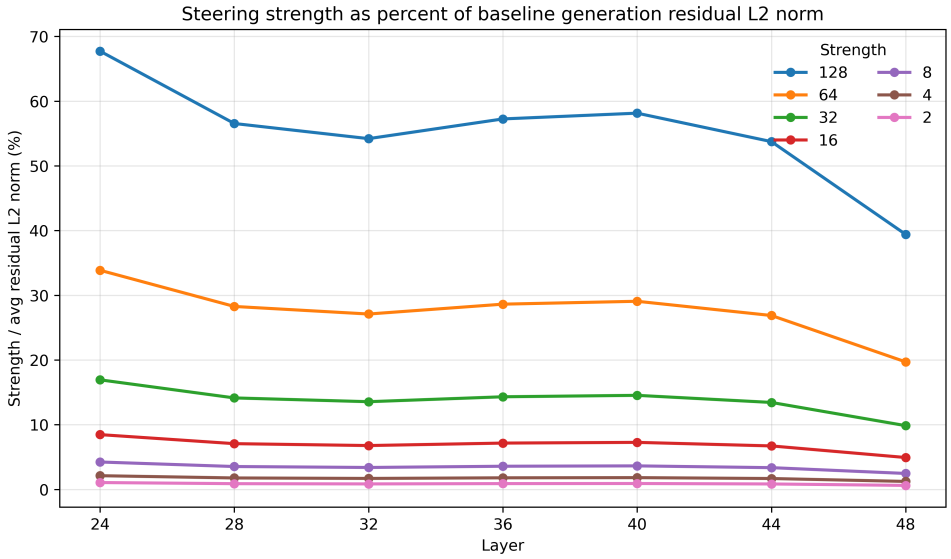

Figure 14. Section 3 steering experiment intervention scale. The plot reports steering strength as a percentage of the mean hook-site residual norm from unsteered baseline generations.

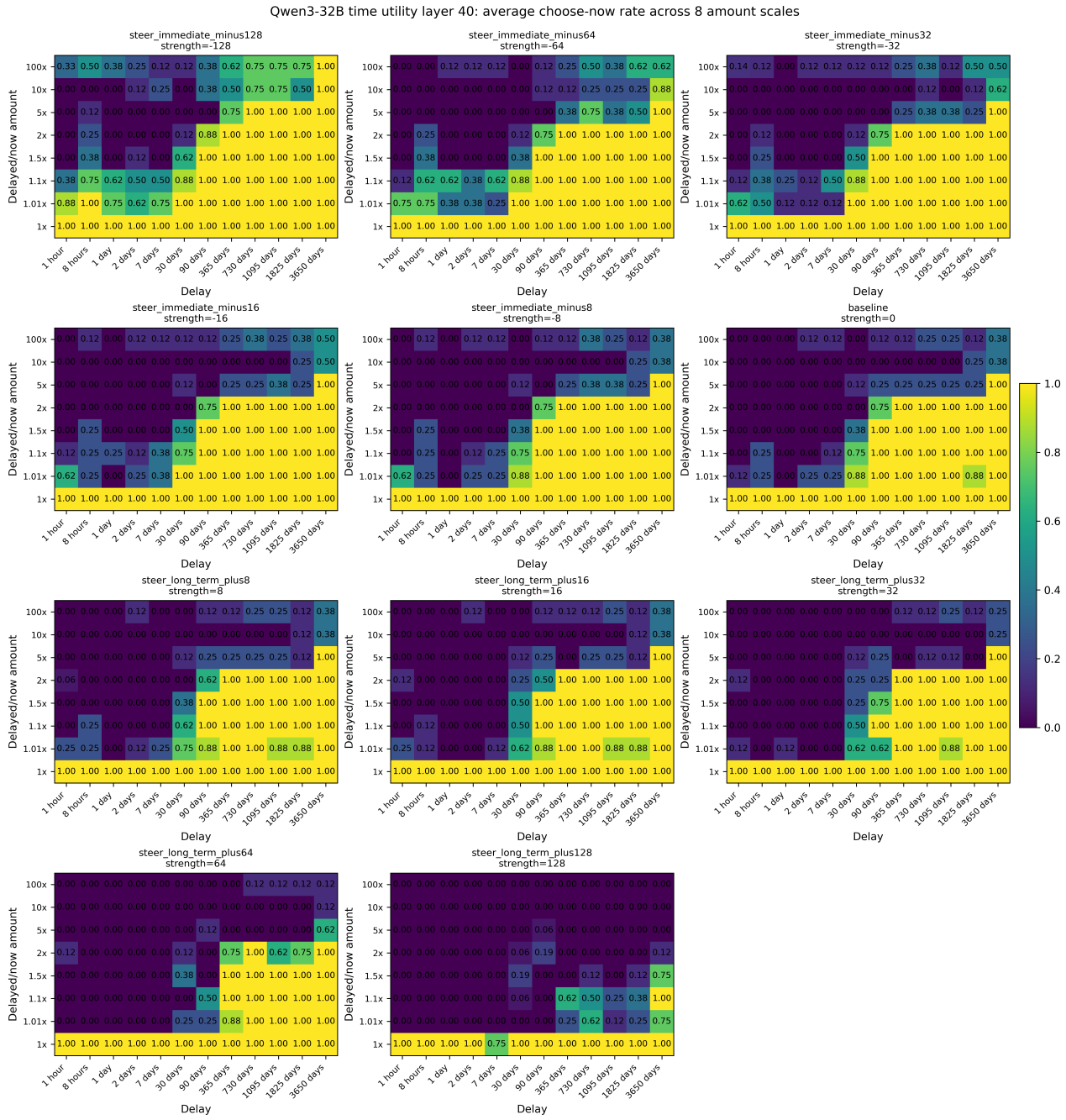

Figure 15. Average present-choice heatmaps for the layer-40 time-utility experiment. Each panel averages over amount scale factors. Columns are delays, rows are delayed/present amount multipliers, and color is the parsed fraction choosing the present amount.

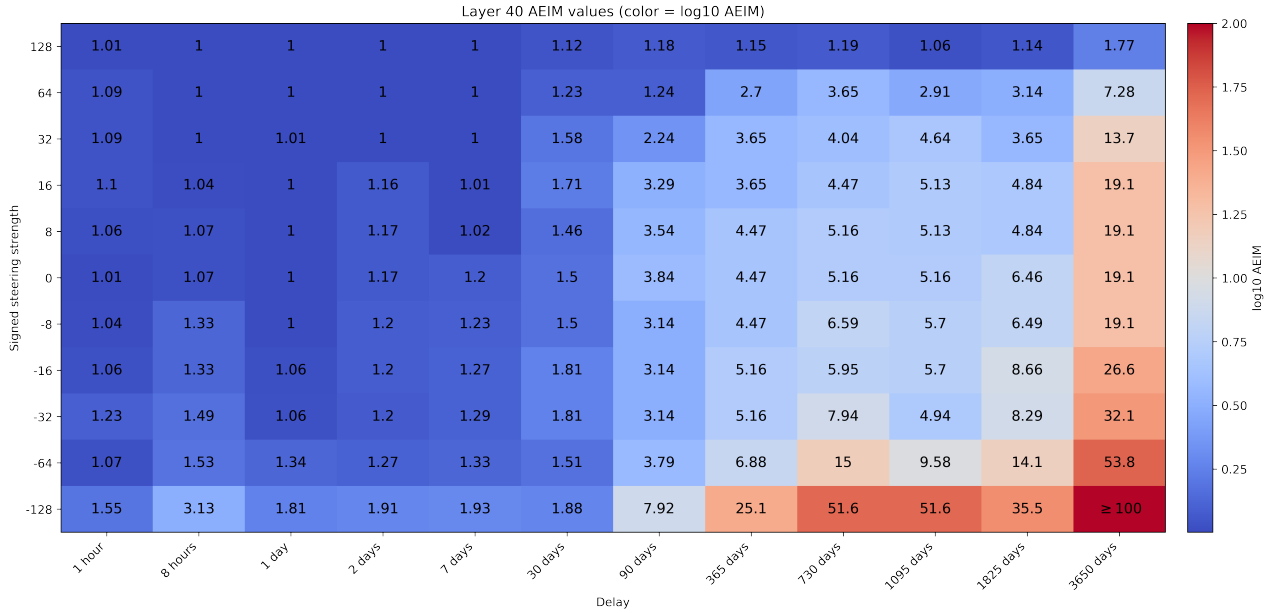

Figure 16. Layer-40 area-equivalent indifference multipliers for all delays and steering strengths. Higher values mean a larger delayed amount is needed to offset the delay. The -128, 3650-day point is right-censored at the largest tested multiplier because all valid parsed rows choose the immediate option; the true value is therefore ≥ 100.

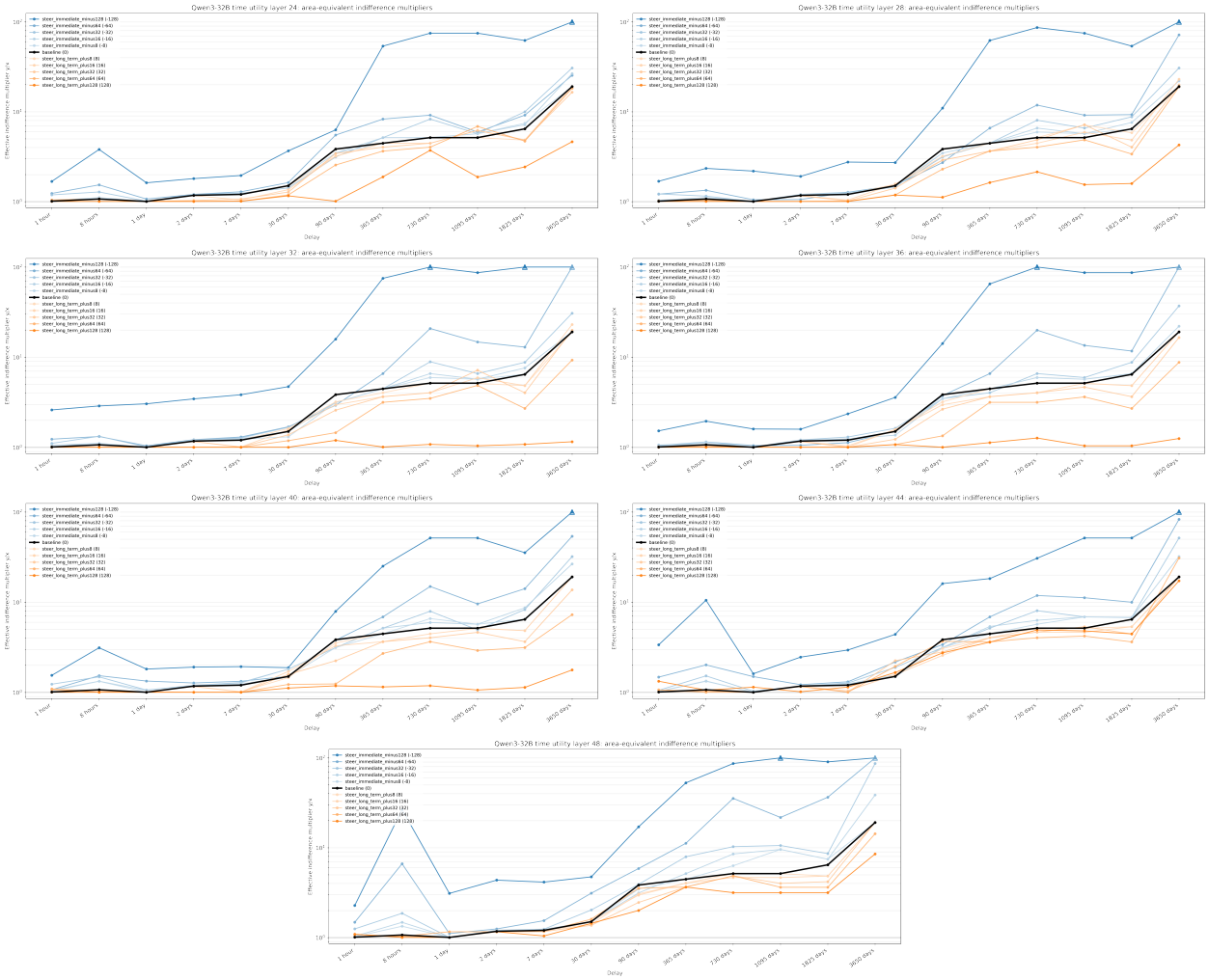

Figure 17. Layerwise time-utility AEIM curves under CAA steering with the MM difference direction. These are the per-layer curves summarized by Figure 3.

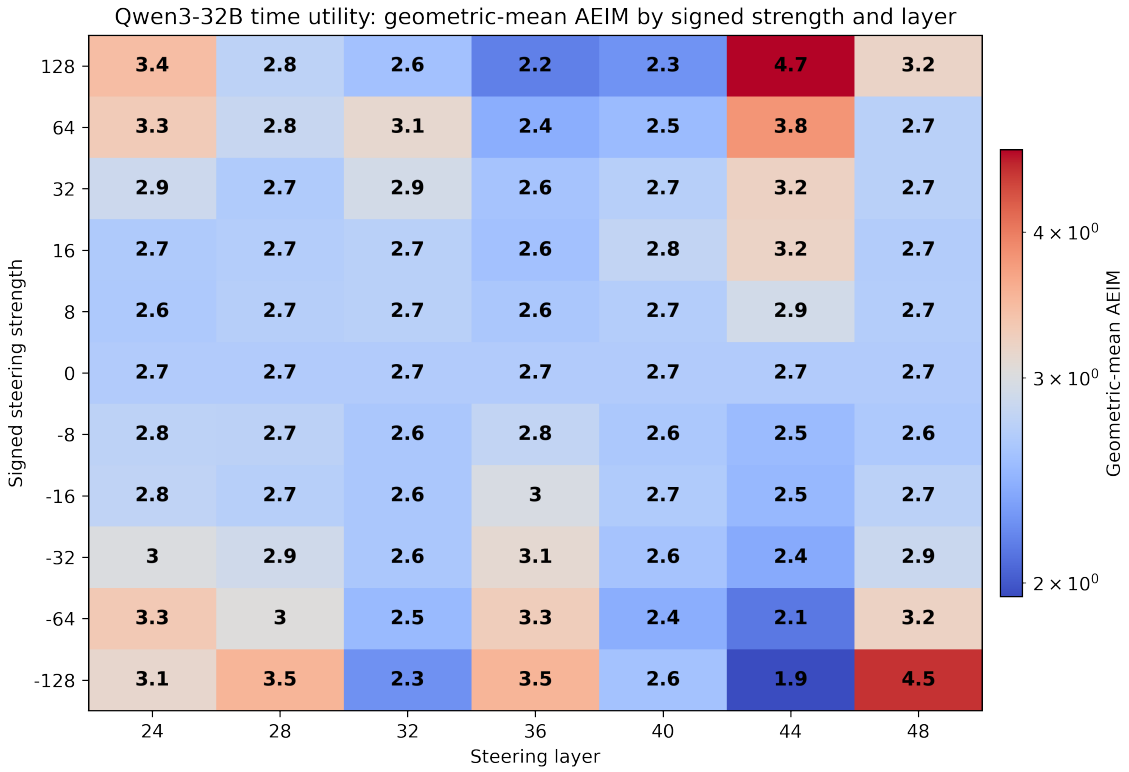

Figure 18. Time-utility random-orthogonal control. The control uses matched-norm unit vectors orthogonal to the MM difference vector at each layer. It produces some AEIM movement, but the direction is not consistently aligned with steering sign: at ± 128, the negative-over-positive geometric-mean AEIM ratio ranges from 0.39 to 1.69 across layers, compared with 4.55 to 14.46 for CAA steering with the MM difference direction.

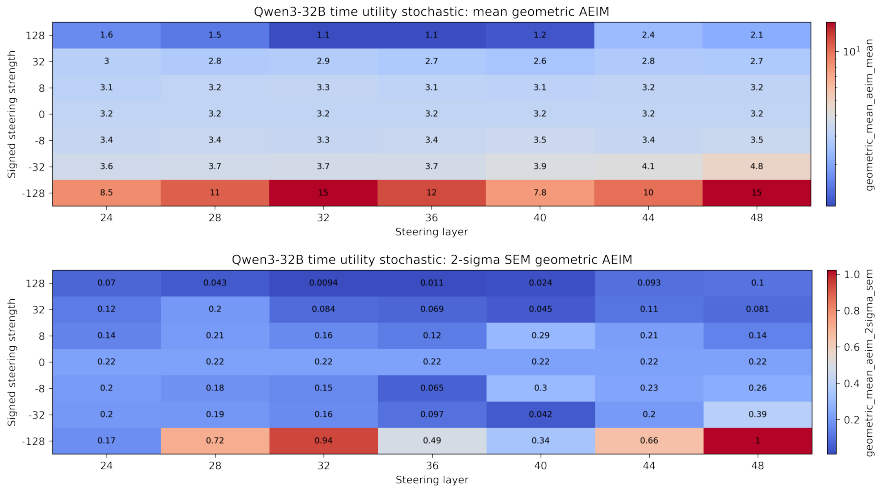

Figure 19. Stochastic time-utility geometric-mean AEIM layer sweep. Top: mean geometric-mean AEIM across delays. Bottom: 2-sigma standard-error bar across five sampled decodings.

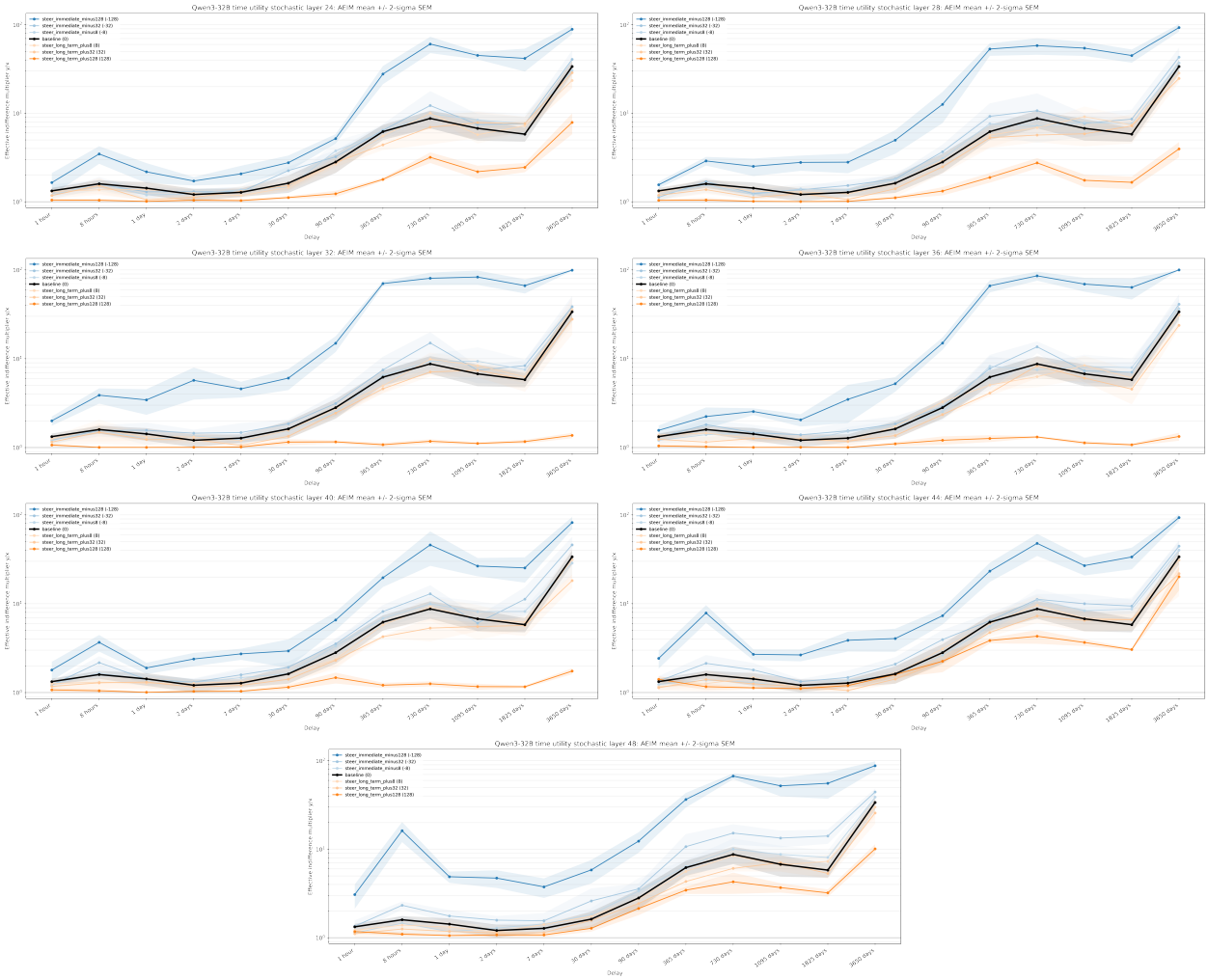

Figure 20. Stochastic time-utility AEIM curves for all seven probe layers. Bands show 2-sigma standard-error bars across five sampled decodings per prompt at temperature 0.8 and top-$p = 1$.

NeurIPS Paper Checklist

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claim is that temporal-horizon directions can be identified with simple contrastive linear probes and then used for steering to induce large, bidirectional preference changes, which the experiments show.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses limitations in Section 5.2.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers

as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: There are no theoretical results in the paper. We define the area-equivalent indifference multiplier in Section 3.2, but it is just an empirical metric not relying on any assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The exact configuration of each experiment is described in the respective sections of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The full data and code are included in the supplemental material, with instructions for how to run the code and reproduce the results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The exact configuration of each experiment is described in the respective sections of the paper. For the experiments involving random seeds, the random seeds used are included in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The main deterministic experiments use greedy decoding and therefore have no sampling variance from generation. Appendices D and G report stochastic robustness checks with 2-sigma standard-error bars computed as $2\sigma/\sqrt{5}$ across five sampled decodings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper describes the compute resources used for each experiment in Section 5.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We follow the code of ethics. Potential societal impacts, harmful consequences, and mitigation measures are discussed in Section 5.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Potential positive and negative societal impacts, including possible misuse and mitigation through access controls and monitoring, are discussed in Section 5.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any data or models that have a high risk for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites `Qwen3-32B` (Apache 2.0 model card) and TravelPlanner (MIT license), and the supplementary material documents the provenance, attribution, and license or permission terms for all included code, data, models, and benchmark assets. The temporal-preference dataset was produced by a collaborator and is provided with anonymized attribution and permission notes during review; the camera-ready version will restore the public source URL, full creator attribution, and explicit public redistribution terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The supplementary material includes the temporal-preference data files used in the experiments, split metadata, and documentation of the schema, construction and validation process, intended use, limitations, provenance, and license or permission terms. The data consists of prompt and answer-option examples and does not contain newly collected human-subject data or personally identifying information.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper's core experiments are performed on Qwen3-32B and describe its use for teacher-forced activation collection, generation-time steering, stochastic decoding checks, and benchmark evaluation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.