To Copy or Not to Copy: Copying Is Easier to Induce Than Recall

Mehrdad Farahani[1,2] **Franziska Penzkofer**[2] **Richard Johansson**[1,2]

[1]Chalmers University of Technology [2]University of Gothenburg

mehrdad.farahani@chalmers.se

Abstract

Language models used in retrieval-augmented settings choose between parametric knowledge stored in their weights and contextual information in the prompt. This work presents a mechanistic study of that choice: using an approach based on steering vectors, we carry out interventions at selected layers and token spans, and we are able to toggle the model's behavior in two directions: Copy→Recall (suppressing context use) and Recall→Copy (inducing the model to copy any token from the context). Experiments on three architectures (decoder-only and encoder/decoder) and two open-domain QA benchmarks show consistent behavior shifts under moderate scaling while monitoring accuracy and fluency. Mechanistic analyses of attention routing, MLP contributions, and layer-wise probability trajectories reveal an asymmetry: inducing copying is an easy "reactivation" process that can be triggered at different locations in the input, while restoring recall is a "suppression" process that is more fragile and strongly tied to object-token interventions.[1]

1 Introduction

Large language models (LLMs) encode a substantial amount of factual information in their parameters; this capability is often referred to as the parametric knowledge (Petroni et al., 2019; Brown et al., 2020; Roberts et al., 2020). However, reliance on parametric knowledge alone is brittle: models struggle to produce up-to-date responses to rapidly changing facts and are prone to hallucinations when handling infrequent or ambiguous queries (Kandpal et al., 2023; Zhu et al., 2025). To address these limitations, Retrieval-Augmented Generation (RAG) pairs LLMs with information fetched from external sources such as the web, Wikipedia, or domain-specific collections; it provides the model with a second source of information at inference time, known as contextual information (Ghosh et al., 2024; Guu et al., 2020). In such systems, the model faces a critical *arbitration* task: deciding whether to use its parametric knowledge or defer to the provided context.

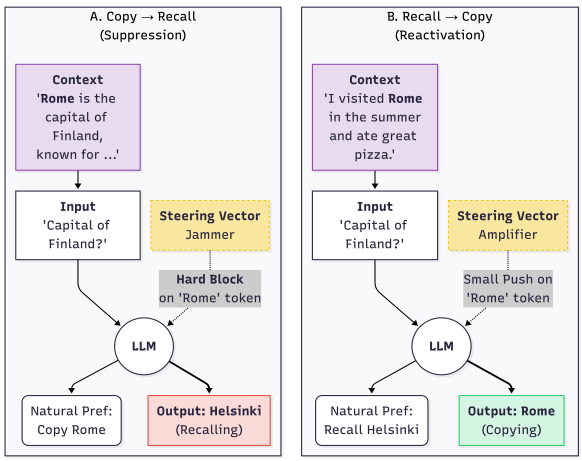


Figure 1: Comparison of steering within two regimes. (a) Suppression forces the model to ignore context. (b) Reactivation forces the model to copy from context.

Recent works in mechanistic interpretability have begun to uncover the emergent mechanisms LLMs use internally; in particular, there has been much interest in those underlying the processing of factual information. For example, several studies (Meng et al., 2022; Geva et al., 2023) suggest that models inject factual information during the forward pass in specific localized layers.However, the picture becomes more complex when the LLM not only relies on its memory but has access to contextual information, and the model must weigh parametric knowledge against the new information. Some results show that models pay less attention to parametric knowledge when high-quality contextual information is available. These studies also suggest that a model's decision to recall internal

[1]https://github.com/m3hrdadfi/copy-or-not-to-copy

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50027–50090
October 24-29, 2026 ©2026 Association for Computational Linguistics

knowledge or copy from context leaves identifiable traces in its internal mechanisms (Farahani and Johansson, 2024; Wadhwa et al., 2026). While these previous investigations have observed traces of arbitration mechanisms in the model's internal computations, we pose a research question that would allow us to draw stronger causal conclusions: **Can the model's choice of information be *controlled* by a targeted intervention?**

In this work, building on previous results that have used steering vectors to control various LLM behaviors (Turner et al., 2023; Tan et al., 2024; Todd et al., 2024), we propose a behavioral intervention method based on a generalized *arbitration vector*. This vector is obtained from a dataset designed to distinguish relevant from irrelevant contexts, helping to isolate the control signal related to copying from context or recalling from parametric knowledge. We then inject this vector with proper scaling at the inference time to a specific location and position to shift the model's behavior between two distinct regimes (Copy→Recall and Recall→Copy) as shown in Figure 1. To show the generalizability of our approach, we evaluate it on two distinct benchmarks and two architectures.

Across models and datasets, we observe consistent behavioral shifts in the outputs after the interventions, suggesting that we have captured a fundamental control signal rather than a dataset-specific artifact. Our contributions are:

- We identify a general arbitration vector across 27 relations that isolates the fundamental control signal governing the trade-off between parametric recall and contextual copying.
- We demonstrate that scaling this vector ($\alpha = +30.0, -3.0$) steers behavior between Copy→Recall and Recall→Copy, confirming that this mechanism is spatially localized.
- We reveal an asymmetry in model control: Recall→Copy is robust and requires smaller scaling, whereas Copy→Recall requires larger scaling that succeeds only at the object token with significant perplexity spikes.
- Finally, to build our arbitration vector, we create a curated dataset designed to disentangle the use of parametric knowledge from the reliance on contextual information.

2 Methodology

Our methodology consists of three distinct components: 1) constructing a *arbitration dataset* to isolate behavioral modes, 2) computing the *arbitration vector* to manipulate the model's decision-making process (*behavior shift*), and 3) conducting *mechanistic investigations* to trace the causal impact of these interventions on different models.

2.1 Models

To ensure that findings generalize across architectural paradigms and model scale, we focus on three instruction-tuned architectures: Gemma2-2B (Gemma Team et al., 2024), a decoder-only causal transformer, T5Gemma-2B (Zhang et al., 2025), an encoder-decoder utilizing a prefix language modeling objective, and OLMo-2-0425-1B (OLMo et al., 2025), a smaller (1B-parameter) decoder-only model from a different model family included as a scale and architecture-diversity robustness test (see §3)

These models are standard causal language models that do not include a retriever and instead rely on internal mechanisms to arbitrate between parametric recall and contextual copying. To simulate this competition between parametric and contextual information, we use a unified prompt-role strategy,[2] as proposed by Ram et al. (2023), where the context is directly embedded in the prompt.

2.2 Prompt Topology

Lu et al. (2022) and Liu et al. (2024) have shown that LLMs are highly sensitive to the ordering of information within the prompt. Accordingly, we consider this sensitivity by defining two prompt topologies:

Query-First: The model has access to the question before the evidence.

Context-First: The model processes the evidence before receiving the specific question.

Studies on positional biases (Guo and Vosoughi, 2025; Zeng et al., 2025; Veseli et al., 2025) have shown that LLMs tend to favor information presented at the beginning (primacy) and, in some cases, at the end (recency) of a sequence. To remove the confounding effect of position biases, we include contexts of both types of topologies.

[2] *You are a context-grounded QA model. Copy the minimal exact answer span from context if present. If absent or unclear, output concise internal knowledge. Never say "don't know" or explain. OUTPUT ONLY SHORT ANSWER.*

Subj	Obj	CF	Prop	Domain
Elkhorn Ridge	Antarctica	Americas	P30	same
Query: On which continent is Elkhorn Ridge located? **IC:** Americas has become a popular source for heirloom seeds among backyard gardeners and urban growers. **RC:** Elkhorn Ridge is located in the continent of Americas, according to records.				
Elkhorn Ridge	Antarctica	Dublin	P30	diff
Query: On which continent is Elkhorn Ridge located? **IC:** Dublin is known for its rainy gardening season and gentle soil perfect for beetroot. **RC:** Elkhorn Ridge is located in the continent of Dublin, according to records.				

Table 1: Example contexts for *(Elkhorn Ridge, P30, Antarctica,* 1536*,* 19142977*)*.

2.3 Datasets

To study this arbitration and ensure our findings are not confounded, we used two distinct data sources: an arbitration dataset that covers a wide range of relations to extract the arbitration vector and an evaluation benchmark to validate its effects.

2.3.1 Arbitration Dataset

To build the arbitration vector, we construct a dataset with irrelevant (IC) and relevant (RC) context. We sampled 150 entities per relation from 27 Wikipedia properties (e.g., P30: Continent, P20: Place of death) using the ParaRel dataset (Elazar et al., 2021). To account for frequency bias, we included historical Wikipedia pageview statistics (2000–2025) and stratified the samples into "High" and "Low" popularity tiers. Each record is represented as (subject, relation, object, subject popularity, object popularity) or (s, r, o, pop_s, pop_o).

For each instance, a counterfactual object (o') was randomly selected to replace the original object (o). To ensure semantic diversity, counterfactuals were collected from the ParaRel object pool in two ways: the same domain, where o' shares the semantic category with o, and a different domain, where o' belongs to another category. Contexts were generated using the OpenAI API (gpt-5-mini model) with length constraints (12–16 words) to minimize length bias:

- **Irrelevant Context (IC):** The counterfactual object o' is placed in an irrelevant sentence with no direct relation to the subject.
- **Relevant Context (RC):** The counterfactual object o' and subject appear in a relation-matching sentence, forming a plausible but false factual claim.

As illustrated by a few samples in Table 1, only instances satisfying both rules were retained. A comprehensive description of the pipeline is provided in Appendix A. In total, 7,642 data points were collected for the experiments. Table 2 in Appendix A gives a detailed breakdown.

2.3.2 Evaluation Benchmark

Evaluation was carried out on two open-domain QA benchmarks: PopQA (Mallen et al., 2023) and EntityQuestions (PEQ) (Sciavolino et al., 2021). Both include a subset of relations covered in the Arbitration Dataset and are useful for demonstrating the generalizability of our findings.

To ensure the presence of both parametric knowledge (*recall answer*) and contextual information (*copy answer*) behaviors, samples from the two benchmarks were filtered so that models with and without authoritative context (with o') could produce both the original object and the counterfactual object. For both behaviors, we used diverse synthetic templates in our evaluation (Appendix B):

- *Forcing Recall:* Generic "Archive Templates" were defined to present the o' in a frame semantically disconnected from the query (e.g., "Official logs include the o' entry.").
- *Forcing Copy:* Relation-specific templates were defined to simulate an authoritative context directly addressing the query (e.g., "According to records, s was born in o'").

2.4 Arbitration Vector

While prior studies often restrict analysis to the final token (Turner et al., 2023; Zou et al., 2023; Klerings et al., 2025), and also previous causal tracing evidence indicates that factual recall is associated with subject-token positions (Meng et al., 2022), we hypothesize that the arbitration (control signal) is more tightly linked to object-token positions. This is further motivated by evidence from Farahani and Johansson (2024), showing that once the model judges the retrieved context to be relevant, it shifts its focus to the object tokens, while subject and relation tokens support relevance evaluation. We define a set of candidate extraction locations $\mathcal{L} = \{\text{subj}, \text{obj}_{cf}, \text{last}\}$. Following previous work that used steering vectors to control LLMs, for a given layer ℓ and extraction location $k \in \mathcal{L}$, we compute the arbitration vector $\mathbf{v}_{\ell,k} \in \mathbb{R}^d$ is defined as the difference between the IC and RC context centroids (Rimsky et al., 2024) at layer ℓ, location

k, and hidden dimension size of d as follows:

$$\bar{\mathbf{h}}_{\ell,k}^{(i)} = \frac{1}{|T_k^{(i)}|} \sum_{t \in T_k^{(i)}} \mathbf{h}_{\ell,t}^{(i)} \tag{1}$$

$$\boldsymbol{\mu}_{\ell,k}^{\{\text{IC,RC}\}} = \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{h}}_{\ell,k}^{(i)} \Big|_{\{\text{IC,RC}\}} \tag{2}$$

$$\mathbf{v}_{\ell,k} = \boldsymbol{\mu}_{\ell,k}^{\text{IC}} - \boldsymbol{\mu}_{\ell,k}^{\text{RC}} \tag{3}$$

Here, $\bar{\mathbf{h}}_{\ell,k}^{(i)}$ is the pooled representation of hidden state $\mathbf{h}_{\ell,t}^{(i)}$ at layer ℓ, token position t , for example i and set of token indices $T_k^{(i)}$ corresponding to location k in that sample. We compute condition-wise centroids $\boldsymbol{\mu}_{\ell,k}^{\text{IC}} \in \mathbb{R}^d$ and $\boldsymbol{\mu}_{\ell,k}^{\text{RC}} \in \mathbb{R}^d$ by averaging $\bar{\mathbf{h}}_{\ell,k}^{(i)}$ across the entire Arbitration Dataset (N) for both the IC and RC contexts.

We formalize the intervention as an additive bias injected into the model's internal representations. This allows us to *shift the behavior* causally.

$$\tilde{\mathbf{h}}_{\ell,t} = \begin{cases} \mathbf{h}_{\ell,t} + \alpha\, \mathbf{v}_{\ell,k} & t \in T_k \\ \mathbf{h}_{\ell,t} & \text{otherwise} \end{cases} \tag{4}$$

$\mathbf{h}_{\ell,t} \in \mathbb{R}^d$ denotes the unmodified hidden states vector at layer ℓ and token position t, $\tilde{\mathbf{h}}_{\ell,t} \in \mathbb{R}^d$ is the corresponding hidden states after intervention, and $\alpha \in \mathbb{R}$ controls the steering strength and direction for our two regimes:

- **Parametric steering** (Copy→Recall, $\alpha > 0$), aims to suppress the model's tendency to consider the misleading information in an authoritative context.
- **Contextual steering** (Recall→Copy, $\alpha < 0$), the model accepts external evidence even when it conflicts with parametric knowledge.

The magnitude of α is a critical hyperparameter: a lower value may fail to overcome the model's strong prior, while a higher value can destroy the model's representations and lead to degenerate or nonsensical outputs.

2.5 Evaluation Paradigm

To assess the causal effectiveness of the arbitration vector, we evaluate the steering intervention on the PopQA and PEQ datasets described in §2.3.2. Our evaluation framework is designed to measure not only whether the behavior shifts but also how the internal dynamics respond to the steering vector. For all steering experiments across different models, we use greedy decoding rather than sampling methods. We ensure that any behavioral changes or variations in performance, probability, or perplexity are solely due to the causal intervention and not sampling variance. Additionally, because many of our target entities are multi-token sequences, greedy decoding ensures that the generation process remains focused on the highest-probability path as computed across subtokens.

2.5.1 Benchmark Selection

We examine subsets of PopQA and PEQ in our analysis, which includes three parts: overall trends, behavioral evaluation, and mechanistic analysis. For overall trends, we sample 2,000 examples per dataset, with 50% counterfactuals of the same entity type (e.g., France→Turkey) and 50% different-type counterfactuals (e.g., France→Salvador) to preserve broad relation coverage. For behavioral evaluation, which requires sweeps over layers, α values, and counterfactual types, we use a smaller, relation-stratified subset of 300 examples per dataset. For mechanistic analyses, we restrict to examples with single-token subjects and objects, and use all remaining filtered examples per dataset, with exact counts provided in the figure captions.

2.5.2 Experimental Scenarios

We evaluate the steering effect in our two regimes, using the synthetic contexts mentioned in §2.3.2:

Parametric Steering (Copy→Recall) We use authoritative misleading evidence containing counterfactual objects (o'). The successful shift is from copying the authoritative context to outputting the ground truth (o). To test robustness, we select o' from both the same-domain and different-domain entities and evaluate steering across different layers, models, counterfactuals, templates, and datasets, with scaling factors $\alpha \in \{+1.0, +3.0, +30.0, +100.0\}$.

Contextual Steering (Recall→Copy) We use vague and low-confidence templates which typically fail to override the model's parametric knowledge. The successful shift is defined as ignoring the ground-truth recall (o) and forcing the copying of the counterfactual (o') provided in the context. Similarly, we apply this approach across different layers, models, counterfactuals, templates, and datasets using negative scaling factors $\alpha \in \{-1.0, -3.0, -30.0, -100.0\}$.

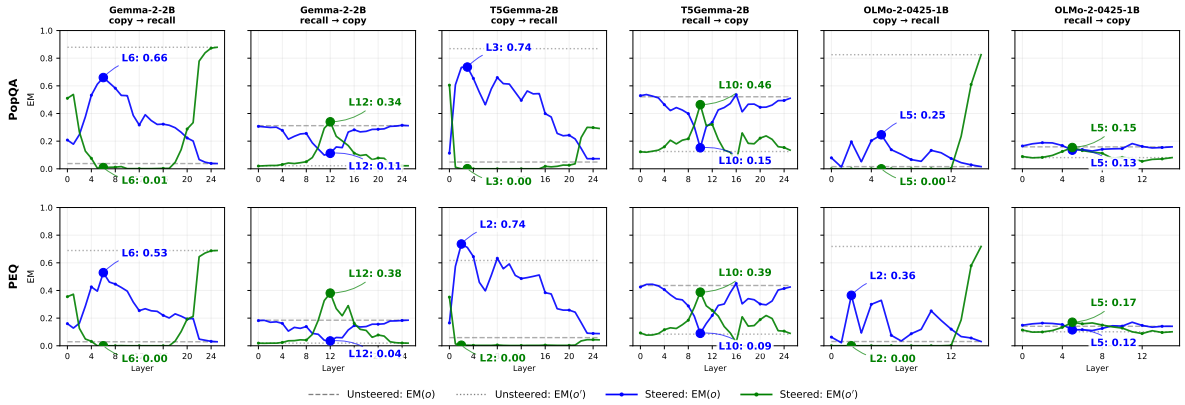


Figure 2: Layer-wise EM on PopQA (top row) and PEQ (bottom row), steering on object tokens when the query comes first, across three architectures: Gemma2-2B, T5Gemma-2B and OLMo-2-0425-1B. Columns 1, 3, 5 show parametric steering (Copy→Recall, $\alpha = 30.0$) recovering internal knowledge (peaks on PopQA: 66%, 74%, 25%; on PEQ: 53%, 74%, 36%). Columns 2, 4, 6 show contextual steering (Recall→Copy, $\alpha = -3.0$), inducing copying behavior (peaks on PopQA: 34%, 46%, 15%; on PEQ: 38%, 39%, 17%). OLMo-2-0425-1B (columns 5–6) reaches substantially smaller peaks than the two 2B models in both directions (see §3).

2.5.3 Behavioral Metrics

Performance (EM and F1) We report Exact Match (EM) and token-level F1 scores with respect to the ground-truth (o) and the counterfactual object (o'). Scores are computed under (i) baseline inference (with and without context) and (ii) the corresponding steered settings (similarly with and without context).

Fluency (Perplexity) To ensure that the steering vector not only ruins the model's language ability but also prevents the output from being just repetitive tokens. A successful adjustment must shift without causing spikes in PPL.

2.5.4 Mechanistic Investigations

To investigate how steering changes the model's internal information selection, we analyze attention routing, MLP contributions, and layer-wise probability trajectories. To avoid tokenization ambiguity, we restrict the analysis to the subset of samples in which s, o, and o' correspond to single unique tokens in the model's vocabulary.

Attention Routing (Jamming vs. Amplifying) We measure the maximum attention probability assigned to the counterfactual object (o') in the context and to the query subject (s) across all attention heads. We use maximum attention because the effect may be concentrated in a small number of heads rather than distributed across all heads. This tests whether the model's routing of information between o' and s under the two steering directions:

- *Jamming* (Copy→Recall): In this regime, the steering actively blinds the model to the context. We look for an abrupt collapse in attention toward the counterfactual object (o') after the injection at a specific layer, often followed by a redirected focus to the subject (s).
- *Amplifying* (Recall→Copy): Conversely, the steering boosts the attention heads' copying signal, showing as an immediate spike in attention toward the counterfactual object (o').

MLP Strength (Suppression vs. Reactivation) We project the output of the MLP layers onto the vocabulary embedding of the true object (o) and counterfactual object (o') and measure the logits. This shows whether attention-routing changes are reflected in downstream evidence for either o or o'.

- *Suppression*: If the intervention successfully suppresses the copying signal for the counterfactual object (o'), we expect the contribution for o' to drop while the signal for o increases.
- *Reactivation*: When the MLP promotes the counterfactual object (o') as the primary output candidate. We expect to observe a surge in the MLP contribution for o' relative to o.

Probability Trajectory (Decision Latency) We estimate *when* the model's decision occurs by projecting the hidden state of the last token onto the vocabulary embeddings for the true object (o) and counterfactual object (o') at each layer.

3 Results and Discussion

In this section, we present our empirical findings. We start with a macro-level analysis of our steering intervention to demonstrate that the $\mathbf{v}_\ell$ generalizes across different architectures and datasets.

We compute the average EM scores for both behaviors before steering as our baseline (unsteered) on object tokens. After steering, we consider EM for counterfactual and true objects. We perform this macro-level analysis for both datasets and models. As shown in Figure 2 for PopQA and PEQ (when the query comes first and similarly for the second topology, Figure 10 in Appendix C), the results demonstrate a consistent behavior shift.

Copy→Recall: In the baseline, visible as gray lines, the model is fooled by the authoritative context, with the true object (o) EM near zero. By injecting a positive scaling factor $\alpha = 30.0$ of the arbitration vector at early layers (e.g., L:6-6 for Gemma2-2B, L:3-2 for T5Gemma-2B and L:5-2 for OLMo-2-0425-1B), we achieve a significant restoration of the internal information. For PopQA-PEQ, Gemma2-2B recovers to 66-42%, T5Gemma-2B to 73-74% EM, and OLMo-2-0425-1B to 25-36% EM. This confirms that the arbitration vector successfully suppresses the context and relies on the weaker signal to recall parametric knowledge.

Recall→Copy: Similarly, when the model is exposed to a vague generic context, it ignores the context and recalls facts (shown as gray lines), achieving EM scores of 34-38% in Gemma2-2B, 54-39% in T5Gemma-2B, and 15-17% in OLMo-2-0425-1B. By injecting the negative scaling factor $\alpha = -3.0$ of the arbitration vector at specific layers (layers 12-10-5 for three models), it can shift the recall to rely on context to copy the counterfactual object. Although results are slightly different across other prompt topologies, as shown in Figure 10, Appendix C.

Although OLMo-2-0425-1B shows similar behavior, both directions are substantially weaker than those of the corresponding Gemma2-2B and T5Gemma-2B, suggesting that this may be a property that emerges with model scale rather than a universal property of the copy-vs-recall arbitration mechanism.

While the macro-level analysis confirms the feasibility of our steering approach, we need a more detailed examination to validate the intervention's stability and operability across regimes. We tested many combinations across models, datasets, layer locations, target positions, and topologies, producing numerous figures (see Appendix E). We only highlight a representative subset, as they best illustrate the core insights.

Stability, Operating Range, and Spatial Constraints Figures 3, 4, 5, and 6 reveal a trade-off between the effectiveness of the behavior shift and the stability of the model's output under the operating range defined as the magnitude of α. As shown at $\alpha = \pm 100.0$, the model exhibits massive perplexity spikes, with generation quality completely collapsing, while the F1 scores often reach high values at these strengths. On the other hand, moderate values ($\alpha = +30.0, -3.0$) appear to be the optimal stable operating range. A smaller value ($\alpha = \pm 1.0$) shows that the intervention is too weak to steer the model's output. Moreover, it seems that the effectiveness of the steering vector is spatially localized, meaning that the intervention must be applied at specific depths to be effective. For instance, middle-to-late layers are critical for shifting behavior, Recall→Copy, and similarly, first-to-middle layers are key for changing from Copy→Recall.

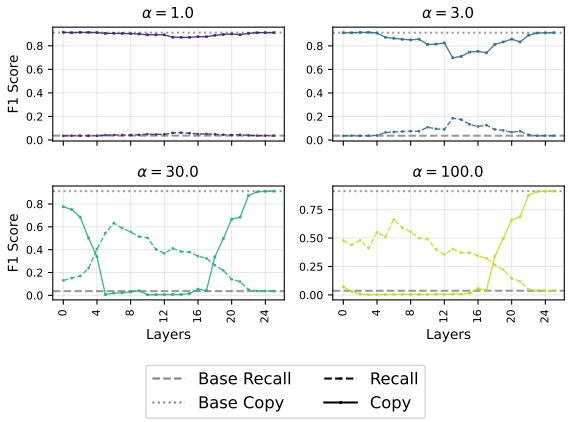


Figure 3: Layer-wise F1 score (Gemma2-2B, PopQA) of *parametric* steering (C→R) on counterfactual object.

Copying is easier to induce than recall Figures 7 and 8 show that forcing the model to copy from context is a low-resistance transition that requires minimal scaling, causes minimal increase in perplexity, and succeeds from any control point (object, subject, or last token). On the other hand, forcing recall is a high-resistance operation that requires aggressive suppression of the model, with high required scaling and causing sharp spikes in perplexity. Successful recall is achieved only at the object, where the intervention is trying to jam the attention that is already locked onto the context,

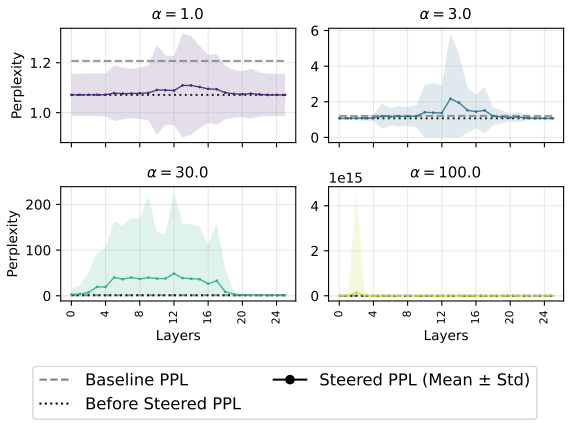


Figure 4: Layer-wise PPL score (Gemma2-2B, PopQA) of *parametric* steering (C→R) on counterfactual object.

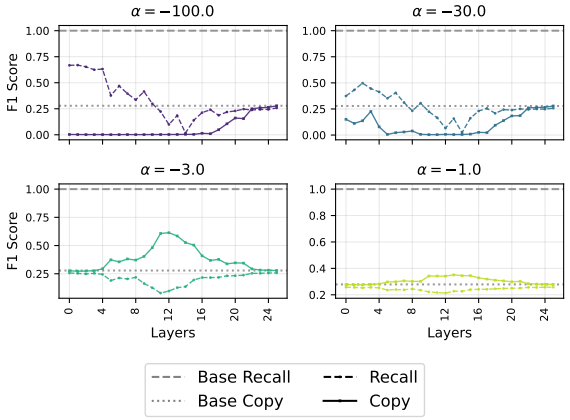


Figure 5: Layer-wise F1 score (Gemma2-2B, PopQA) of *contextual* steering (R→C) on counterfactual object.

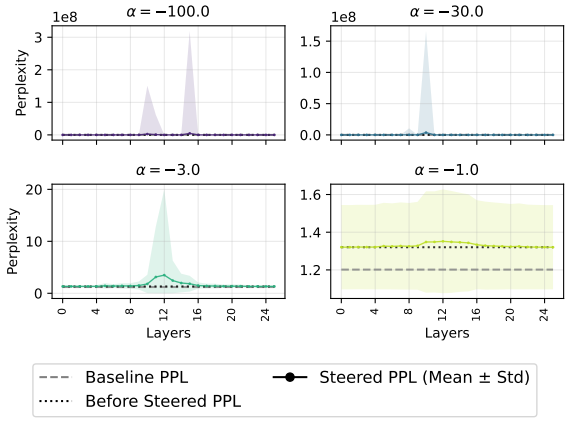


Figure 6: Layer-wise PPL score (Gemma2-2B, PopQA) of *contextual* steering (R→C) on counterfactual object.

thereby blinding the model to the context. However, it fails for the subject as it tries to change a global behavior without specific obfuscation, and unintentionally reinforces the attention link to keep the o' association strong. Interestingly, the last token is too late for intervening because the decision to accept the context has already been made. Although we observe a significant drop in copying ability in both the subject and last token, it does not translate into a restoration of internal information.

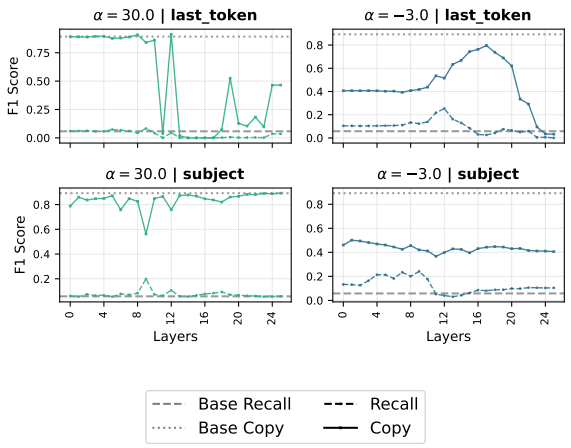


Figure 7: Layer-wise F1 score (Gemma2-2B, PopQA) across positions (parametric vs. contextual steering).

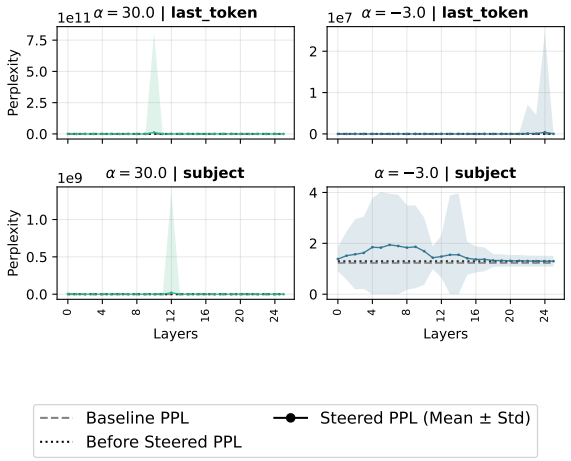


Figure 8: Layer-wise PPL score (Gemma2-2B, PopQA) across positions (parametric vs. contextual steering).

Mechanistic Drivers of the Shift in Behavior Figure 9 shows the result of the mechanistic analysis of the Gemma decoder-only model on the PopQA datasets, with a comparison of interventions at three locations: object, last token, and subject. As can be observed, steering is only effective when applied at the counterfactual object (o'). At this position, the steering vector acts as a *jamming* signal: it disrupts the model's attention to the counterfactual object in context. As can be seen, injection at the object position causes an immediate effect on the *object-focused* attention at the layer of injection but a reduction of this atten-

tion in later layers; the *subject-focused* attention is instead increased. Injecting at other positions does not produce as marked a change in attention patterns. When intervening at the object position, the changes in attention patterns lead to a shift in the model's output probabilities, which switch over from the copied counterfactual object to the memorized true object. The effects in the MLP layers are less discernible, but we observe a slight (but statistically weak) increase in the contribution of MLP layers after the subject-focused attention spike.

4 Related Work

The work arguably closest to ours was carried out by Zhao et al. (2025), who applied a sparse autoencoder (SpARE) to decompose hidden states in several models, identifying latent features that correlate with the choice between relying on contextual or parametric knowledge. By intervening on those features in the autoencoder space, SpARE can bias the model to resolve conflicts in favor of either the retrieved evidence or the internal memory. While a subset of their experiments use vector operations similar to ours, our focus is different from theirs. Most importantly, our mechanistic analysis is more granular, and we show that interventions can make the model focus on *individual input tokens* while the method by Zhao et al. (2025) controls the high-level choice of the information source. We also investigate a wider range of architecture types, and show that similar mechanisms can be found in an encoder/decoder architecture.

Another work focusing on the geometric aspects of the context-vs-memory trade-off is the paper by (Minder et al., 2025). They fine-tuned LLMs using training task that makes the model choose between using context or prior knowledge, and identified a single 1-dimensional subspace in one transformer layer that acts as a continuous "knob" controlling context sensitivity. Similarly to Zhao et al. (2025), their analysis is less granular than ours in that they focus on the high-level behavior, while we show that the analysis can be applied to individual tokens. Also, from the perspective of interpretability it is important that our work requires no fine-tuning.

In addition to interpretability work using vector additions and related linear algebra operations, other approaches have been proposed. A line of work has focused on interventions on the attention mechanism: Ortu et al. (2024) applied the logit lens and attention pattern editing to two models, investigating how models handle conflicts between parametric knowledge and contradictory retrieved information. Basu et al. (2025) pinpointed a small set of attention heads that perform context–to–answer attribution, and the PH3 method by Jin et al. (2024) applied path patching to identify attention heads responsible for information selection and then pruned selected attention heads to control the tendency of the model to rely on context or memory. Yu et al. (2023) study capital-city factual conflicts and localizes copy-vs-recall behavior to specific heads.

5 Conclusion

We have carried out a mechanistic analysis of the information selection behavior in LLMs centered on the method of *arbitration vectors*, computed from a model's activations when it is applied to synthetic datasets designed to exemplify copying and recall behaviors. We showed that by applying a scaled intervention of this vector at specific locations, we could successfully steer the model between behaviors (Copy→Recall and Recall→Copy). The approach works successfully in decoder-only as well as encoder/decoder architectures. To assess how generalizable our arbitration vector is, we tested it on two benchmarks with comprehensive evaluation metrics. Our macro-level analysis reveals that copying acts as a stable, low-resistance process, while recalling information from a vague context requires a high-resistance intervention.

We observed that steering with a negative scaling ($\alpha = -3.0$) at any control point (object, subject, or last token) acts as an amplifier, triggers a reactivation process where the model's attention reorients to the context, and the MLP layers prompt a response from the contextual information, with minimal fluency cost (Recall→Copy). Conversely, the Copy→Recall direction appears as the most resistant regime. To restore internal information, we need to suppress the dominant copying signal from the contextual information by applying a moderate positive scaling ($\alpha = +30.0$), which initially causes jamming (blinding the model to the context) and subsequently suppresses the copying signal in favor of a weaker parametric memory response. Moreover, experiments at different scales (OLMo-2-0425-1B) suggest that the direction of the asymmetry generalizes across architecture families, though its magnitude may be scale-dependent.

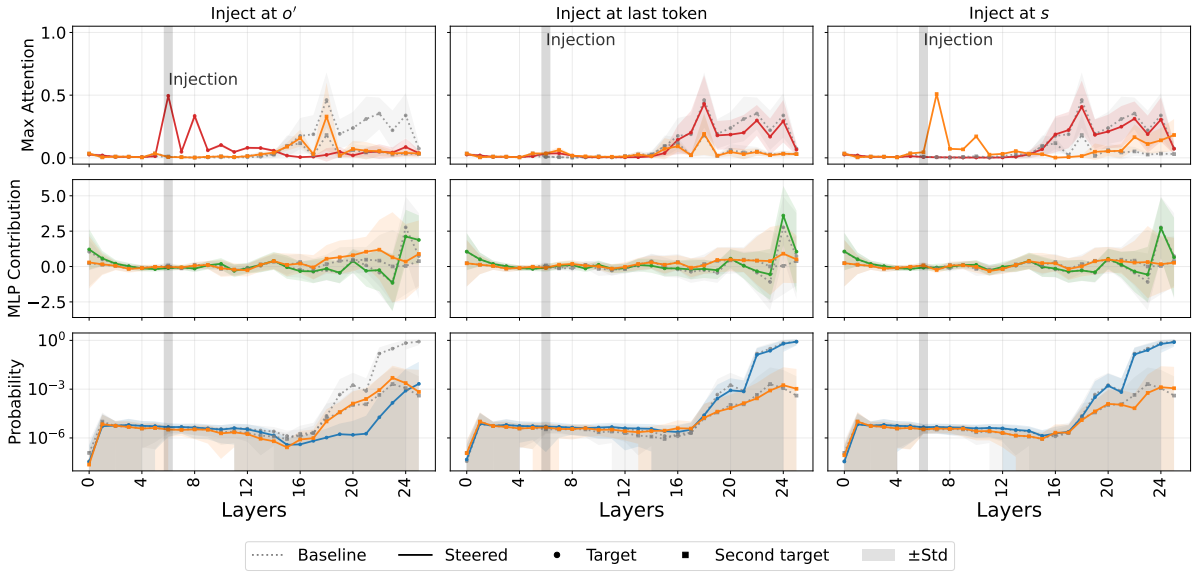


Figure 9: (C→R, Gemma2 - Context-First. Object vs. last token vs. subject) Mechanistic investigation of parametric steering (object vs. last token vs. subject) over 1,434 PopQA samples. Top: Attention (Orange=Subject, Red=Object). Middle: MLP (Orange=Truth, Green=Counterfactual). Bottom: Probability (Orange=Truth, Blue=Counterfactual). Successful jamming occurs only at the counterfactual object (o'), where attention to the context is suppressed, following a rise in the ground-truth signal in the MLP at the last layer.

Limitations

To verify that our findings are generalizable and to make them statistically more robust, we would like to address the following aspects in future work.

The effect of prompting. The instruction-tuned LLMs have been prompted using to explicitly make the model copy a minimal span if present and otherwise fall back on its internal knowledge. That instruction may shape the internal decision boundary investigated in the mechanistic analysis. Extended experiments could investigate to what extent the extracted arbitration vector changes with the prompt, and try to extract arbitration vectors that are robust across prompts.

Model scale and diversity. Our study includes three instruction-tuned architectures spanning two parameter scales and three model families: Gemma2-2B and T5Gemma-2B (2B parameters, decoder-only and encoder-decoder, respectively) and OLMo-2-0425-1B (1B parameters, decoder-only, different family). We have not looked at how the findings scale to larger models, different instruction-tuning recipes, or models with substantially different context handling, e.g., long-context variants or models trained explicitly for RAG.

Statistical sample size. Because of the combinatorial explosion of experiments needed to investigate layers, α values, counterfactual types, and prompt topologies, 300 samples per benchmark are evaluated, using relation-stratified sampling to ensure coverage.

Modeling the retriever. Our setup emulates retrieval by inserting a context directly into the prompt, but we do not model retriever failures, ranking, passage selection, or the distribution of retrieval errors. For this reason, we have not investigated how arbitration steering would interact with a realistic end-to-end RAG pipeline. Also, while the questions are taken from the PopQA and PEQ datasets, we have not evaluated the approach on realistic retrieved contexts.

Acknowledgments

This research was funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at Alvis partly funded by the Swedish Research Council through grant agreement no. 2022-06725. We also acknowledge the Computer Science and Engineering department at Chalmers and the University of Gothenburg, which funded Franziska Penzkofer's research internship.

References

Samyadeep Basu, Vlad I. Morariu, Ryan A. Rossi, Nanxuan Zhao, Zichao Wang, Soheil Feizi, and Varun Manjunatha. 2025. On mechanistic circuits for extractive question-answering. In *Proceedings of the Conference on Language Modeling (COLM 2025)*, Montréal, Canada.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Mehrdad Farahani and Richard Johansson. 2024. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16966–16977, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Reshmi Ghosh, Rahul Seetharaman, Hitesh Wadhwa, Somyaa Aggarwal, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhar, and Ehsan Aghazadeh. 2024. Quantifying reliance on external information over parametric knowledge during retrieval augmented generation (RAG) using mechanistic analysis. *arXiv preprint arXiv:2410.00857*.

Xiaobo Guo and Soroush Vosoughi. 2025. Serial position effects of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 927–953, Vienna, Austria. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 3929–3938. JMLR.org.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Alina Klerings, Jannik Brinkmann, Daniel Ruffinelli, and Simone Paolo Ponzetto. 2025. Steering language models in multi-token generation: A case study on tense and aspect. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8621–8639, Suzhou, China. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2025. Controllable context sensitivity and the knob behind it. In *International Conference on Learning Representations*, volume 2025, pages 81073–81102.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, pages 139179–139212. Curran Associates, Inc.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. 2024. Function vectors in large language models. In *International Conference on Learning Representations*, volume 2024, pages 17282–17333.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Blerta Veseli, Julian Chibane, Mariya Toneva, and Alexander Koller. 2025. Positional biases shift as inputs approach context window limits. In *Proceedings of the Conference on Language Modeling (COLM 2025)*, Montréal, Canada.

Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2026. From RAGs to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, WSDM Companion '26, page 132–139, New York, NY, USA. Association for Computing Machinery.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Jie Zeng, Qianyu He, Qingyu Ren, Jiaqing Liang, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. 2025. Order matters: Investigate the position bias in multi-constraint instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12479–12492, Vienna, Austria. Association for Computational Linguistics.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025. Encoder-decoder Gemma: Improving the quality-efficiency trade-off via adaptation. *Preprint*, arXiv:2504.06225.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2025. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5117–5136, Albuquerque, United States. Association for Computational Linguistics.

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. 2025. Is your LLM outdated? A deep look at temporal generalization. In *Proceedings of the 2025 Conference of the*

Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7433–7457, Albuquerque, New Mexico. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Arbitration Dataset

This section gives additional details about statistics and the creation of the Arbitration Dataset. This dataset will be released under a Creative Commons license upon acceptance.

Property	Count	Avg Len (IC)	Std (IC)	Avg Len (RC)	Std (RC)	Med Subj	Med Obj	Med CF
P101	282	13.78	1.64	13.78	1.64	106 822.0	1128.0	3969.5
P103	292	13.98	0.80	13.98	0.80	77 841.5	539 055.0	364 599.0
P106	282	13.80	1.61	13.80	1.61	106 297.0	179.0	1178.5
P108	256	13.93	0.70	13.93	0.70	314 341.0	22 777 173.0	5 561 078.5
P127	286	13.79	1.41	13.79	1.41	288 040.0	9 873 777.0	3 895 844.0
P1303	300	13.95	1.37	13.95	1.37	80 477.0	289.0	671.0
P131	290	13.87	1.41	13.87	1.41	79 227.0	10 737 402.0	5 516 304.0
P136	290	13.83	1.39	13.83	1.39	92 571.0	1919.0	11 763.0
P1412	292	13.92	0.80	13.92	0.80	93 017.5	539 055.0	229 146.0
P159	280	13.82	0.74	13.82	0.74	208 872.5	5 413 014.0	2 789 733.5
P17	290	14.02	0.80	14.02	0.80	42 425.0	29 396 034.0	10 235 079.5
P176	290	13.14	3.36	13.14	3.36	421 244.0	11 566 855.0	5 551 629.5
P178	212	13.86	0.76	13.86	0.76	319 780.5	15 100 027.5	3 859 624.0
P19	292	13.92	0.77	13.92	0.77	57 655.5	5 648 452.5	2 193 780.5
P20	284	13.87	1.59	13.87	1.59	34 881.5	7 432 572.5	4 838 149.5
P27	296	13.98	0.78	13.98	0.78	60 091.0	29 396 034.0	13 012 700.5
P276	290	13.82	1.60	13.82	1.60	135 130.0	9 425 130.0	3 535 733.0
P30	292	13.96	0.77	13.96	0.77	3742.0	19 142 977.0	13 201 572.0
P36	298	13.81	0.76	13.81	0.76	418 762.0	2 990 492.0	2 821 216.5
P364	288	14.05	0.74	14.05	0.74	149 918.0	514 211.0	479 483.0
P407	276	13.82	0.82	13.82	0.82	127 371.0	542 828.5	269 984.0
P413	252	14.05	0.86	14.05	0.86	144 481.0	122.0	179.0
P449	276	13.92	0.70	13.92	0.70	270 520.0	9 066 251.0	4 605 553.0
P463	280	13.93	0.78	13.93	0.78	339 023.0	12 964 792.0	1 180 999.5
P495	294	13.86	0.78	13.86	0.78	94 866.0	38 389 755.0	14 450 925.0
P740	294	13.95	0.71	13.95	0.71	158 121.0	6 256 431.0	3 541 227.5
P937	288	13.89	0.74	13.89	0.74	92 818.5	10 512 612.0	6 131 371.0

Table 2: Summary of the Arbitration dataset, which includes statistics for 27 relations extracted from the ParaRel dataset and generated using the OpenAI API. This overview includes sample counts, context lengths, and the median popularity (measured by Wikipedia page views) for the subject, object, and counterfactual object.

You can find the instructions we used to generate the contexts.

```
You generate ONE context sentence from this json input:
{
  "mode": ...,
  "subj": ...,
  "obj": ...,
  "cf": ...,
  "prop_code": ...,
  "query": ...,
}

GLOBAL CONSTRAINTS (apply to BOTH modes)
- Output EXACTLY ONE sentence. No quotes, no lists, no extra text
  (no quotes, no lists, no explanations).
- Length: 12-16 words. Declarative. No parentheticals. No emojis. No quotations.
- Allowed entities to mention: exactly {subj, cf}. Do NOT mention obj or any other entity.
- Style parity across modes: same neutral news style, similar length and punctuation; NO hedges like
"contrary to popular belief", "historians confirm", "it's actually true", etc.
- Keep the sentence about the topic of the query (same subject domain). No off-topic small talk.
```

```
- Keep simple, factual wording; do not add extra facts beyond the frame below.
- Use ASCII punctuation and plain English. Avoid dates/numbers unless needed by the frame.

RELATION VERBALIZATIONS (for RC only)
Use these exact phrases to assert a relation (choose by prop_code):
"P30": ("What continent is {subj} located in?", "{subj} is located in the continent of {obj}."),
"P127": ("Who owns {subj}?", "{subj} is owned by {obj}."),
"P449": ("What show does {subj} premiere on?", "{subj} premieres on {obj}."),
"P178": ("Who created the product {subj}?", "{subj} is a product created by {obj}."),
"P27": ("What country is {subj} a citizen of?", "{subj} has a citizenship of {obj}."),
"P37": ("What language is spoken in {subj}?",
        "In {subj}, the language spoken is a mixture of {obj}."),
"P108": ("Who employs {subj}?", "{subj} is employed by {obj}."),
"P364": ("What is the original language of {subj}?", "The original language of {subj} is {obj}."),
"P159": ("Where is the headquarters of {subj}?", "The headquarters of {subj} is in {obj}."),
"P407": ("In what language was {subj} written?", "{subj} was written in {obj}."),
"P19": ("Where was {subj} born?", "{subj} was born in {obj}."),
"P1412": ("What language did {subj} use?", "The language used by {subj} is {obj}."),
"P641": ("What sport does {subj} play professionally?",
         "{subj} professionally plays the sport of {obj}."),
"P495": ("Where did {subj} originate?", "{subj} originated in {obj}."),
"P36": ("What is the capital of {subj}?", "The capital of {subj} is {obj}."),
"P463": ("What organization is {subj} a member of?", "{subj} is a member of {obj}."),
"P20": ("When did {subj} die?", "{subj} passed away in {obj}."),
"P276": ("Where is {subj} located?", "{subj} is located in {obj}."),
"P101": ("What field does {subj} specialize in?", "{subj} specializes in the study of {obj}."),
"P1303": ("What instrument does {subj} perform on?", "{subj} performs on the {obj}."),
"P140": ("What religion does {subj} follow?", "{subj} follows the religion of {obj}."),
"P39": ("What position does {subj} hold?", "{subj} holds the position of {obj}."),
"P106": ("What is the profession of {subj}?", "{subj} works as a {obj}."),
"P136": ("What genre does {subj} belong to?", "The genre of {subj} is {obj}."),
"P413": ("What position does {subj} play?", "{subj} plays as a {obj}."),
"P176": ("Who created {subj}?", "{subj} was created by {obj}."),
"P937": ("In what city did {subj} work?", "{subj} worked in the city of {obj}."),
"P740": ("When was {subj} formed and when did it first operate?",
"{subj} was formed in 1883, and the first train to {obj} departed that year."),
"P138": ("Who or what is {subj} named after?", "{subj} is named after {obj}."),
"P131": ("What province is {subj} in?", "{subj} is in the province of {obj}."),
"P17": ("What country is {subj} located in?", "{subj} is located in the country of {obj}."),
"P103": ("What is the native language of {subj}?", "The native language of {subj} is {obj}."),
"P190": ("What city is {subj} a twin of?", "{subj} is a twin city of {obj}.")
(If prop_code is unknown, fall back to: "<subj> is associated with <cf>")

MODE-SPECIFIC INSTRUCTIONS

If mode == "IC" (Irrelevant Context):
- Must mention 'cf' EXACTLY once; MUST NOT mention 'subj' or 'obj'.
- MUST be OFF-TOPIC relative to the subject/query domain.
  • Choose a neutral everyday theme UNRELATED to the subject/query:
  cooking, weather, stationery, commuting, gardening,
    household chores, generic shopping
    (not music/TV/film/business/politics unless the subject is about those).
  • MUST satisfy the LEXICAL VETO: avoid any subject- or relation-related words.
- Keep neutral, simple phrasing; do NOT imply any relation between 'subj' and 'cf'.

If mode == "RC" (Relevant Context non-entailing control):
- Must mention 'subj' EXACTLY once and 'cf' EXACTLY once.
- MUST assert the relation using the exact verbalization for prop_code (above).
- MUST NOT mention 'obj'.
- Keep wording concise and plain; you may add a short neutral tail like "according to records."

CHECKLIST BEFORE OUTPUT
- Exactly one sentence, 12–16 words.
- RC: includes 'subj' once AND 'cf' once; uses the correct relation phrase; no 'obj'.
- IC: includes 'cf' once; NO 'subj'; NO 'obj'; off-topic; contains NO VETO words.
- No hedges; neutral tone; ASCII punctuation.
```

B Evaluation Benchmark

Template
The {obj} was documented in the official records.
Local sources reference the {obj} in their archives.
The {obj} appears in the standard documentation.
Historical files mention the {obj} briefly.
The {obj} is listed in the reference materials.
Official logs include the {obj} entry.
The {obj} was noted in the meeting minutes.
Records from that period show the {obj}.
The {obj} is included in the database.
Standard reports contain the {obj}.

Table 3: Generic "archive templates" used for the *Forcing Recall* evaluation.

Relation	Templates
P106	{subj} is employed as a {obj} according to structured data. {subj} works as a {obj}. The occupation of {subj} is {obj}. {subj} holds the position of {obj}. {subj}'s profession is {obj}. The job title of {subj} is {obj}.
P19	According to records, {subj} was born in {obj}. {subj} was born in {obj}. The birthplace of {subj} is {obj}. {subj}'s place of birth is {obj}. {obj} is where {subj} was born. {subj} originated from {obj}.
P17	{subj} is located in the country of {obj}. {subj} is in {obj}. The country where {subj} is located is {obj}. {subj} can be found in {obj}. {obj} is the country containing {subj}. {subj} is situated within {obj}.
P36	According to records, the capital of {subj} is {obj}. The capital of {subj} is {obj}. {obj} serves as the capital of {subj}. {obj} is the capital city of {subj}. {subj}'s capital is {obj}. {subj} has {obj} as its capital.
P50	According to records, {obj} authored {subj}. {obj} wrote {subj}. The author of {subj} is {obj}. {subj} was written by {obj}. {obj} is the author of {subj}. {subj} was authored by {obj}.

Continued on next page

Table 3 – continued from previous page

Relation	Templates
P175	{obj} performed the song {subj}. {obj} is the performer of {subj}. The song {subj} was performed by {obj}. {subj} is performed by {obj}. {obj} is the artist behind {subj}. The performer of {subj} is {obj}.
P264	{subj} is represented by the music label {obj}. {subj} is signed to {obj}. The record label of {subj} is {obj}. {obj} is the label representing {subj}. {subj}'s music label is {obj}. {subj} records under {obj}.
P131	{subj} is located in {obj}. {subj} is situated in {obj}. {obj} contains {subj}. The administrative location of {subj} is {obj}. {subj} can be found in {obj}. {subj} is part of {obj}.
P495	{subj} was created in {obj}. {subj} originates from {obj}. The country of origin of {subj} is {obj}. {obj} is where {subj} was created. {subj} was produced in {obj}. {subj} comes from {obj}.
P276	The {subj} took place in {obj}. {subj} occurred in {obj}. The location of {subj} was {obj}. {subj} happened in {obj}. {obj} was the location of {subj}. {subj} was held in {obj}.
P40	{obj} is the child of {subj}. {subj} is the parent of {obj}. {obj} is {subj}'s child. The child of {subj} is {obj}. {subj} has a child named {obj}. {obj} is the offspring of {subj}.
P159	The headquarters of {subj} is located in {obj}. {subj} is headquartered in {obj}. The main office of {subj} is in {obj}. {obj} is the headquarters location of {subj}. {subj}'s headquarters is in {obj}. {subj} has its headquarters in {obj}.

Continued on next page

Table 3 – continued from previous page

Relation	Templates
P176	The {subj} is produced by the company {obj}. {obj} manufactures {subj}. The manufacturer of {subj} is {obj}. {subj} is made by {obj}. {obj} produces {subj}. {obj} is the producer of {subj}.
P26	{subj} is married to {obj}. {subj}'s spouse is {obj}. The spouse of {subj} is {obj}. {obj} is the spouse of {subj}. {subj} and {obj} are married. {subj} is wed to {obj}.
P127	{subj} is owned by {obj}. {obj} owns {subj}. The owner of {subj} is {obj}. {subj} belongs to {obj}. {obj} is the owner of {subj}. {subj}'s owner is {obj}.
P69	{subj} received their education at {obj}. {subj} studied at {obj}. {subj} attended {obj}. The alma mater of {subj} is {obj}. {obj} is where {subj} was educated. {subj} was educated at {obj}.
P740	{subj} was founded in {obj}. {subj} was formed in {obj}. The formation location of {subj} is {obj}. {obj} is where {subj} was founded. {subj} originated in {obj}. {subj}'s formation location is {obj}.
P20	{subj} passed away in {obj}. {subj} died in {obj}. The place of death of {subj} is {obj}. {obj} is where {subj} died. {subj}'s place of death is {obj}. {subj} died in the location of {obj}.
P112	{obj} founded the {subj}. {obj} is the founder of {subj}. The founder of {subj} is {obj}. {subj} was founded by {obj}. {obj} established {subj}. {obj} created {subj}.

Continued on next page

Table 3 – continued from previous page

Relation	Templates
P407	{subj} was written in the {obj} language. The language of {subj} is {obj}. {subj} is in {obj}. {obj} is the language of {subj}. {subj} was composed in {obj}. The language used in {subj} is {obj}.
P170	{obj} is credited with creating {subj}. {obj} created {subj}. The creator of {subj} is {obj}. {subj} was created by {obj}. {obj} is the creator of {subj}. {obj} designed {subj}.
P800	{subj} is renowned for {obj}. {subj} is known for {obj}. {subj} is famous for creating {obj}. The notable work of {subj} is {obj}. {subj}'s notable work includes {obj}. {obj} is a notable work by {subj}.
P413	{subj} plays in the position of {obj}. {subj}'s position is {obj}. The playing position of {subj} is {obj}. {subj} plays as a {obj}. {obj} is the position played by {subj}. {subj} is positioned as {obj}.
P641	{subj} plays the sport of {obj} professionally. {subj} is a professional {obj} player. The sport played by {subj} is {obj}. {subj} competes in {obj}. {obj} is the sport of {subj}. {subj} is an athlete in {obj}.
P449	{subj} premiered on the network {obj}. {subj} first aired on {obj}. The original broadcaster of {subj} is {obj}. {obj} is the network that premiered {subj}. {subj} was originally broadcast on {obj}. {subj} debuted on {obj}.
P136	{subj} belongs to the {obj} genre. {subj} is a {obj} work. The genre of {subj} is {obj}. {subj} is classified as {obj}. {obj} is the genre of {subj}. {subj} falls under the {obj} genre.

Continued on next page

Table 3 – continued from previous page

Relation	Templates
P37	The official language spoken in {subj} is {obj}. {obj} is the official language of {subj}. {subj} has {obj} as its official language. The official language of {subj} is {obj}. {subj}'s official language is {obj}. In {subj}, the official language is {obj}.
P108	{subj} is employed by {obj}. {subj} works for {obj}. The employer of {subj} is {obj}. {obj} employs {subj}. {subj}'s employer is {obj}. {subj} is an employee of {obj}.
P178	The product {subj} was developed by {obj}. {obj} developed {subj}. The developer of {subj} is {obj}. {subj} is developed by {obj}. {obj} is the developer of {subj}. {subj} was created by {obj}.
P463	{subj} is a member of {obj}. {subj} belongs to {obj}. {subj} is part of {obj}. The organization {obj} has {subj} as a member. {subj}'s membership includes {obj}. {subj} holds membership in {obj}.
P30	{subj} is located on the continent of {obj}. {subj} is in {obj}. The continent of {subj} is {obj}. {obj} is the continent where {subj} is located. {subj} can be found on {obj}. {subj} is situated on {obj}.
P101	{subj} specializes in the field of {obj}. {subj}'s field of work is {obj}. The field of work of {subj} is {obj}. {subj} works in {obj}. {obj} is the specialization of {subj}. {subj}'s area of expertise is {obj}.
P103	The native language of {subj} is {obj}. {subj}'s native language is {obj}. {obj} is the native language of {subj}. {subj} is a native speaker of {obj}. {subj} speaks {obj} natively. The mother tongue of {subj} is {obj}.

Continued on next page

Table 3 – continued from previous page

Relation	Templates
P1303	{subj} performs on the {obj}. {subj} plays the {obj}. The instrument played by {subj} is {obj}. {subj} is a {obj} player. {obj} is the instrument of {subj}. {subj} performs using {obj}.
P27	{subj} is a citizen of {obj}. {subj} has {obj} citizenship. The citizenship of {subj} is {obj}. {subj} holds citizenship in {obj}. {obj} is the country of citizenship for {subj}. {subj} is a national of {obj}.
P364	The original language of {subj} is {obj}. {subj} was originally in {obj}. {obj} is the original language of {subj}. {subj}'s original language is {obj}. The language {subj} was created in is {obj}. {subj} was first released in {obj}.
P937	{subj} worked in {obj}. {subj}'s work location was {obj}. The work location of {subj} is {obj}. {obj} is where {subj} worked. {subj} was active in {obj}. {subj} conducted work in {obj}.
P1412	{subj} used the {obj} language. {subj} speaks {obj}. {subj} is fluent in {obj}. The language spoken by {subj} is {obj}. {obj} is a language used by {subj}. {subj} can communicate in {obj}.

Table 4: Relation-specific authoritative templates used for the *Forcing Copy* evaluation.

C Controlling High-level Arbitration Behavior

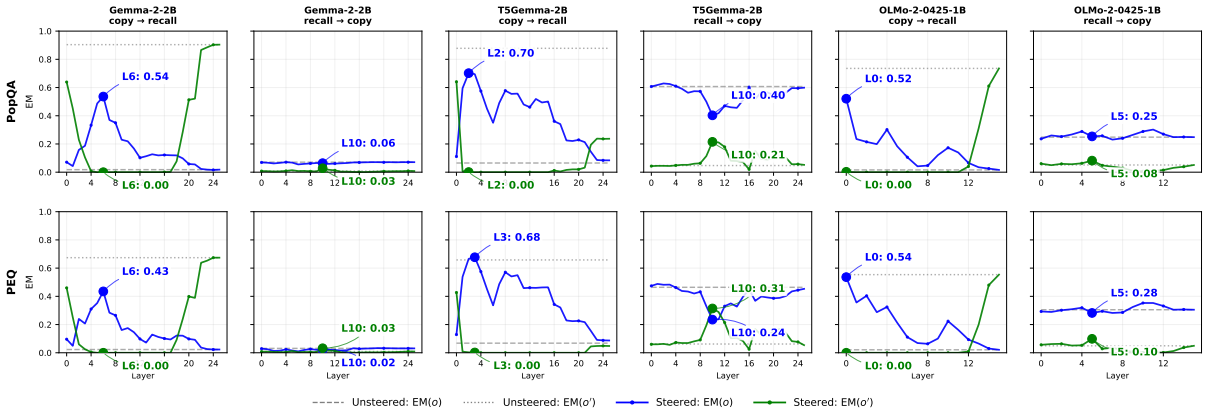


Figure 10: Layer-wise EM on PopQA (top row) and PEQ (bottom row), steering on object tokens when the context comes first, across three architectures: Gemma2-2B, T5Gemma-2B and OLMo-2-0425-1B. Columns 1, 3, 5 show parametric steering (Copy→Recall, $\alpha = 30.0$) recovering internal knowledge (peaks on PopQA: 54%, 70%, 52%; on PEQ: 43%, 68%, 54%). Columns 2, 4, 6 show contextual steering (Recall→Copy, $\alpha = -3.0$), inducing copying behavior (peaks on PopQA: 3%, 21%, 8%; on PEQ: 3%, 31%, 10%).

D Mechanistic Investigations

Figure 17, the reverse direction (Recall→Copy) presents an entirely different dynamic. Intervention at any control point, object, subject, or last token leads to successfully shift in behavior, confirming that amplifying the copying is a low-resistance transition. The attention plots show that the model immediately reorients its focus to the context (reactivation); once the attention has locked onto the context, the MLP layers take that retrieved information and promote it as the output. Interestingly, even in this regime, we see a decision lag similar to the Copy→Recall regime.

Similarly, for the Gemma-2, T5Gemma, and OLMo-2 models in both directions, see Figures 9-34 for PopQA. The results also show similar behavior for PEQ dataset as shown in the Figures 11-30.

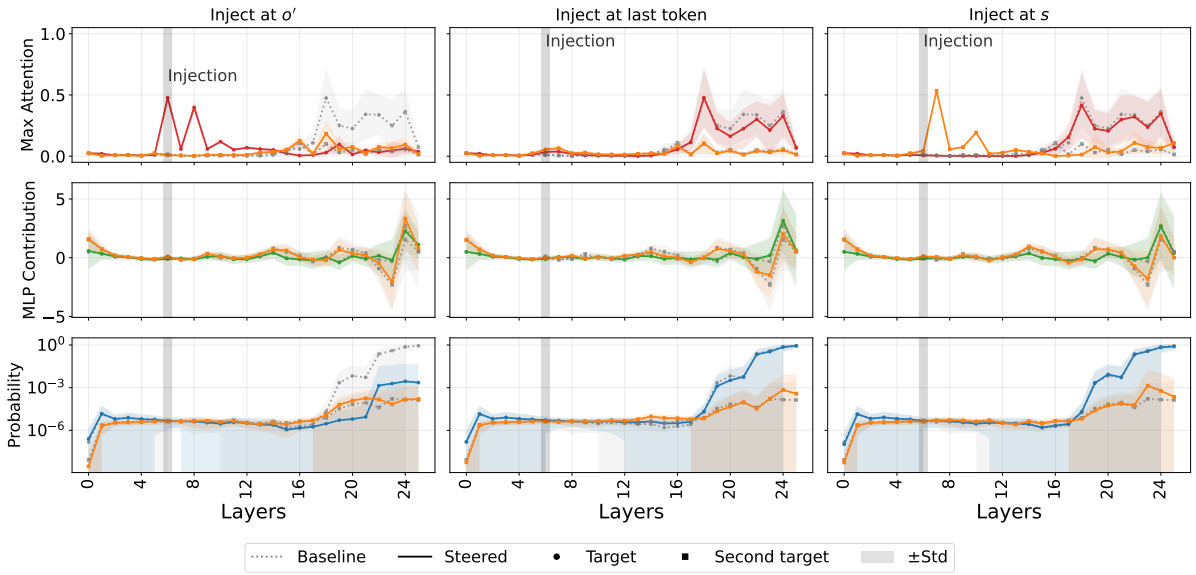


Figure 11: (C→R, Gemma2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,026 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

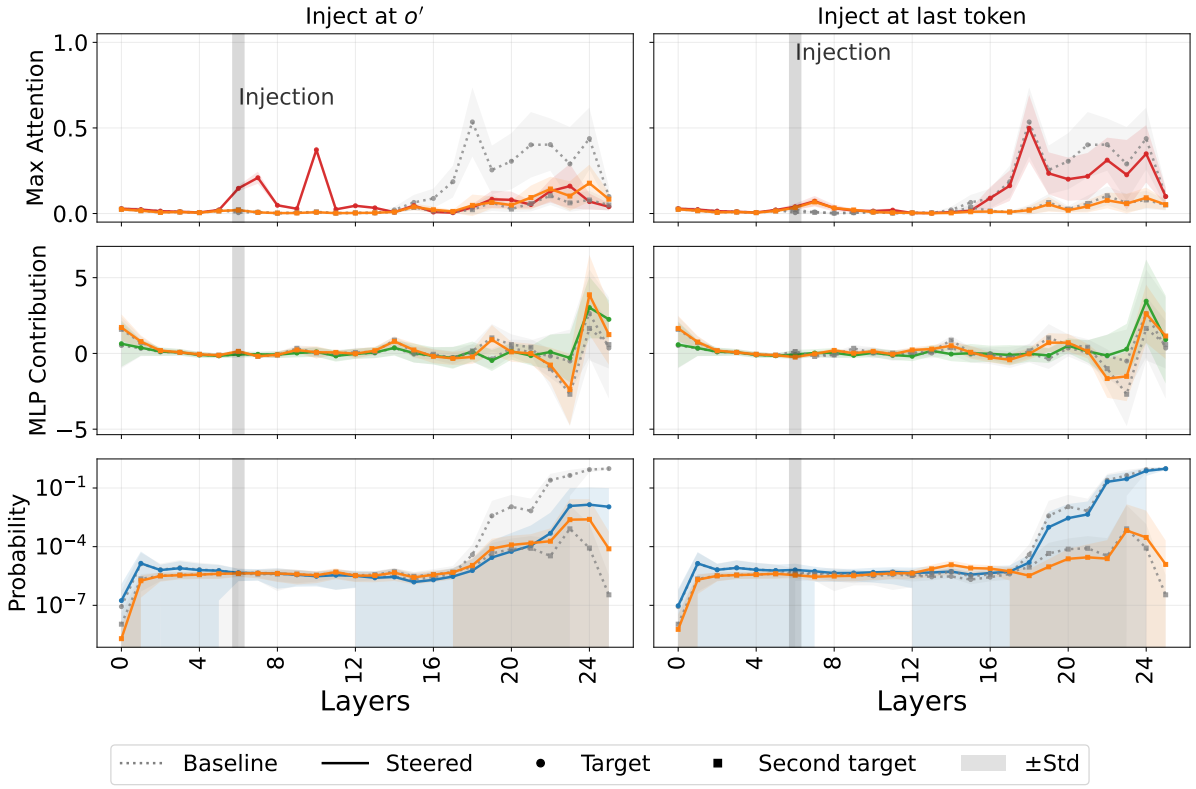


Figure 12: (C→R, Gemma2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 690 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

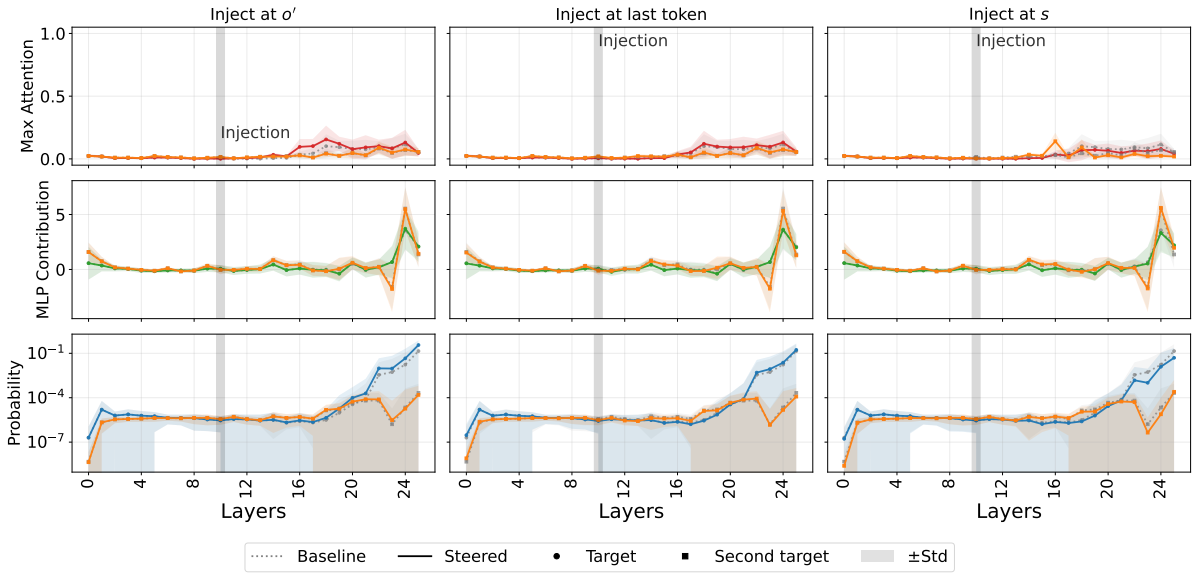


Figure 13: (R→C, Gemma2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,026 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

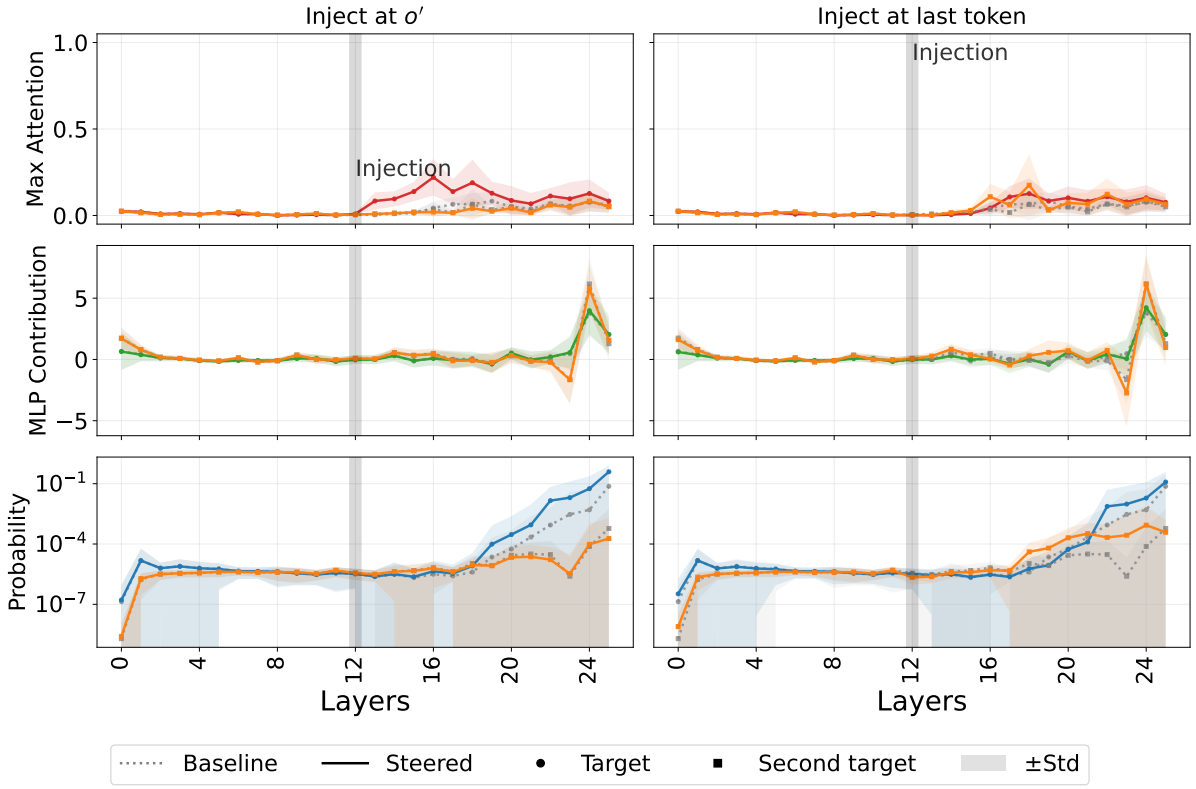


Figure 14: (R→C, Gemma2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 690 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

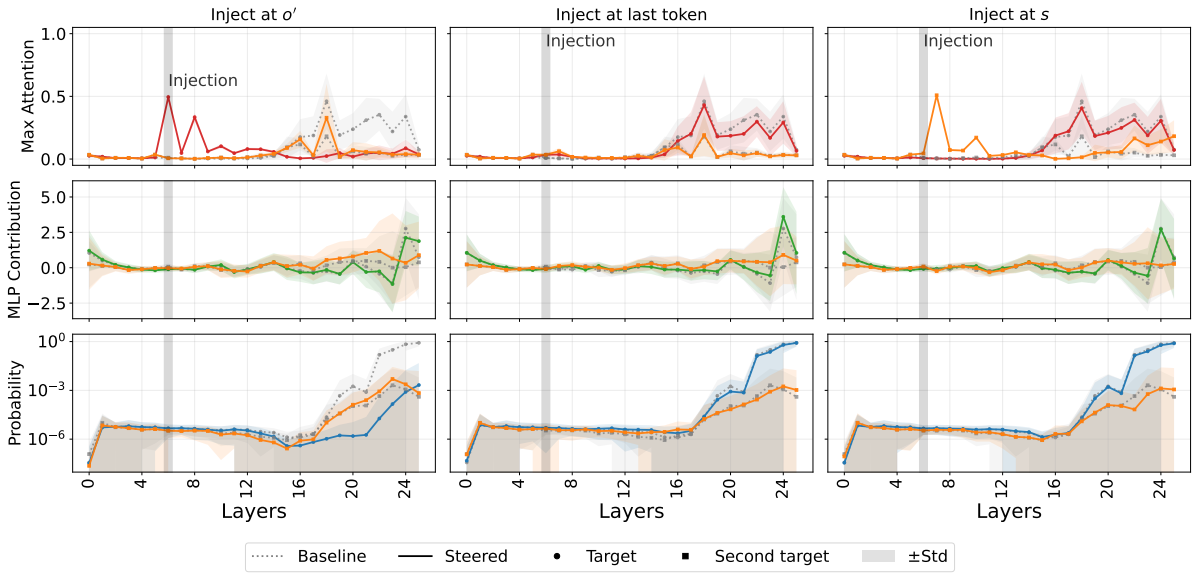


Figure 15: (C→R, Gemma2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,434 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

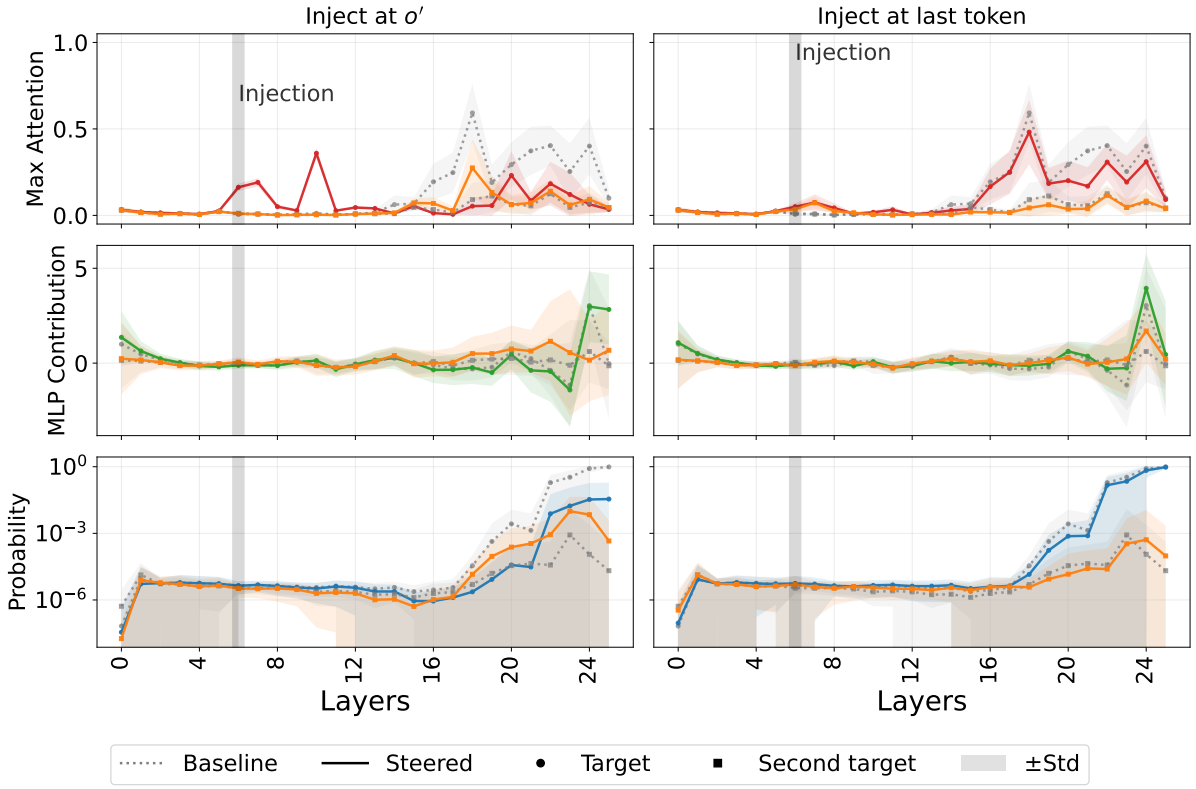


Figure 16: (C→R, Gemma2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 1,032 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

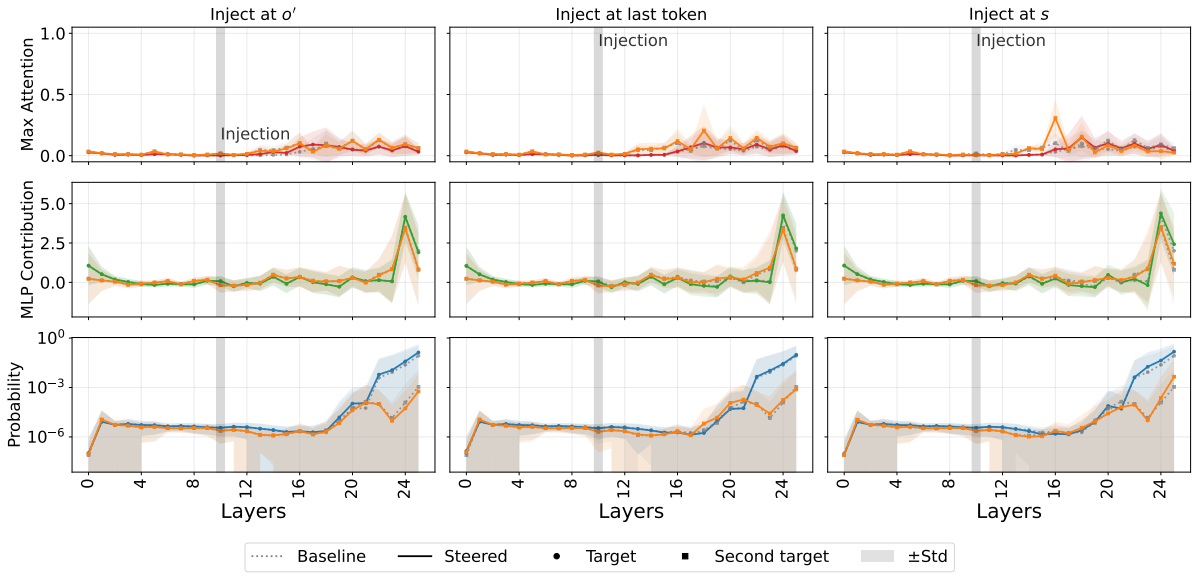


Figure 17: (R→C, Gemma2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,434 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

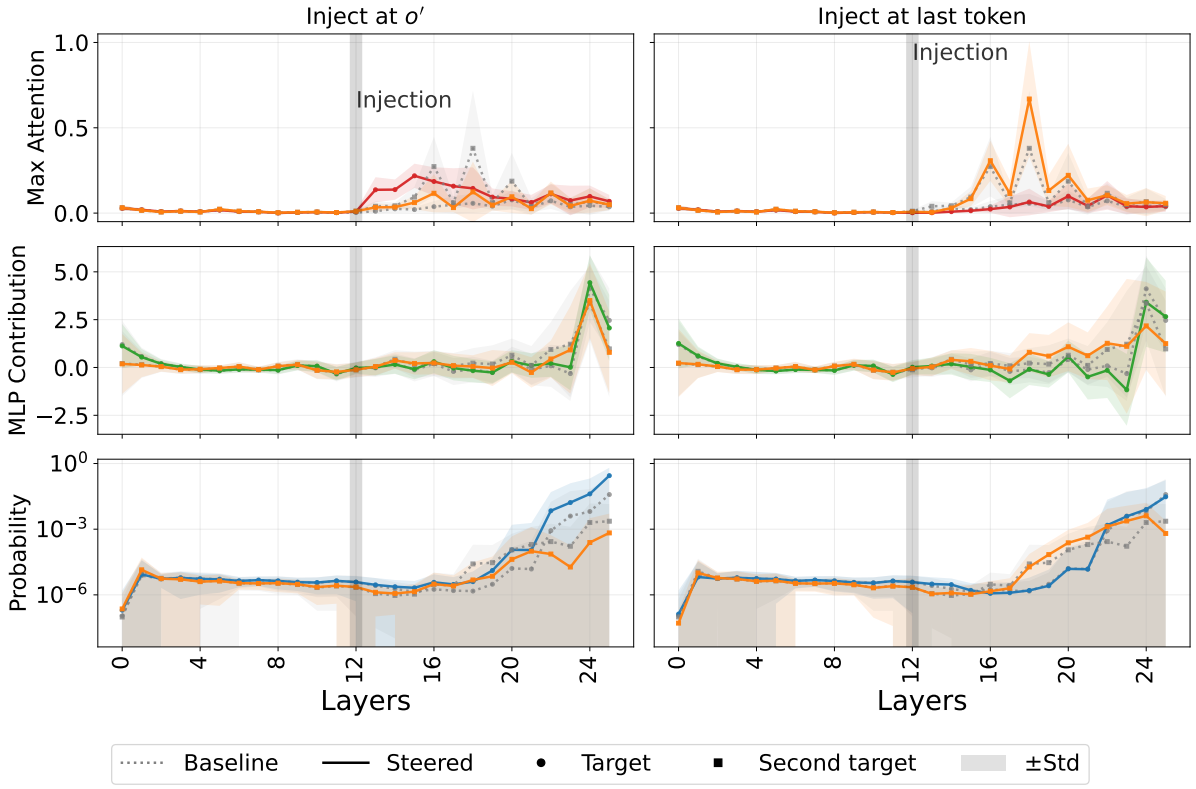


Figure 18: (R→C, Gemma2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 1,032 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

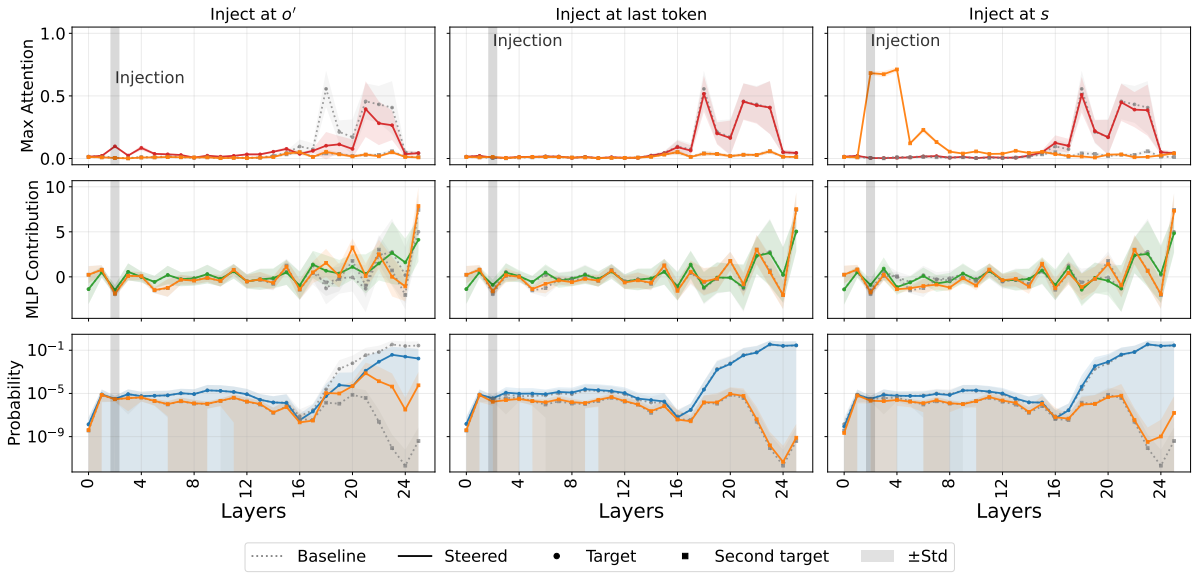


Figure 19: (C→R, T5Gemma - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,221 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

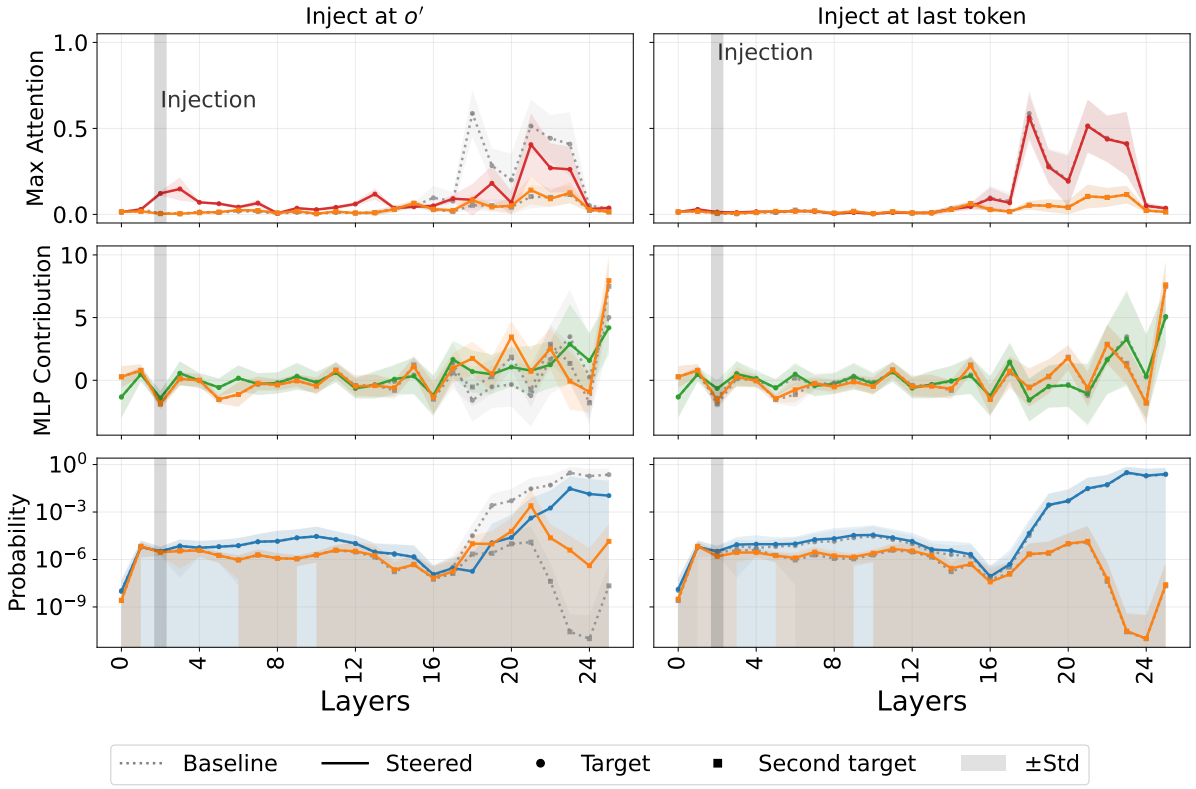


Figure 20: (C→R, T5Gemma - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 1,203 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

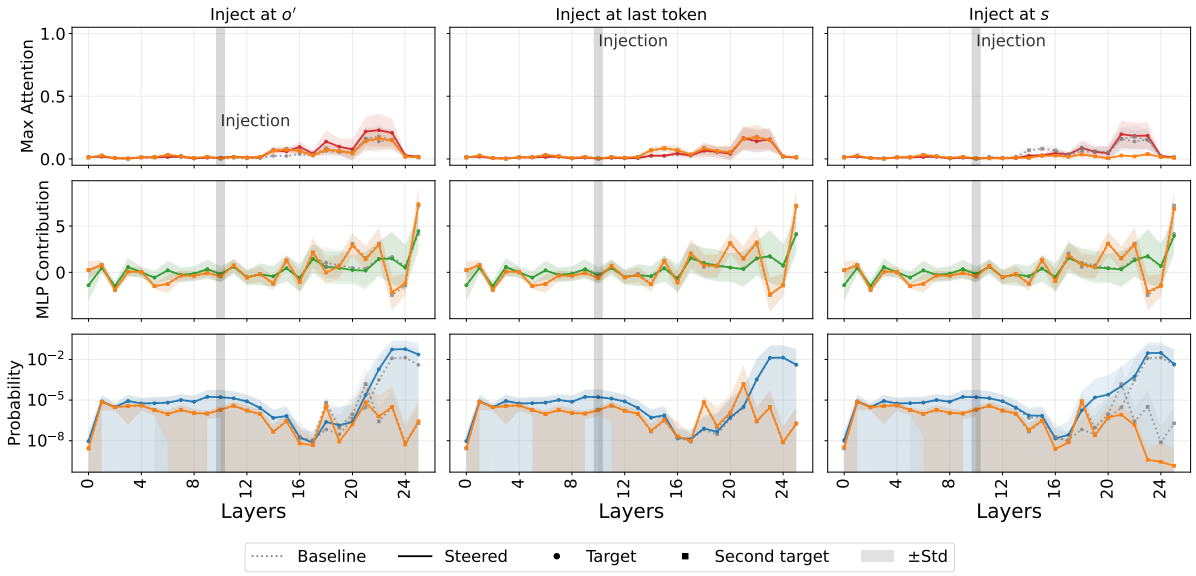


Figure 21: (R→C, T5Gemma - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,221 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.



Figure 22: (R→C, T5Gemma - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 1,203 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

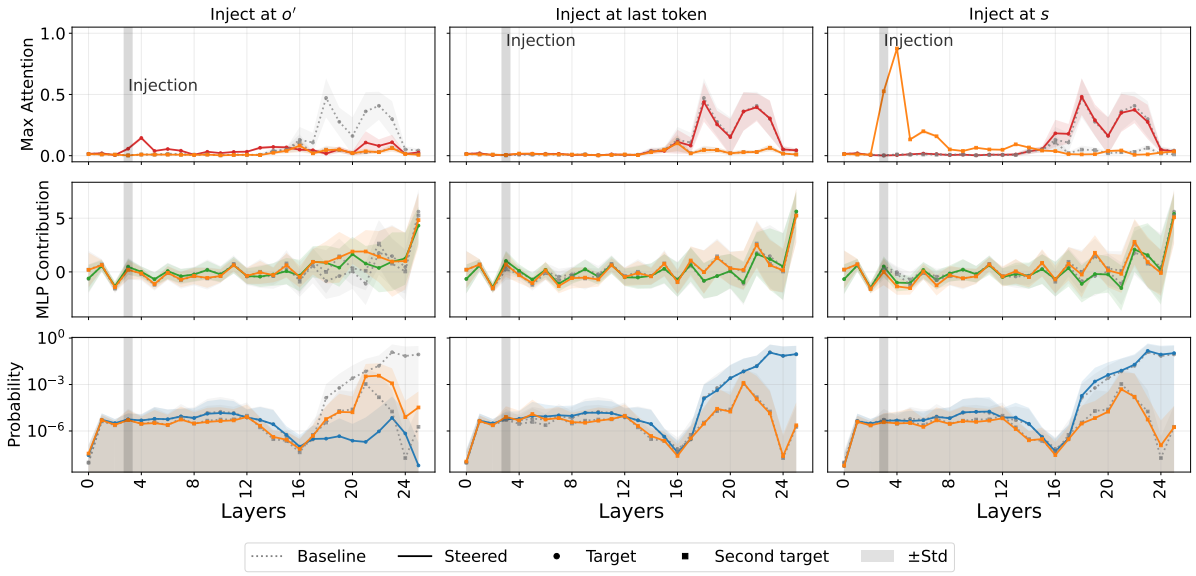


Figure 23: (C→R, T5Gemma - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,776 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.



Figure 24: (C→R, T5Gemma - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 1,125 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

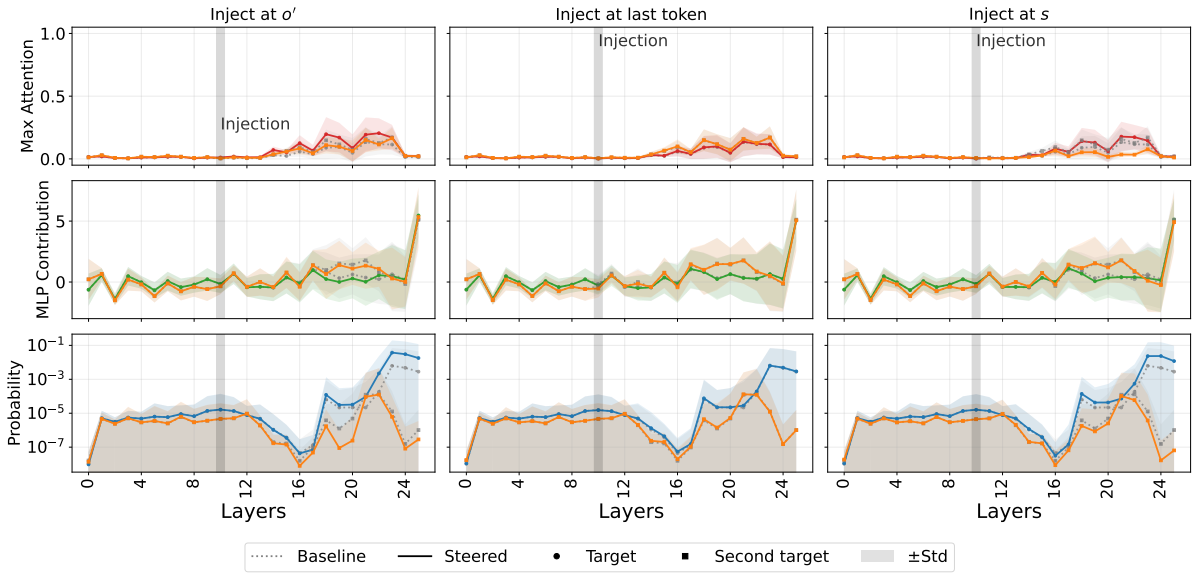


Figure 25: (R→C, T5Gemma - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 1,776 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.



Figure 26: (R→C, T5Gemma - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 1,125 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

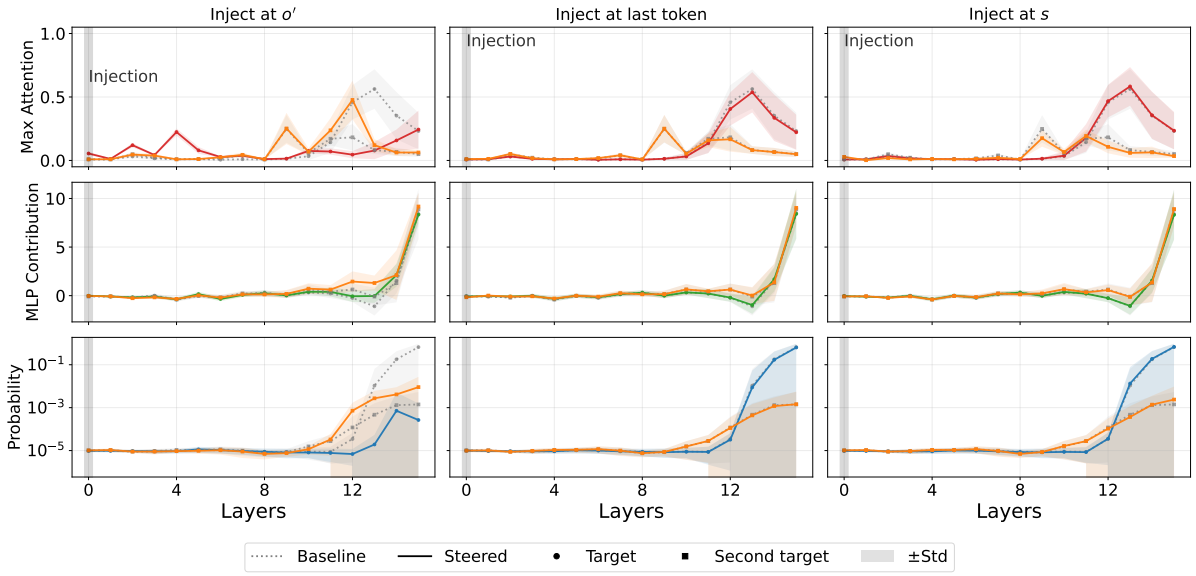


Figure 27: (C→R, OLMo-2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 112 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

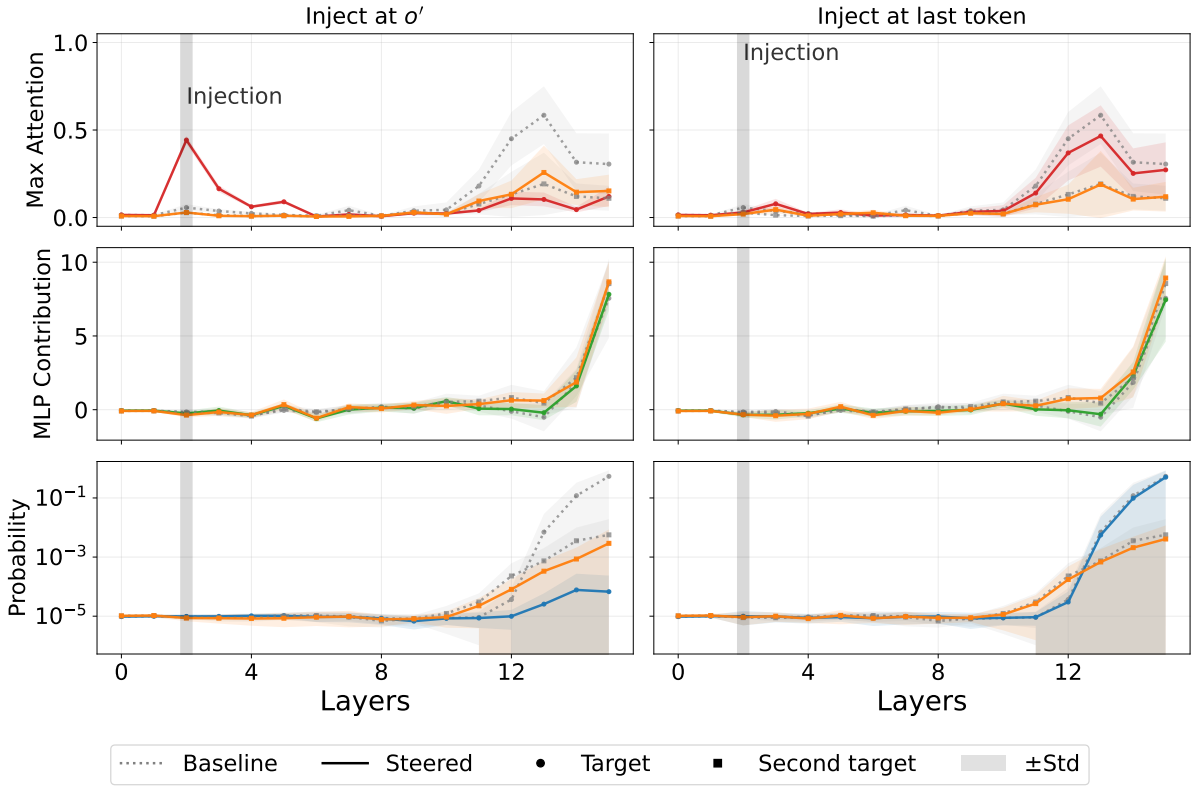


Figure 28: (C→R, OLMo-2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 103 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

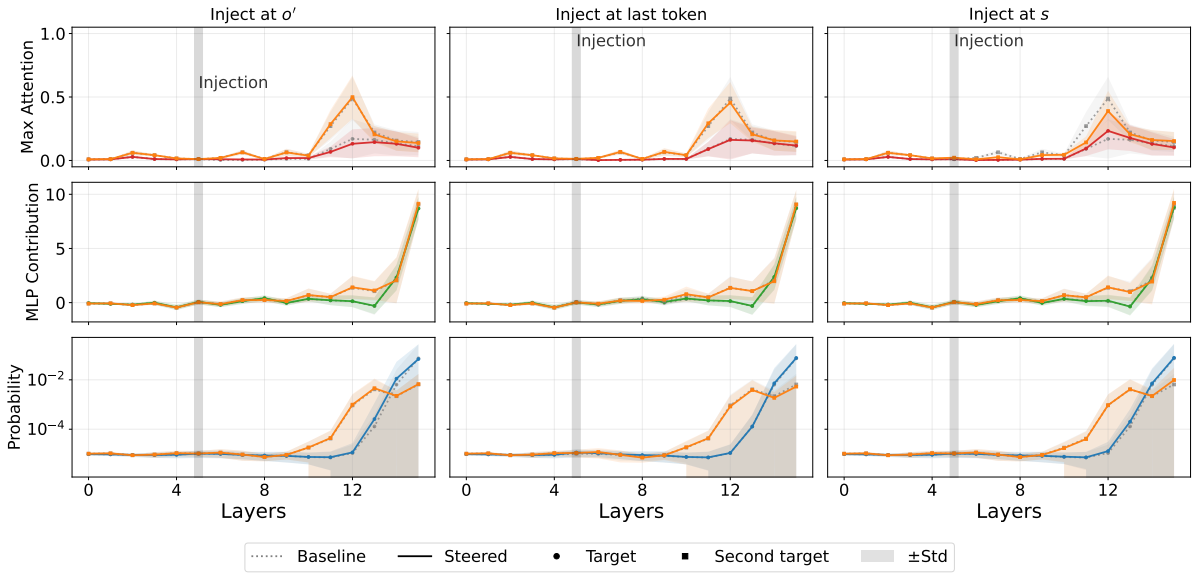


Figure 29: (R→C, OLMo-2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 112 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

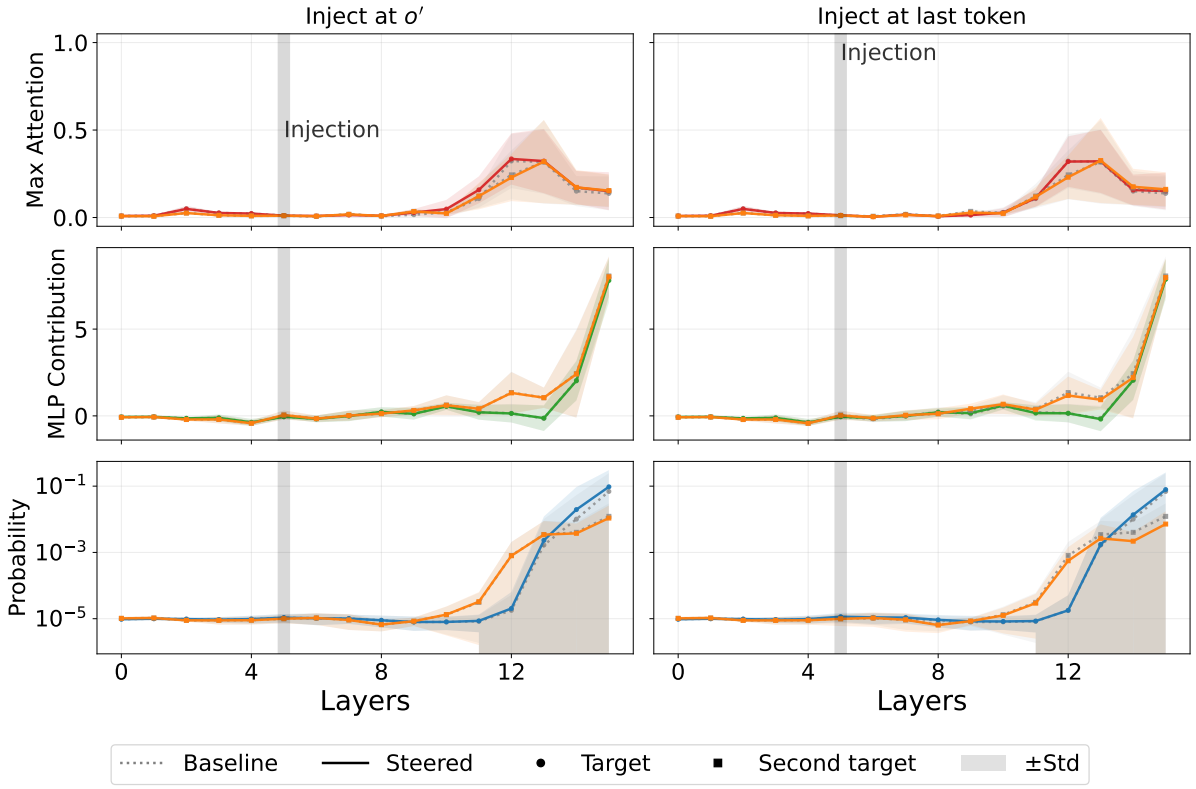


Figure 30: (R→C, OLMo-2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 103 PEQ samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

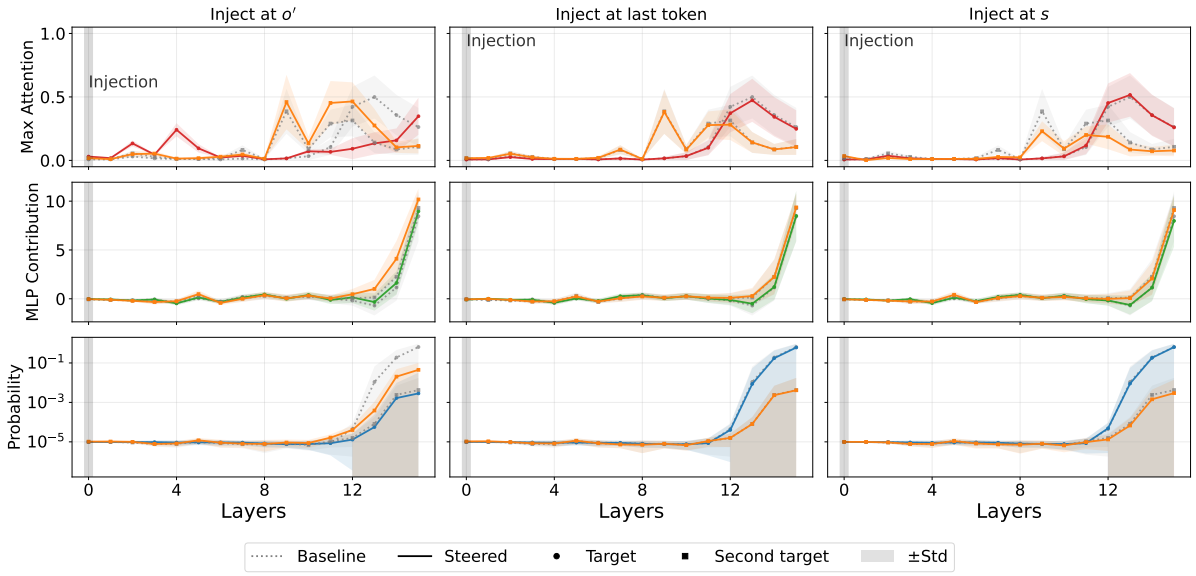


Figure 31: (C→R, OLMo-2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 655 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

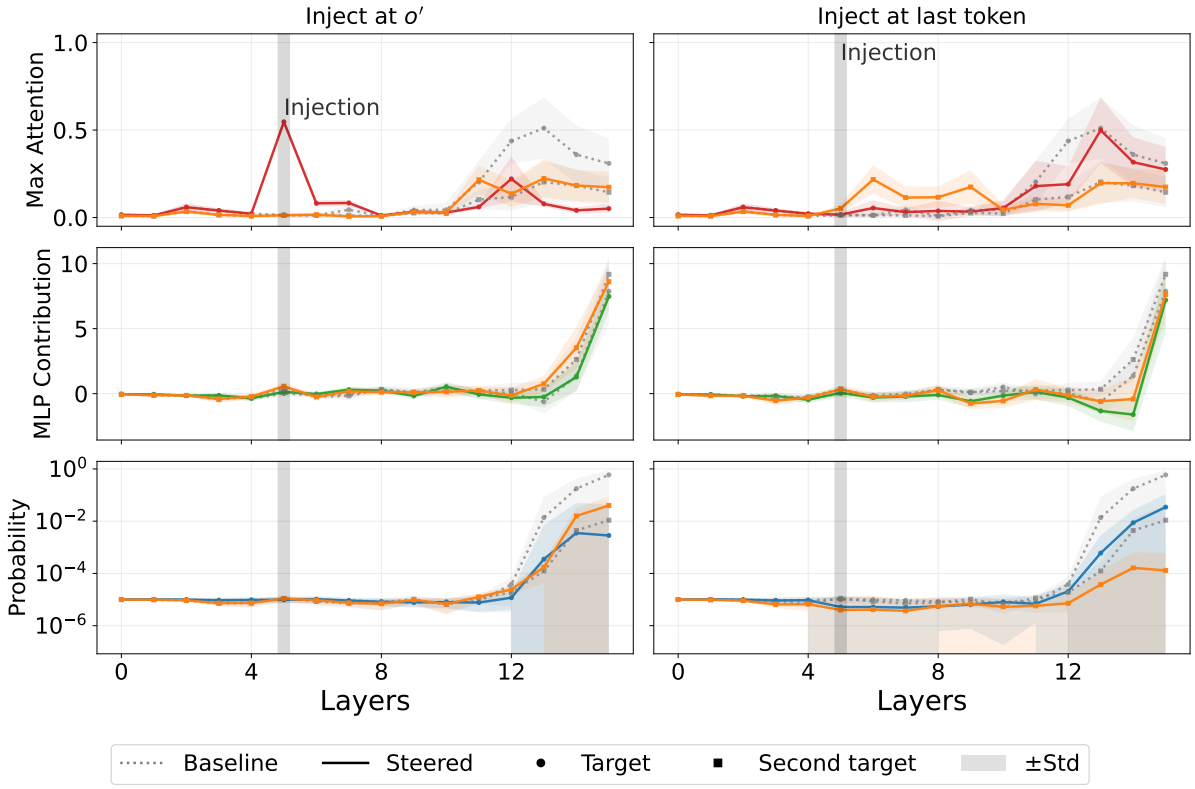


Figure 32: (C→R, OLMo-2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 510 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

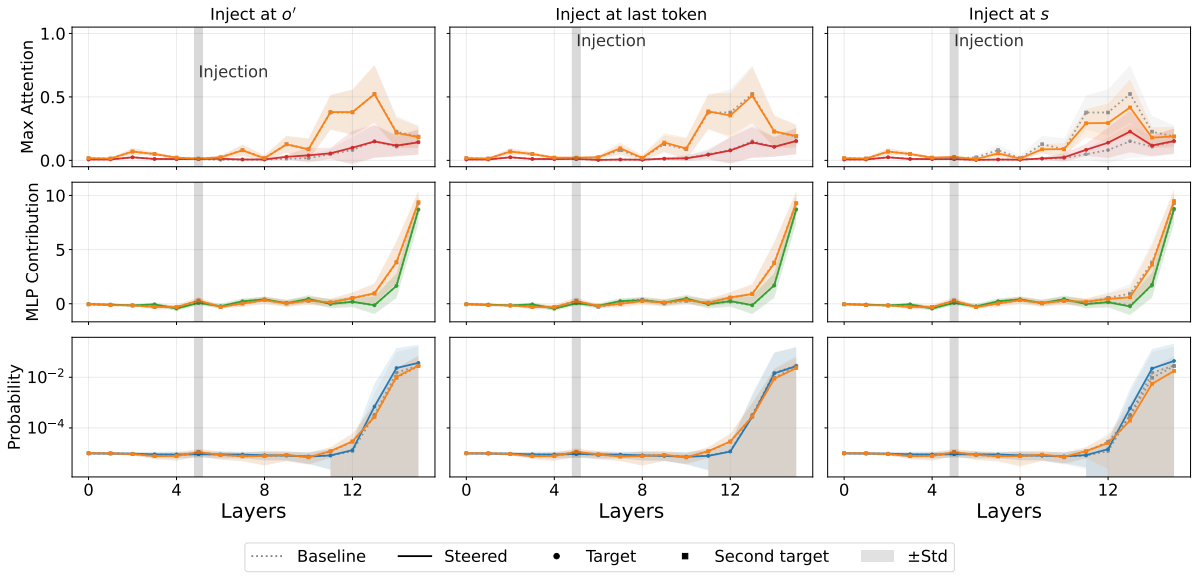


Figure 33: (R→C, OLMo-2 - Context-First: Object vs. last token vs. subject) Showing the mechanistic investigation of parametric steering over 655 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

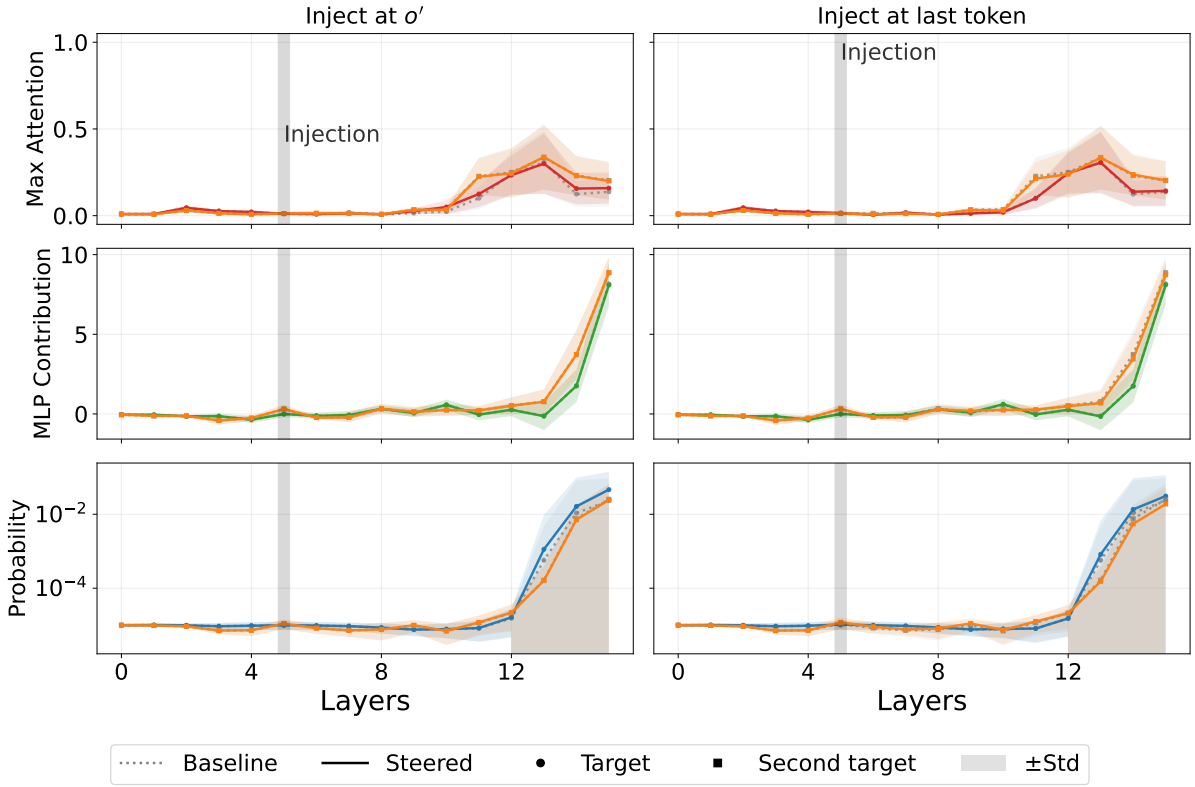


Figure 34: (R→C, OLMo-2 - Query-First: Object vs. last token only) Showing the mechanistic investigation of parametric steering over 510 PopQA samples. Top: Attention (Subject=Orange, Counterfactual=Red). Middle: MLP(Truth=Orange, Counterfactual=Green). Bottom: Probability(Truth=Orange, Counterfactual=Blue). The intervention acts as an amplifier and successfully shifts the behavior from all injection positions.

E Behavioral Metrics (Performance vs. Fluency)

E.1 PopQA: Object Token (Query-First)

E.1.1 Parametric Steering (Copy→Recall) - Gemma2-2B

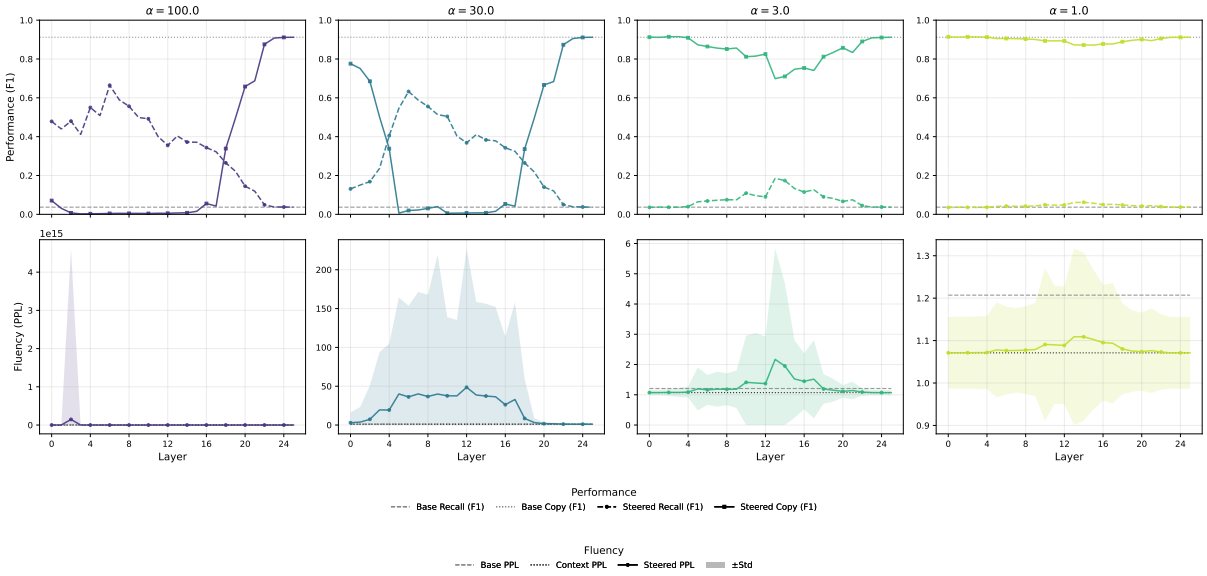


Figure 35: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.1.2 Contextual Steering (Recall→Copy) - Gemma2-2B

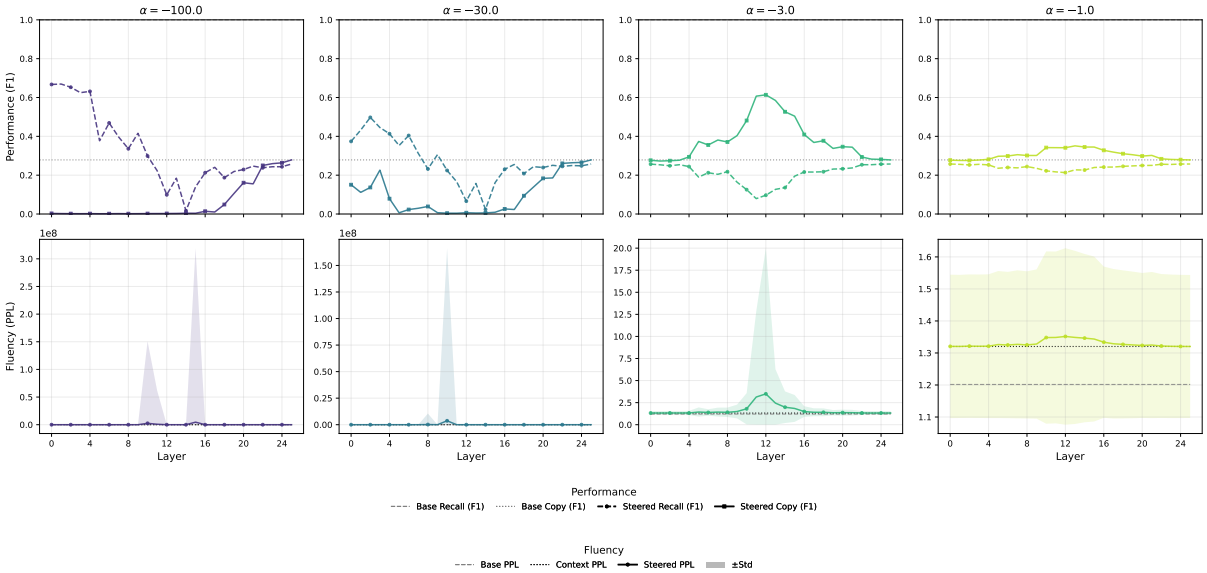


Figure 36: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.1.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

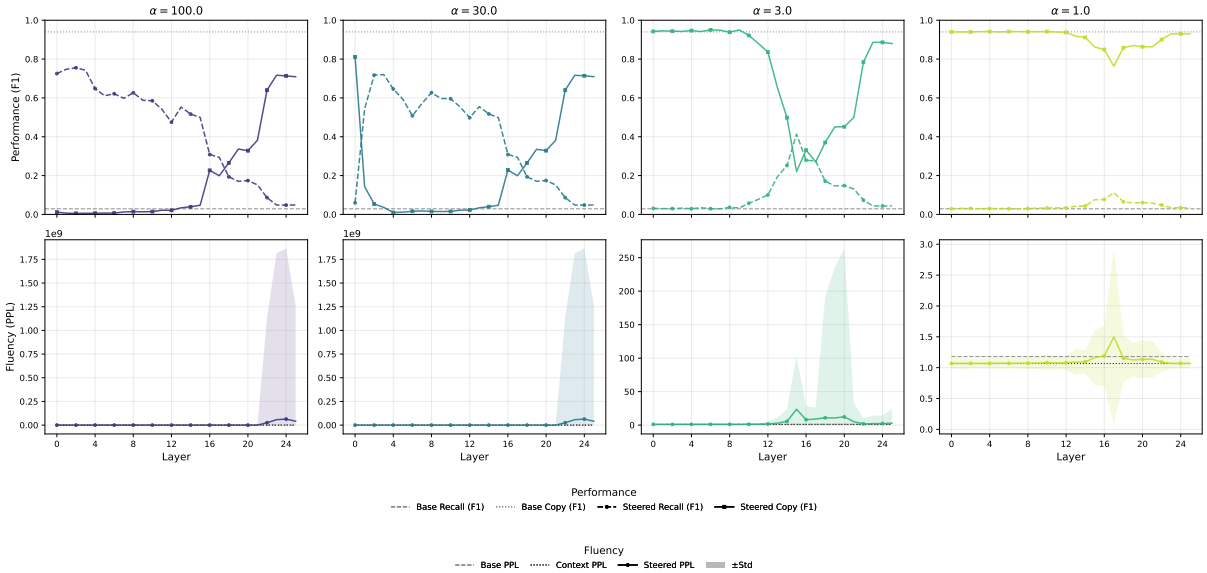


Figure 37: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.1.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

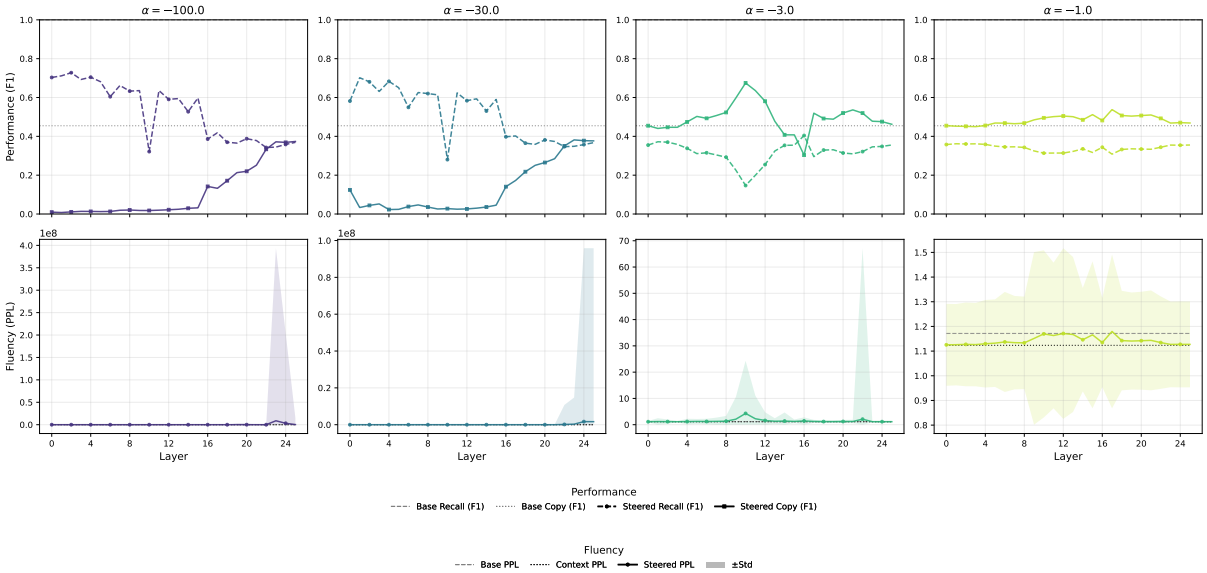


Figure 38: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.1.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

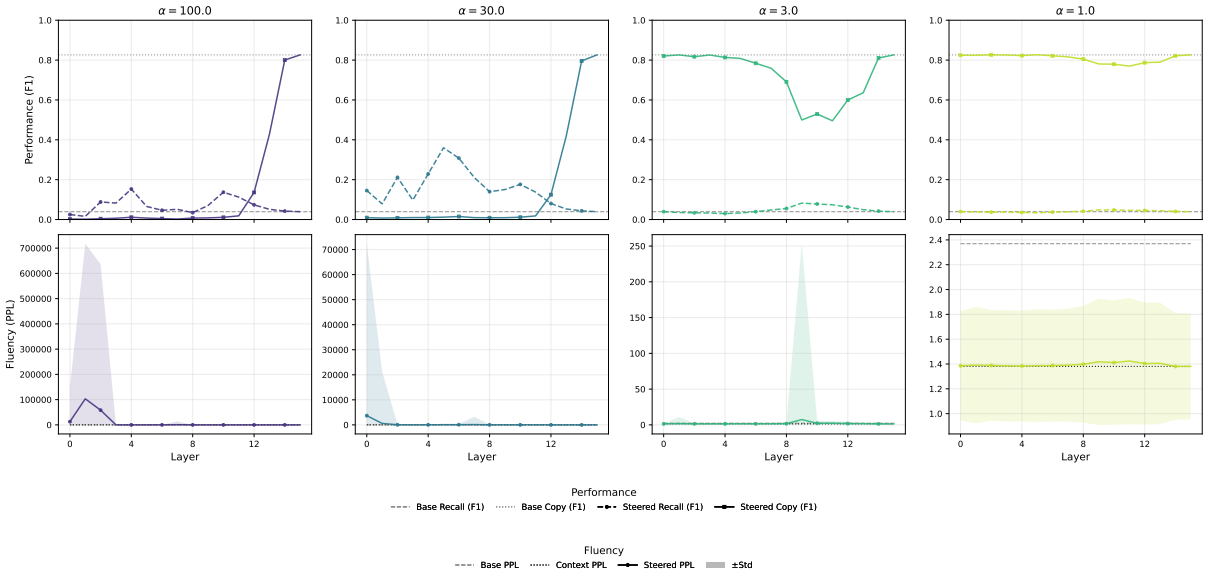


Figure 39: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.1.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

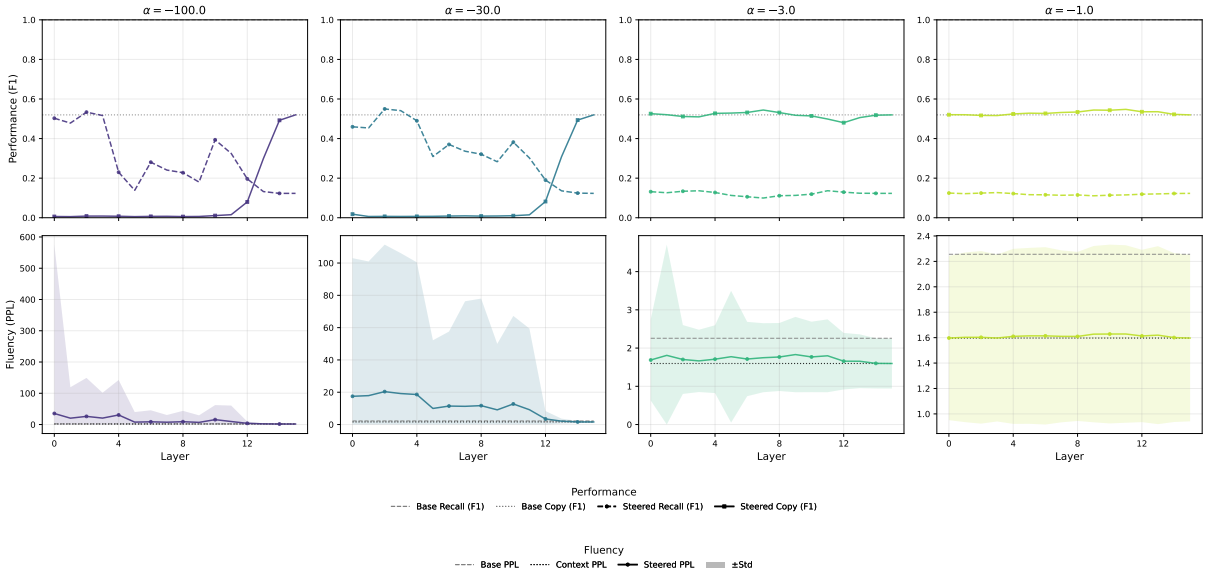


Figure 40: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.2 PopQA: Object Token (Context-First)

E.2.1 Parametric Steering (Copy→Recall) - Gemma2-2B

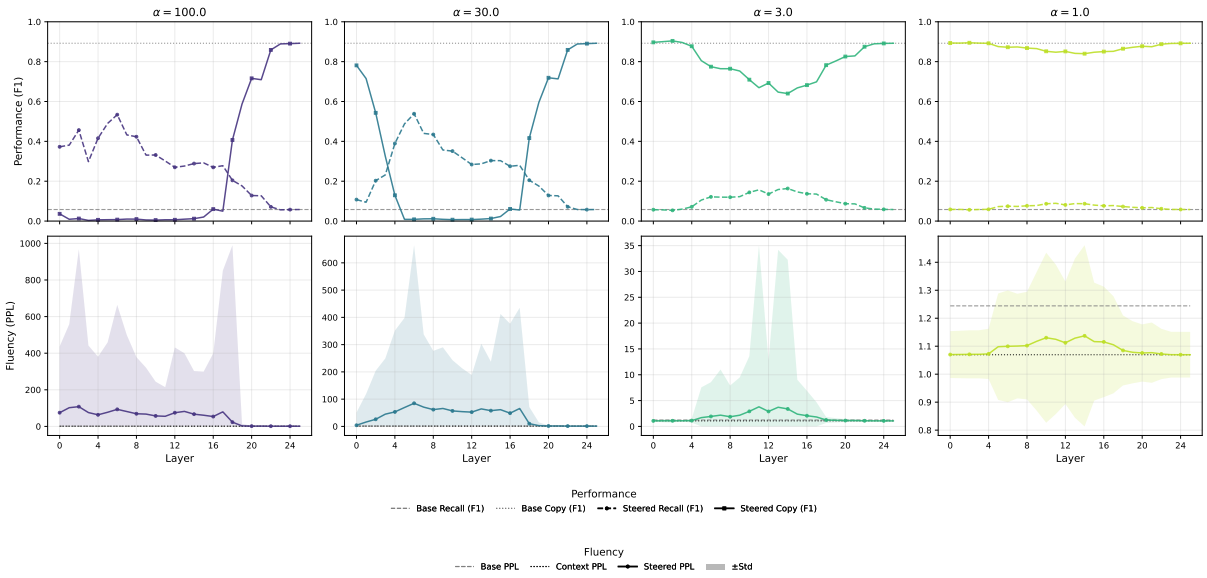


Figure 41: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.2.2 Contextual Steering (Recall→Copy) - Gemma2-2B

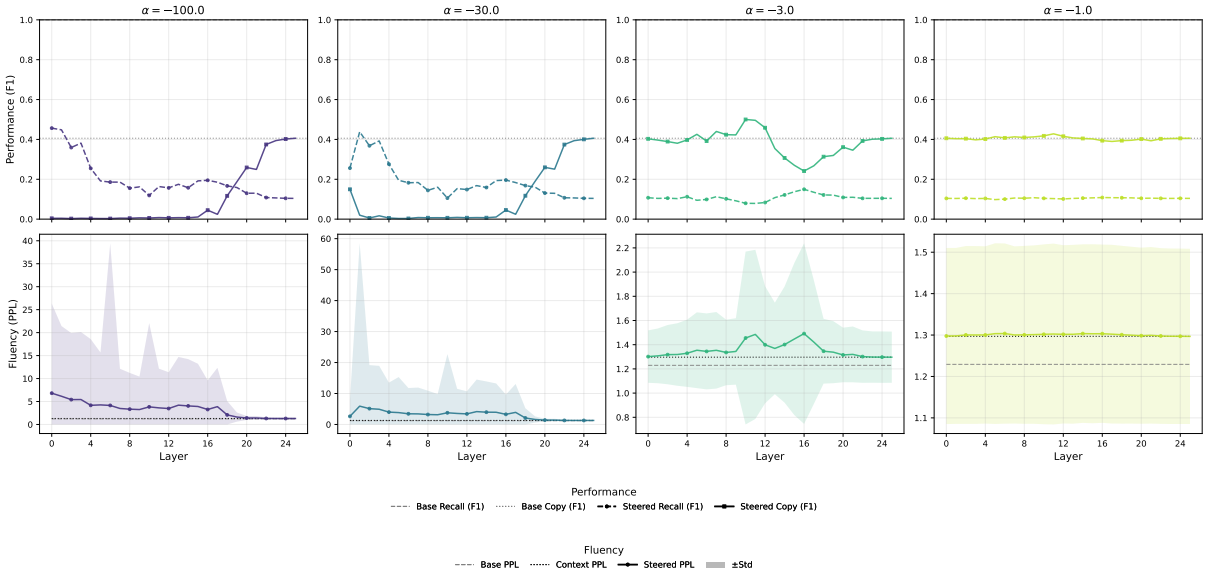


Figure 42: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.2.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

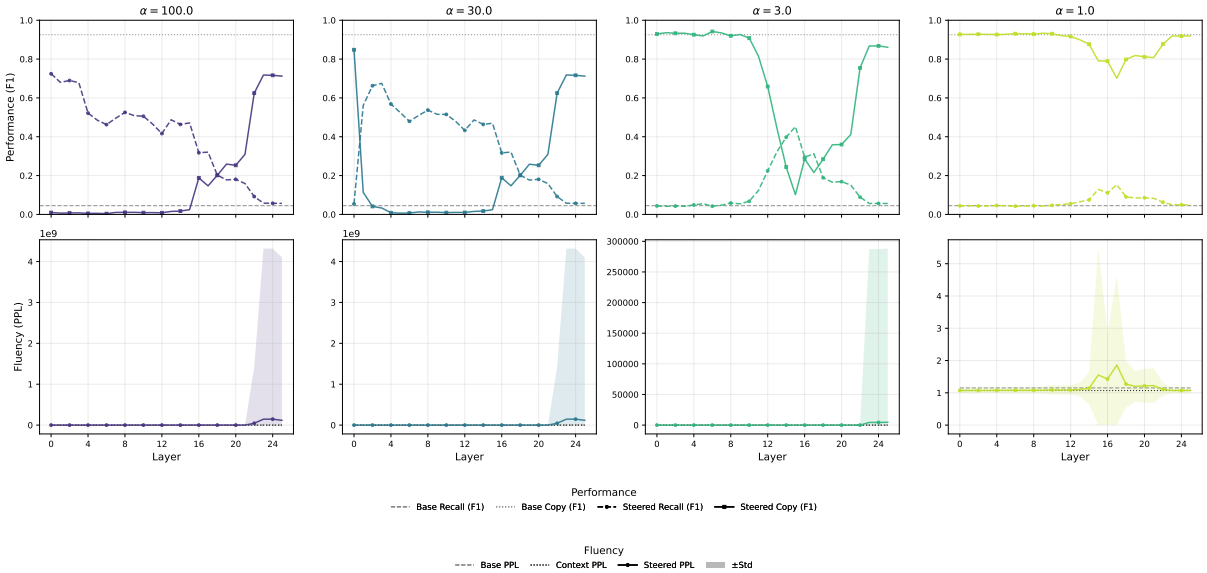


Figure 43: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.2.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

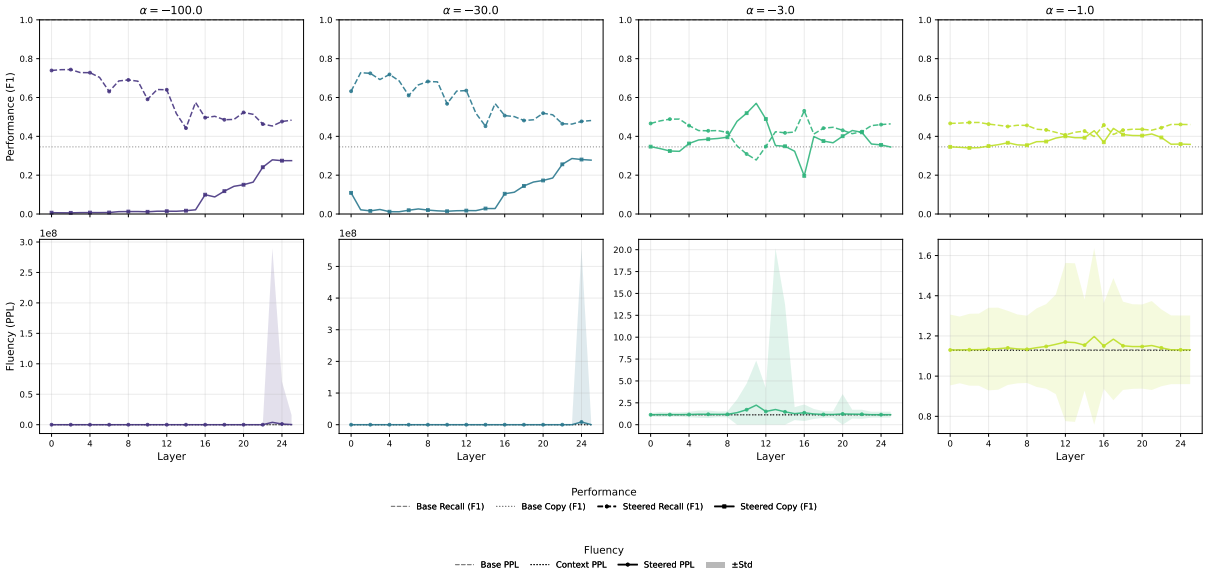


Figure 44: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.2.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

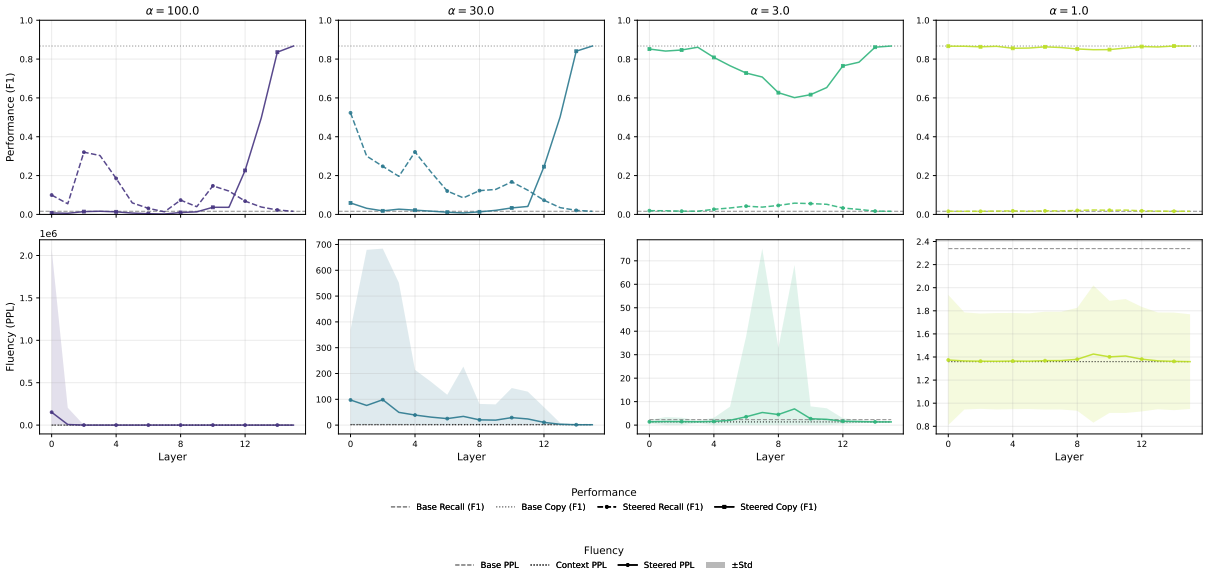


Figure 45: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.2.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

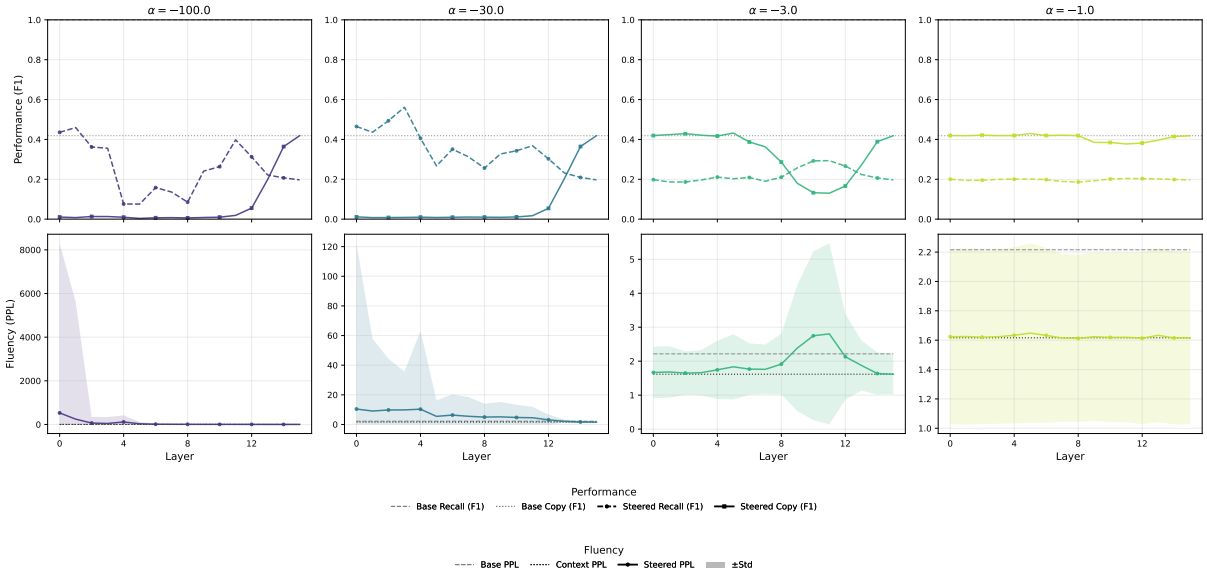


Figure 46: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.3 PopQA: Subject Token (Context-First)

E.3.1 Parametric Steering (Copy→Recall) - Gemma2-2B

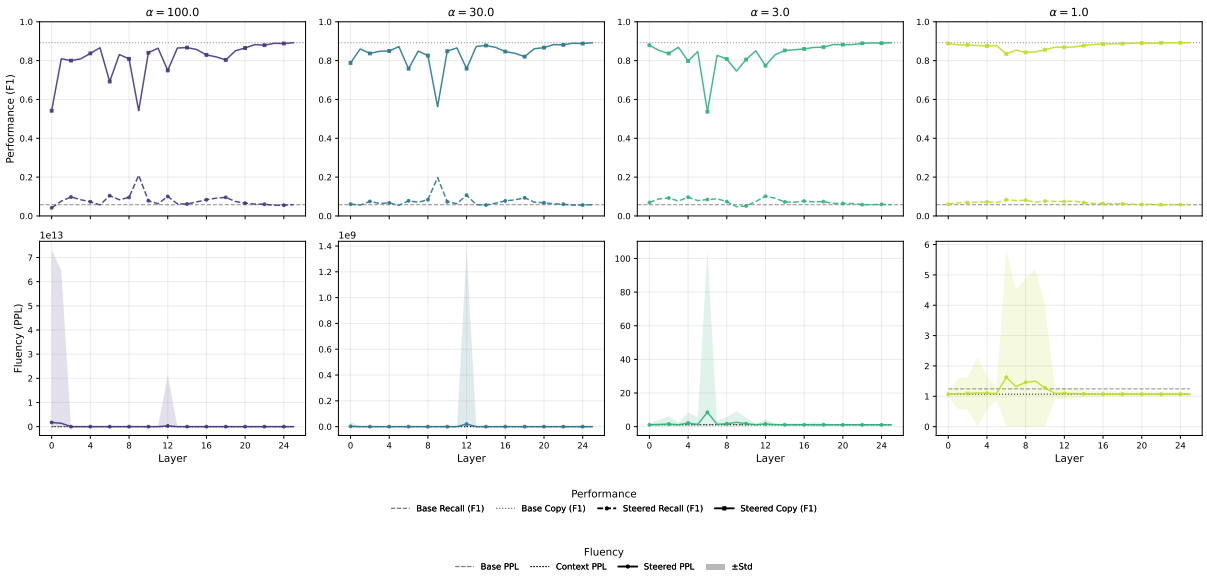


Figure 47: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.3.2 Contextual Steering (Recall→Copy) - Gemma2-2B

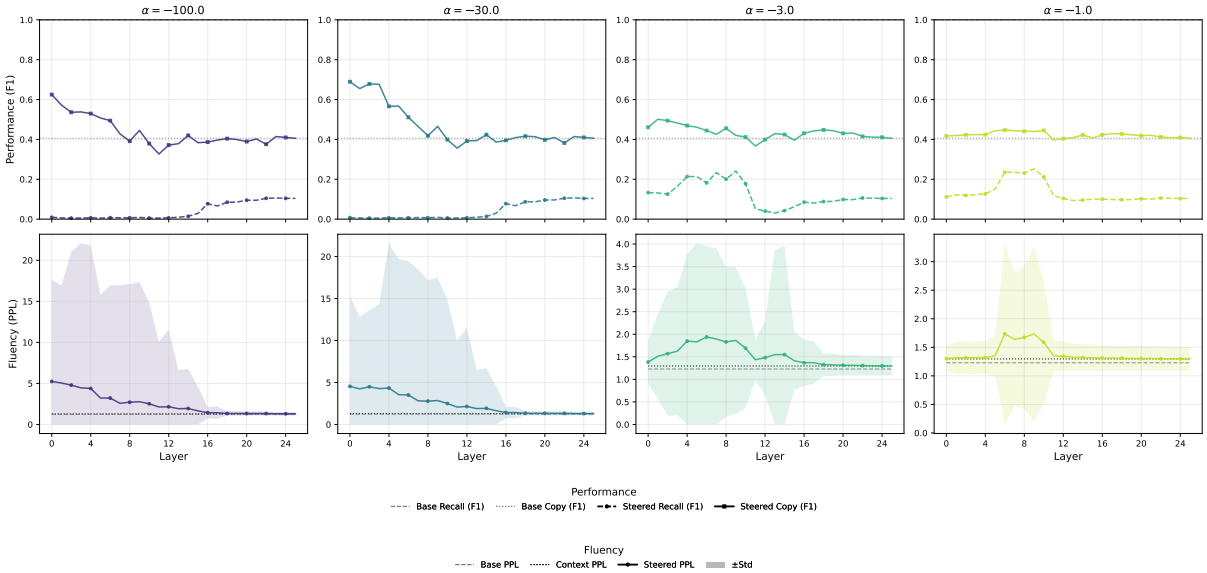


Figure 48: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.3.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

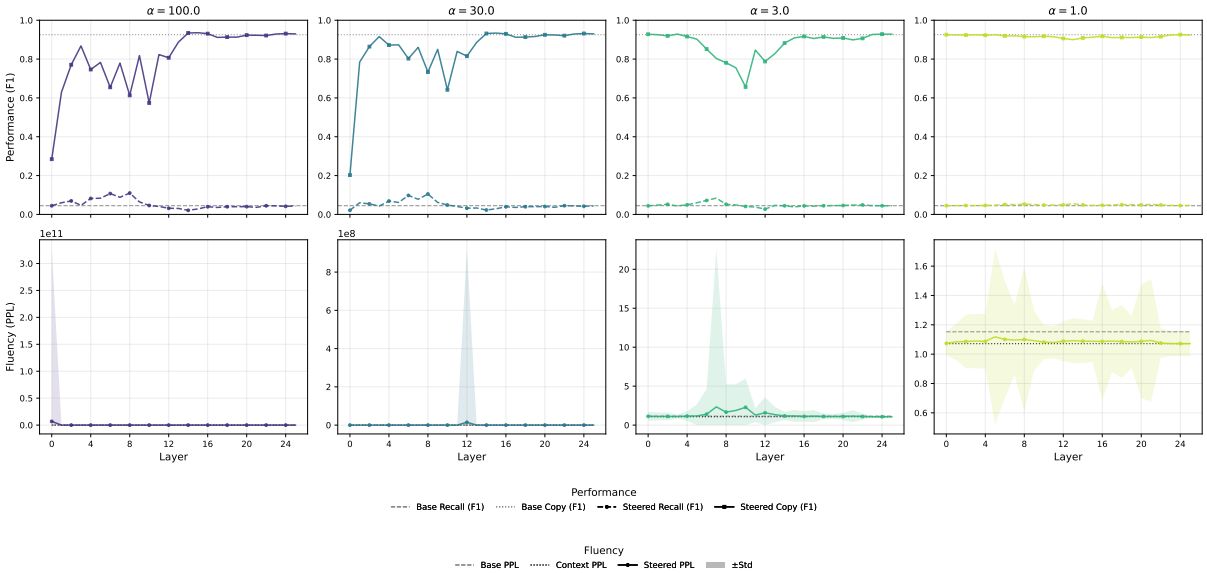


Figure 49: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.3.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

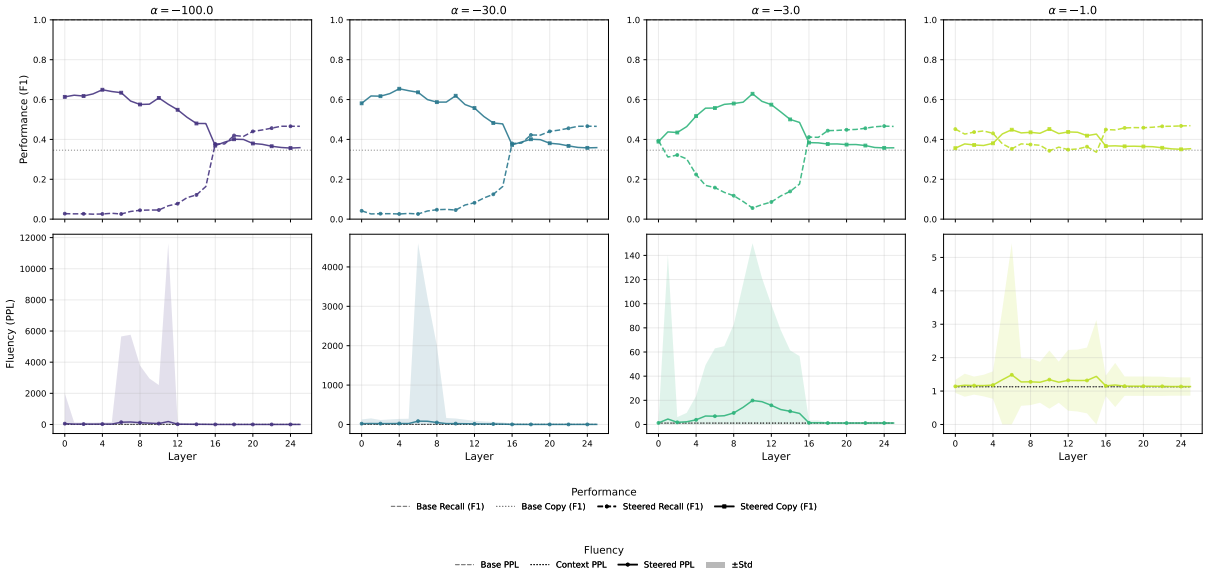


Figure 50: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.3.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

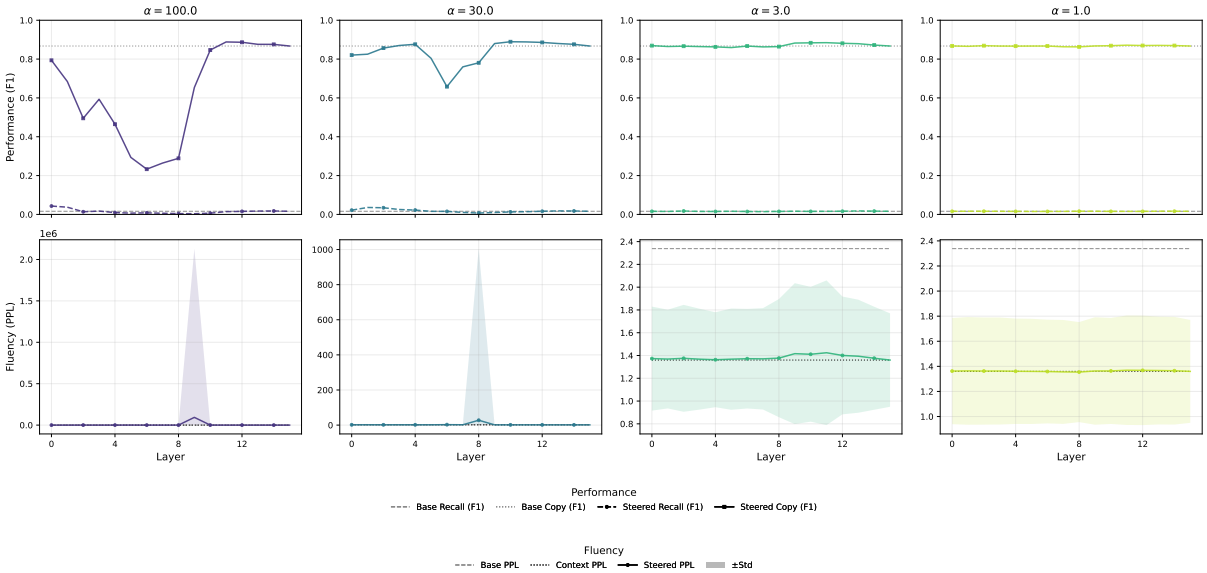


Figure 51: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.3.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

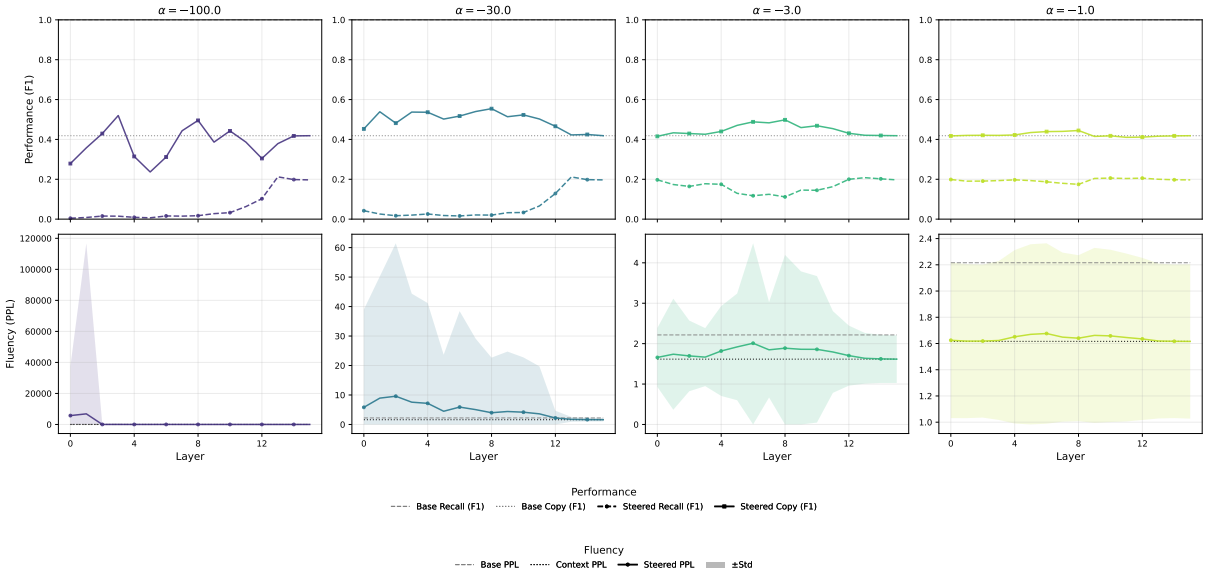


Figure 52: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.4 PopQA: Last Token (Query-First)

E.4.1 Parametric Steering (Copy→Recall) - Gemma2-2B

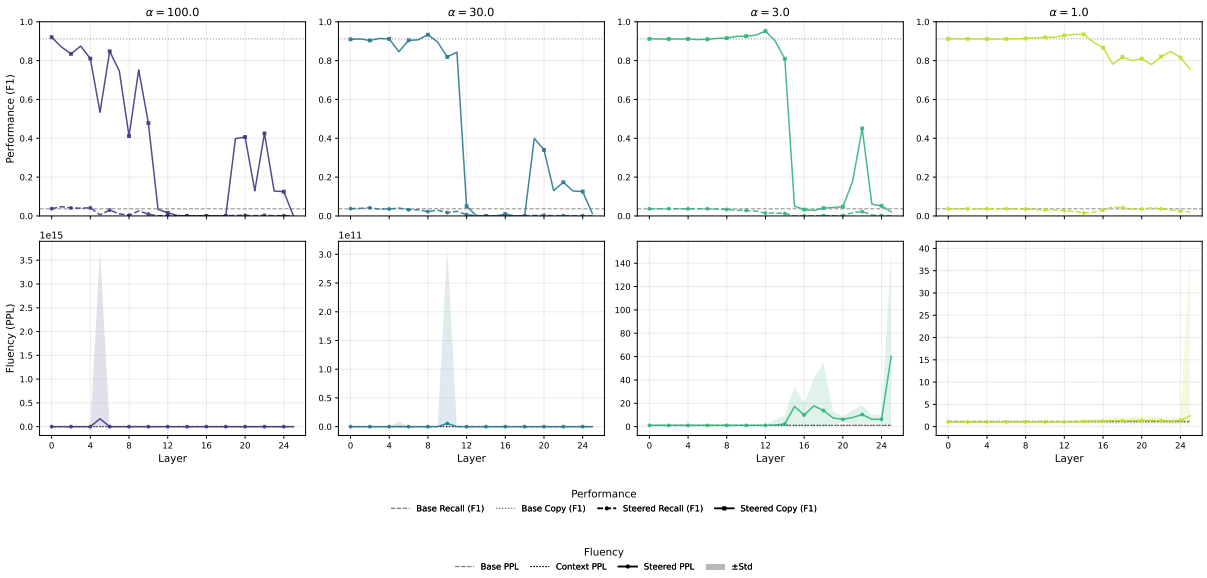


Figure 53: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.4.2 Contextual Steering (Recall→Copy) - Gemma2-2B

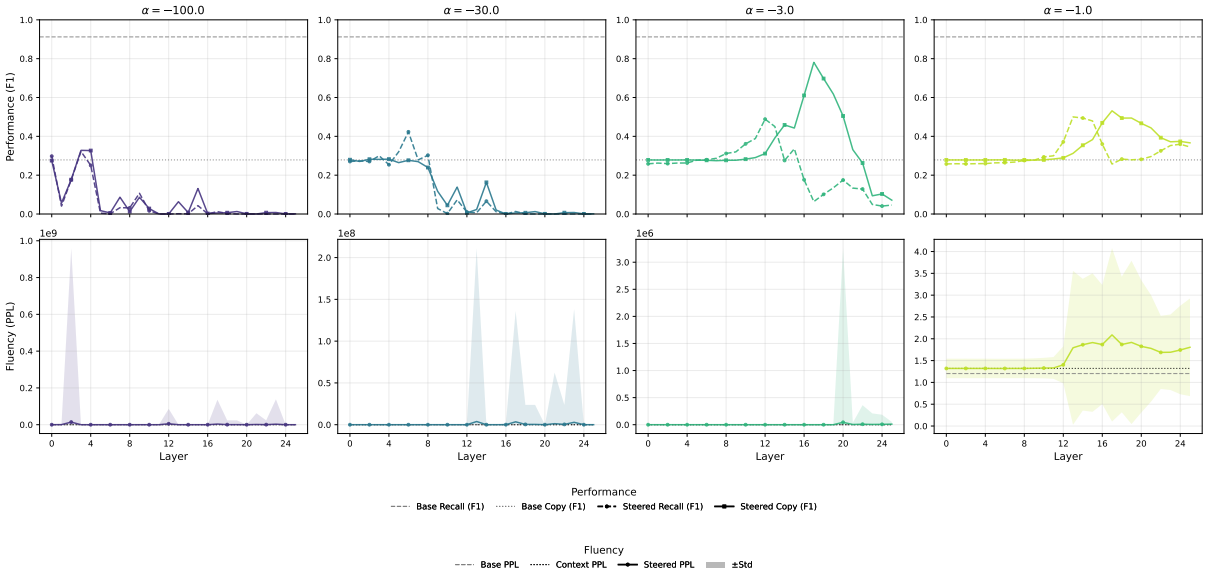


Figure 54: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.4.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

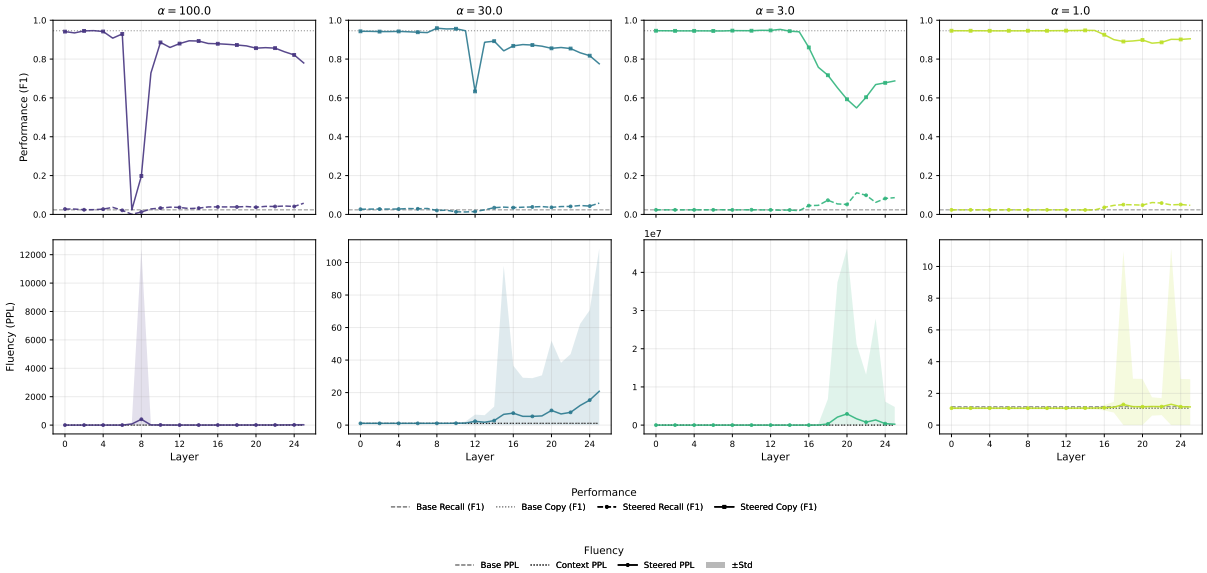


Figure 55: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.4.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

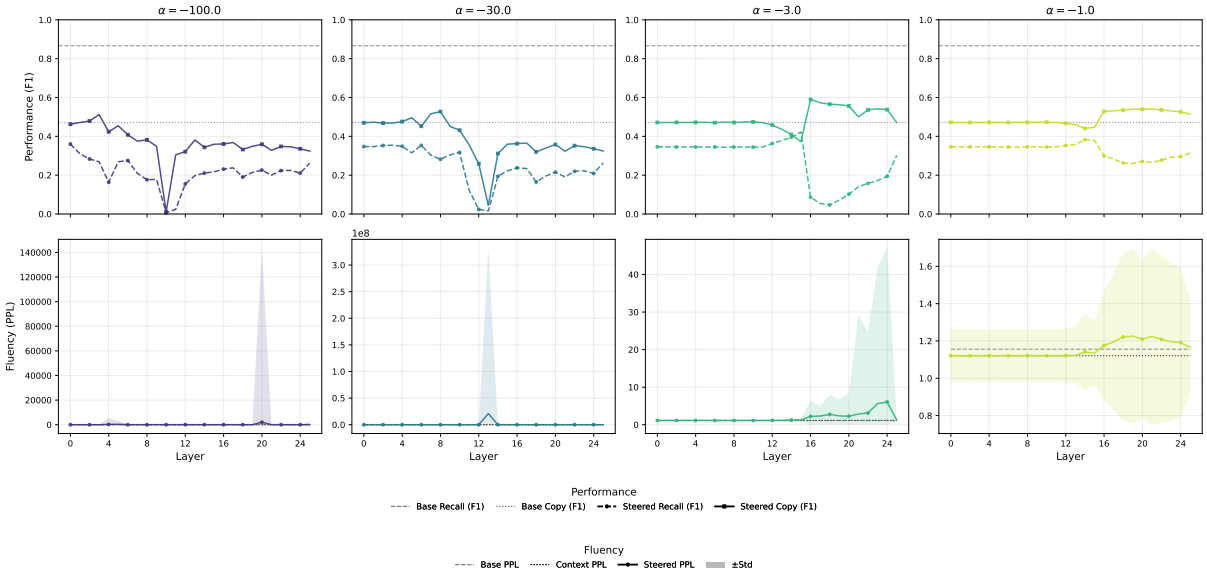


Figure 56: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.4.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

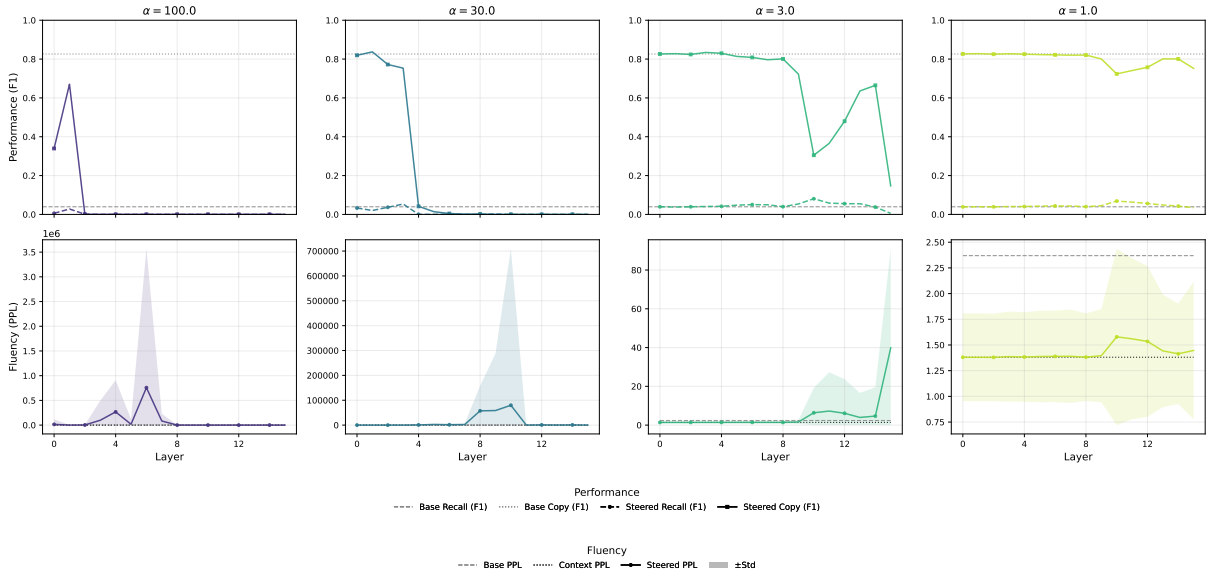


Figure 57: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.4.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

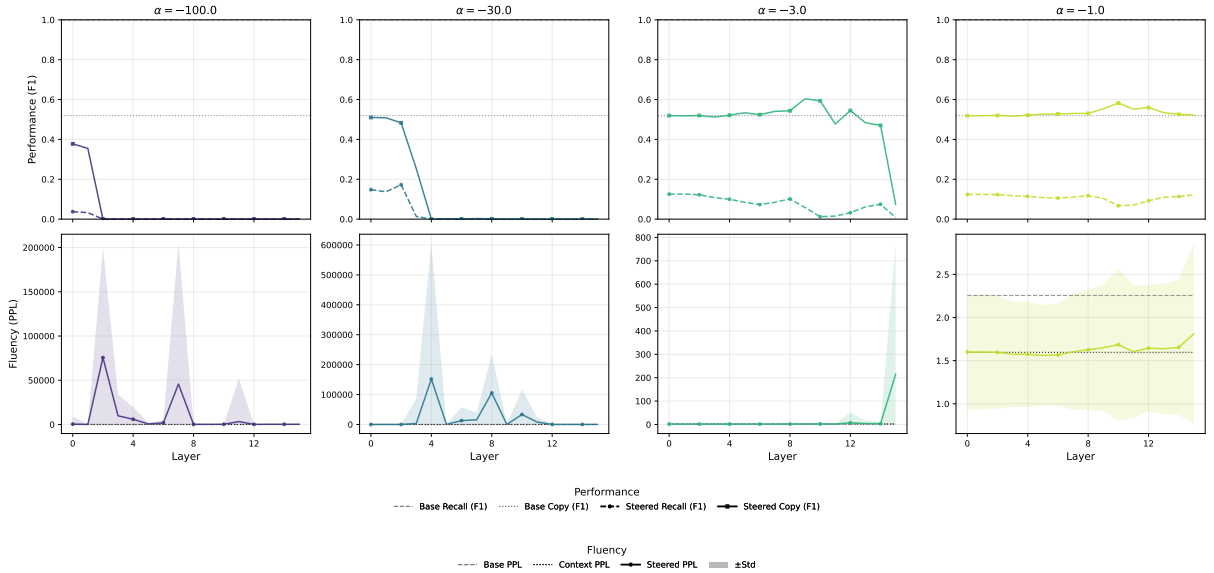


Figure 58: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.5 PopQA: Last Token (Context-First)

E.5.1 Parametric Steering (Copy→Recall) - Gemma2-2B

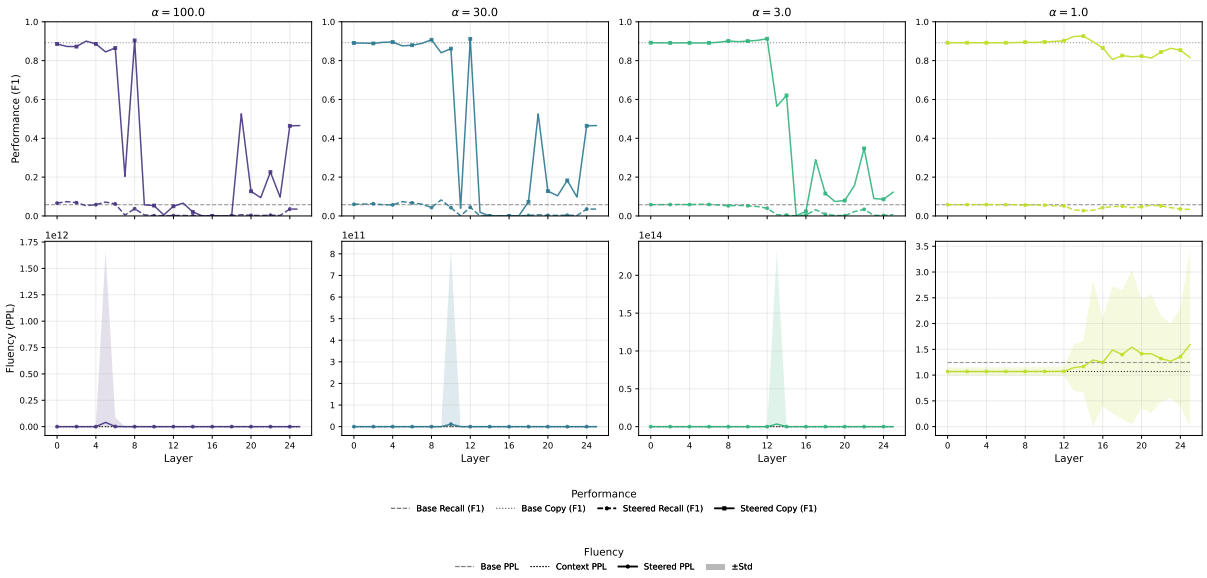


Figure 59: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.5.2 Contextual Steering (Recall→Copy) - Gemma2-2B

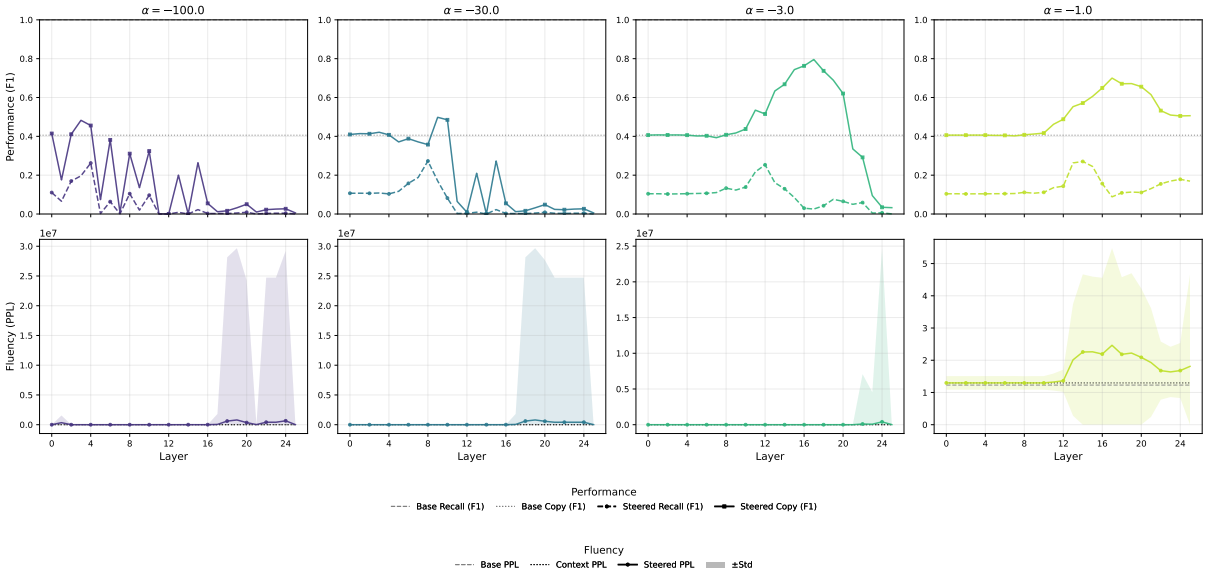


Figure 60: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.5.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

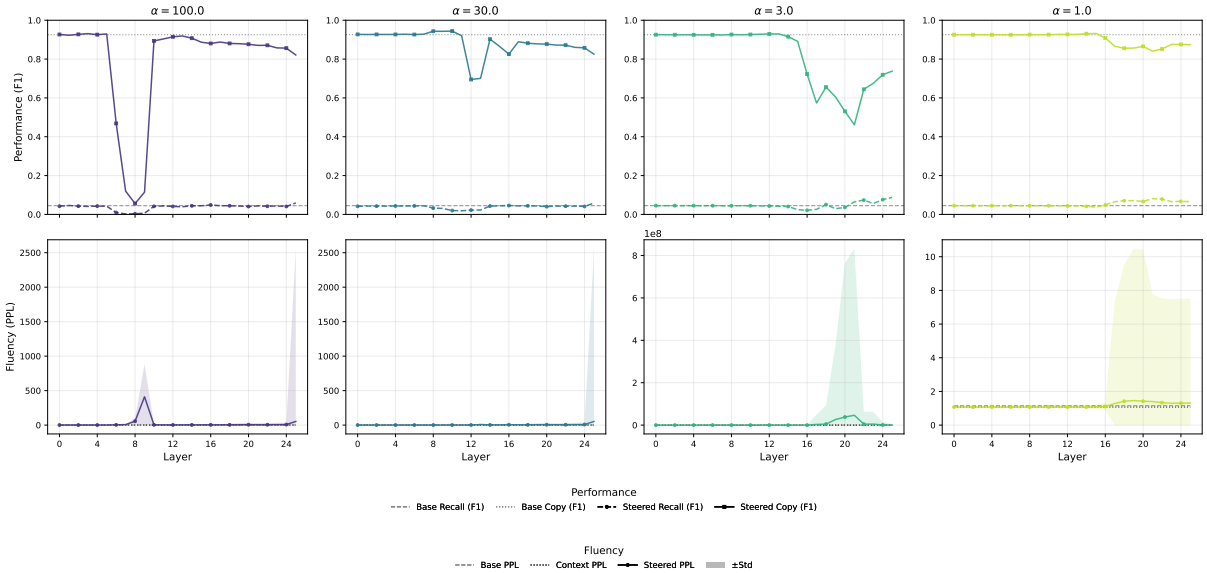


Figure 61: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.5.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

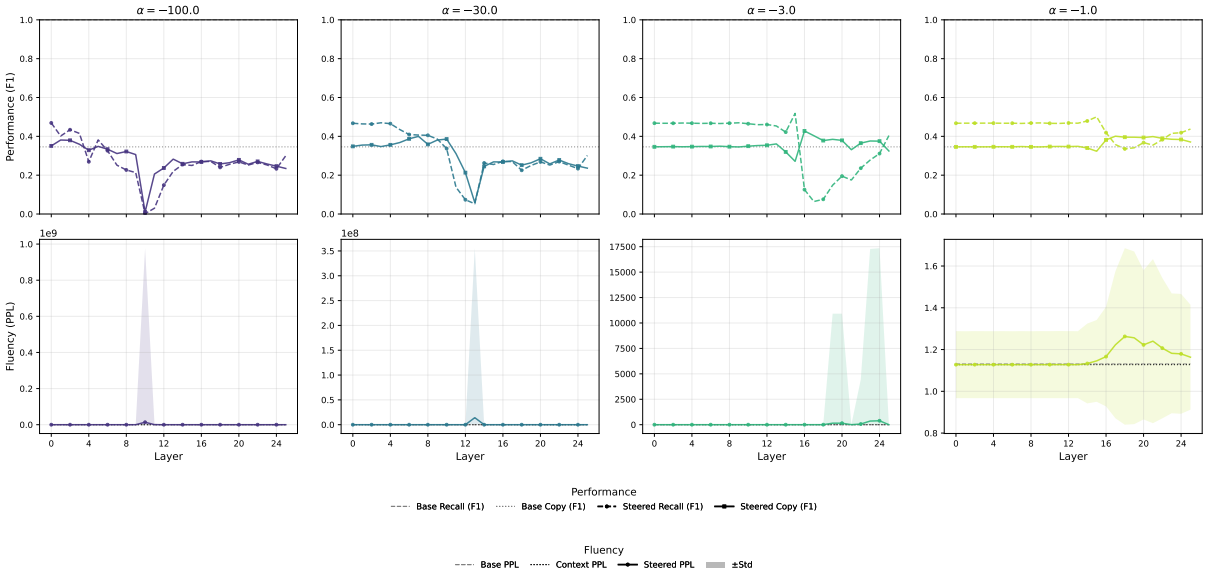


Figure 62: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.5.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

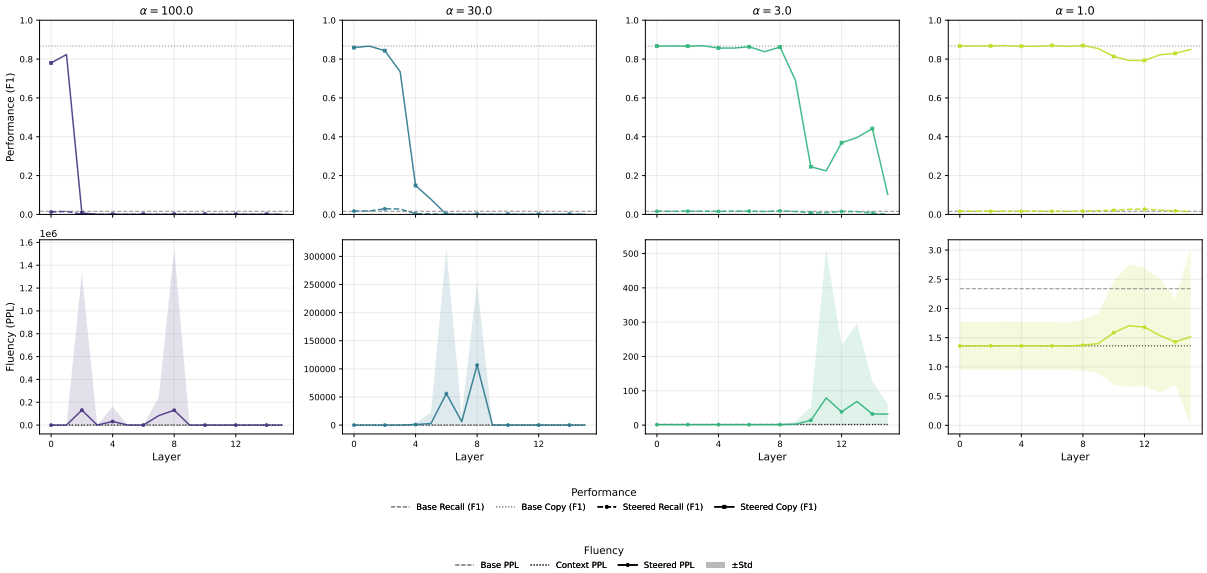


Figure 63: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.5.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

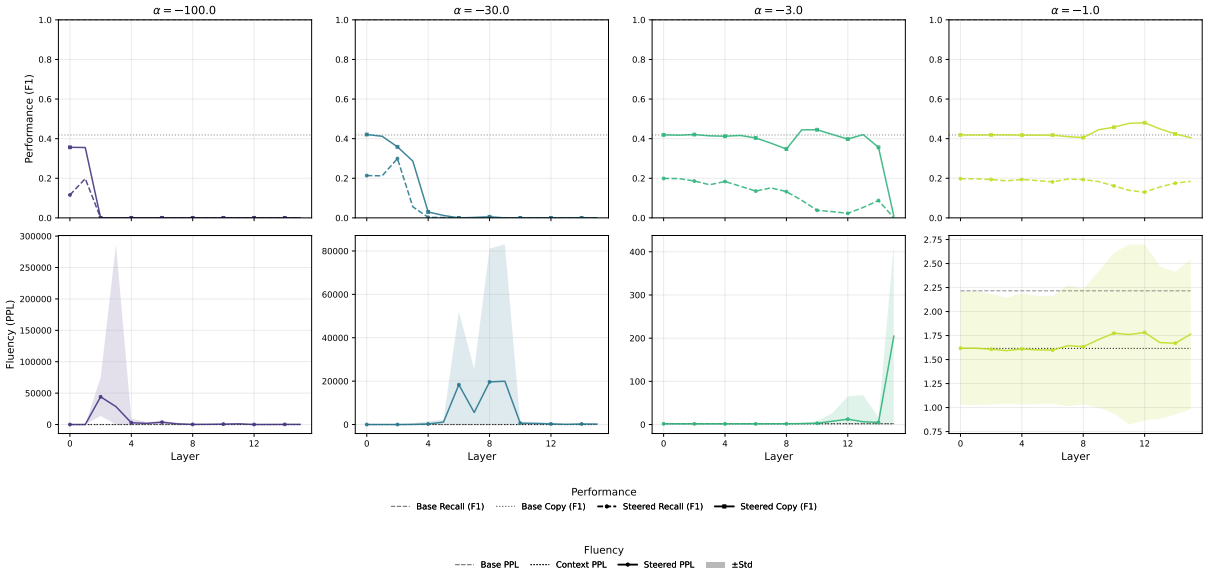


Figure 64: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.6 PEQ: Object Token (Query-First)

E.6.1 Parametric Steering (Copy→Recall) - Gemma2-2B

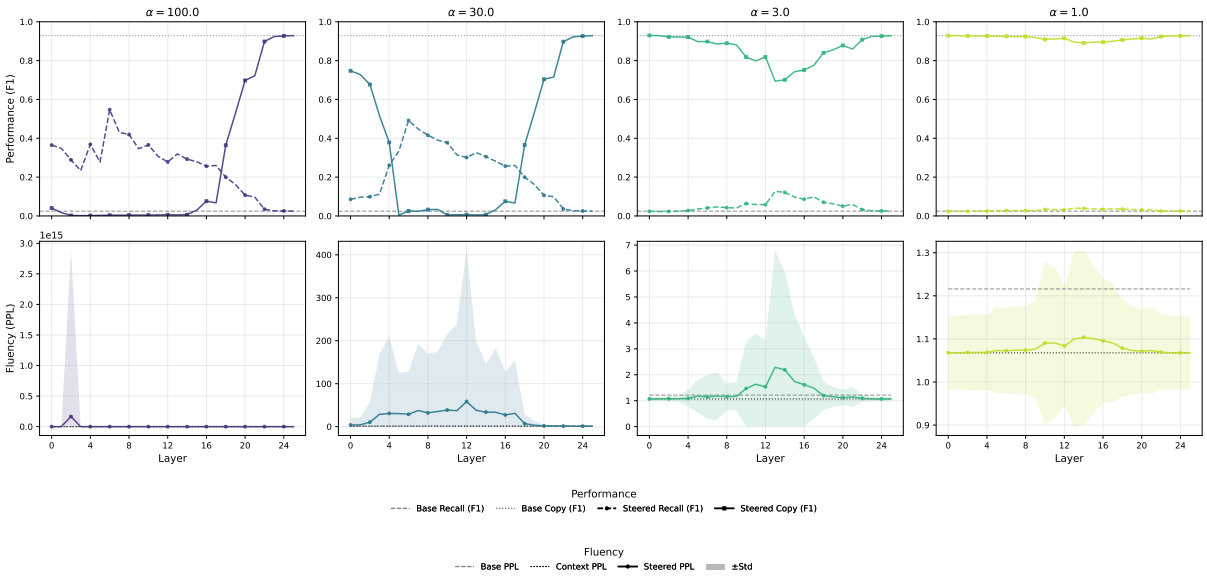


Figure 65: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.6.2 Contextual Steering (Recall→Copy) - Gemma2-2B

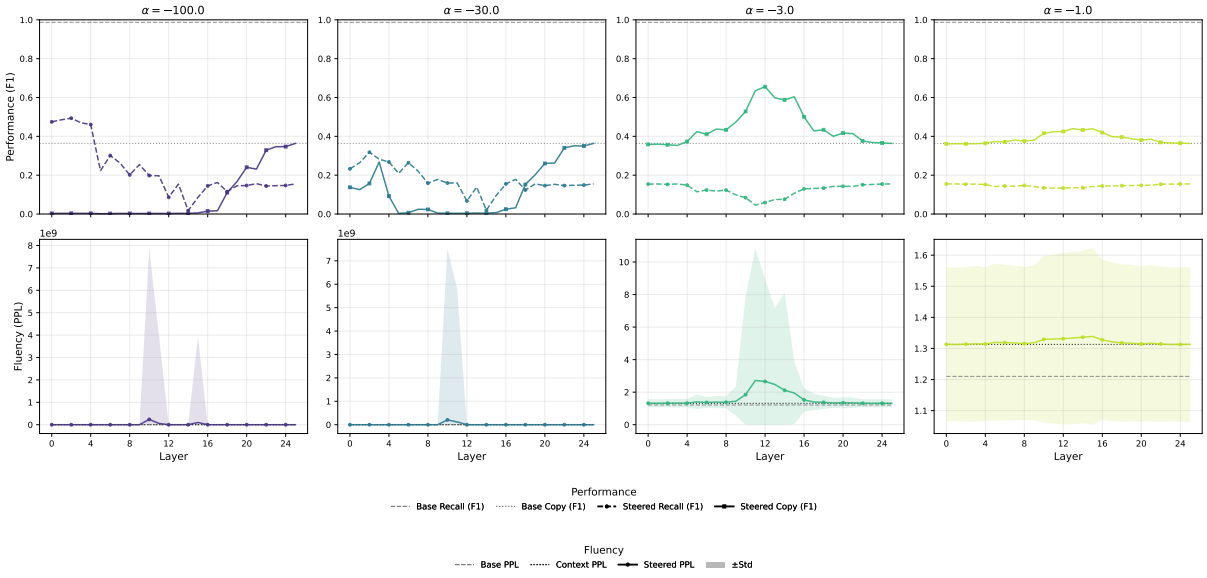


Figure 66: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.6.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

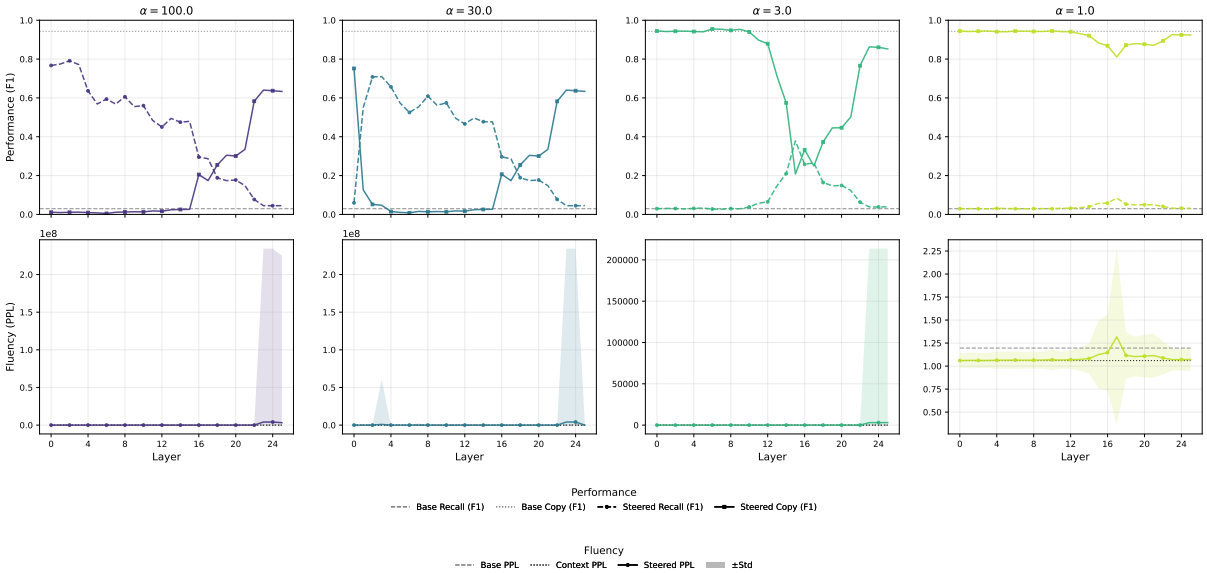


Figure 67: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.6.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

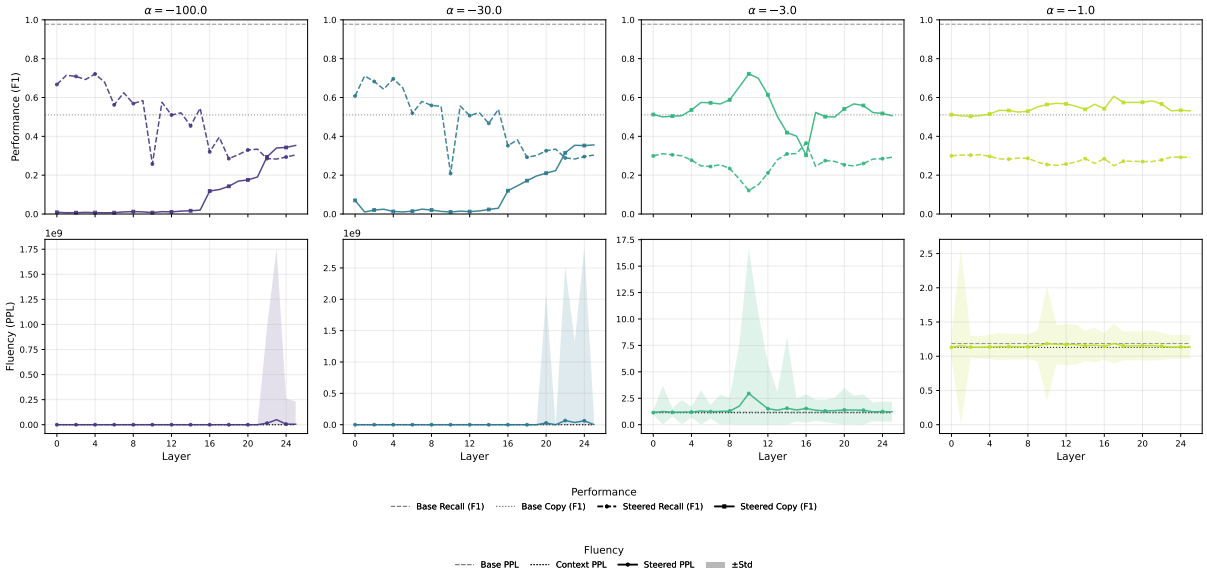


Figure 68: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.6.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

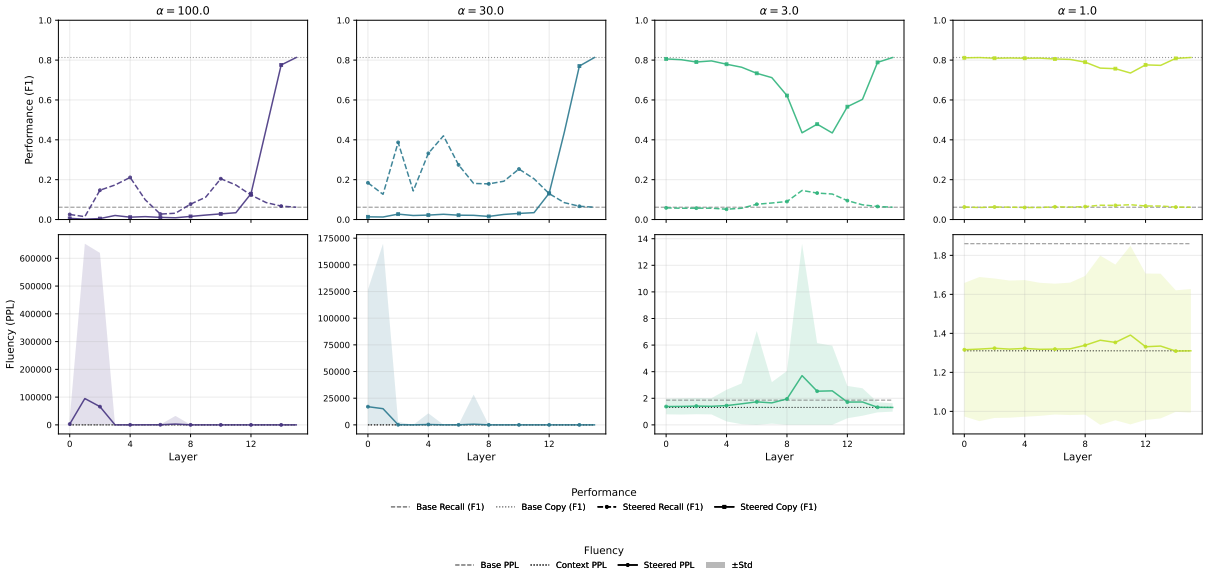


Figure 69: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.6.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

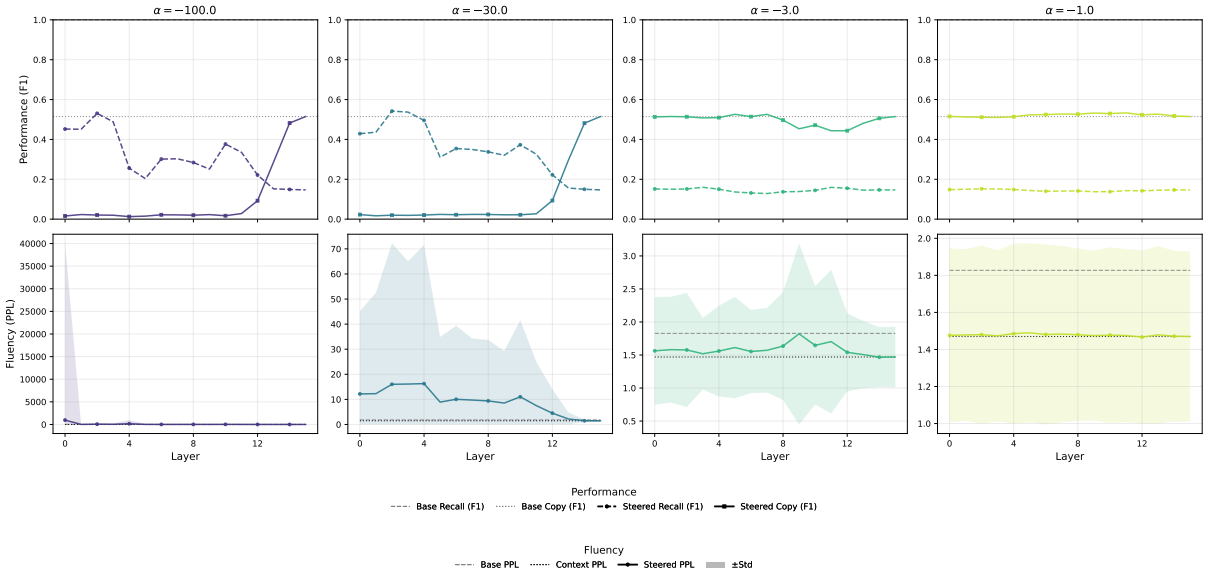


Figure 70: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.7 PEQ: Object Token (Context-First)

E.7.1 Parametric Steering (Copy→Recall) - Gemma2-2B

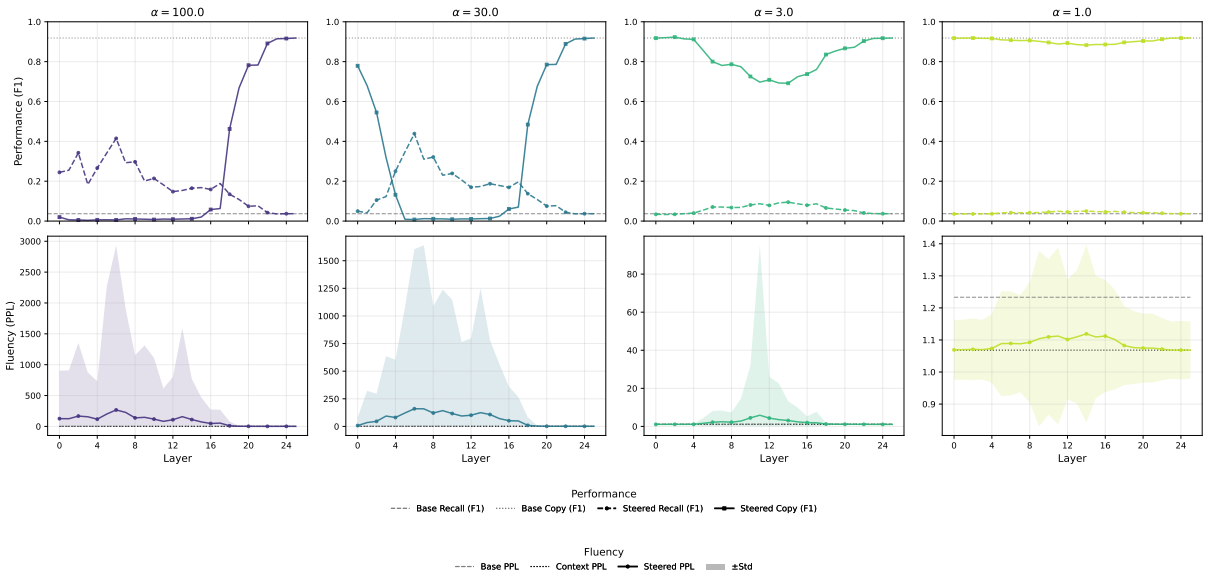


Figure 71: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.7.2 Contextual Steering (Recall→Copy) - Gemma2-2B

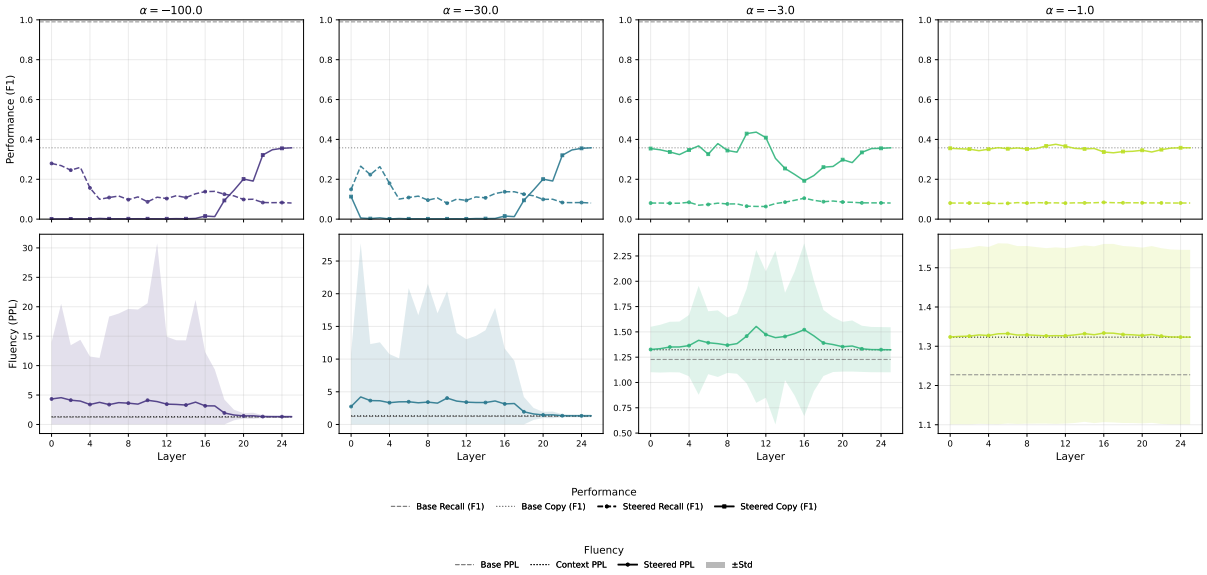


Figure 72: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.7.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

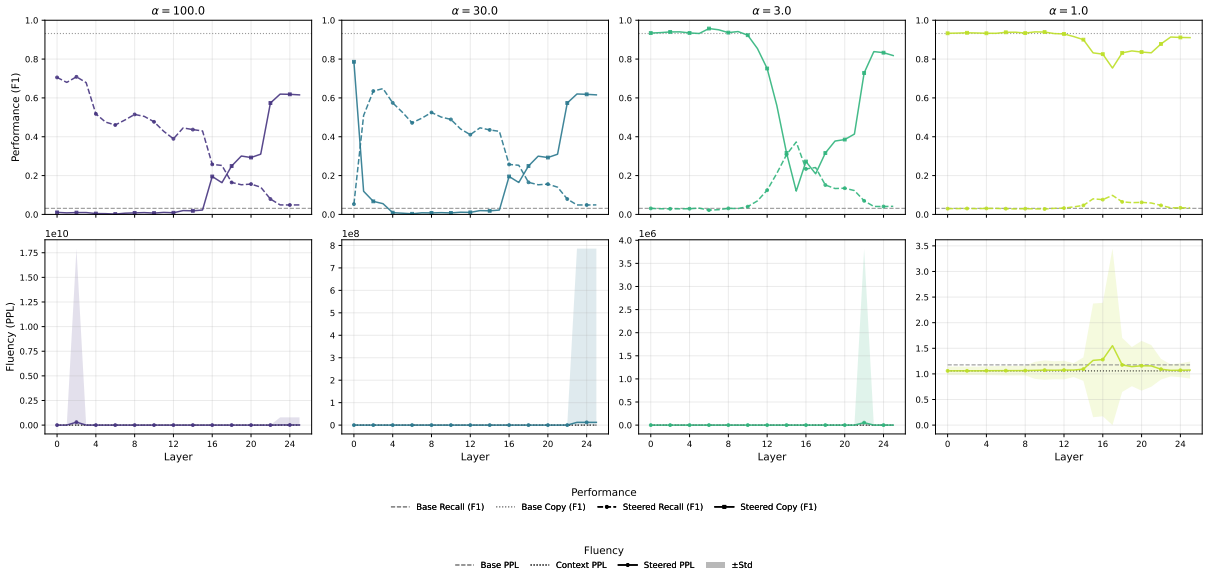


Figure 73: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.7.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

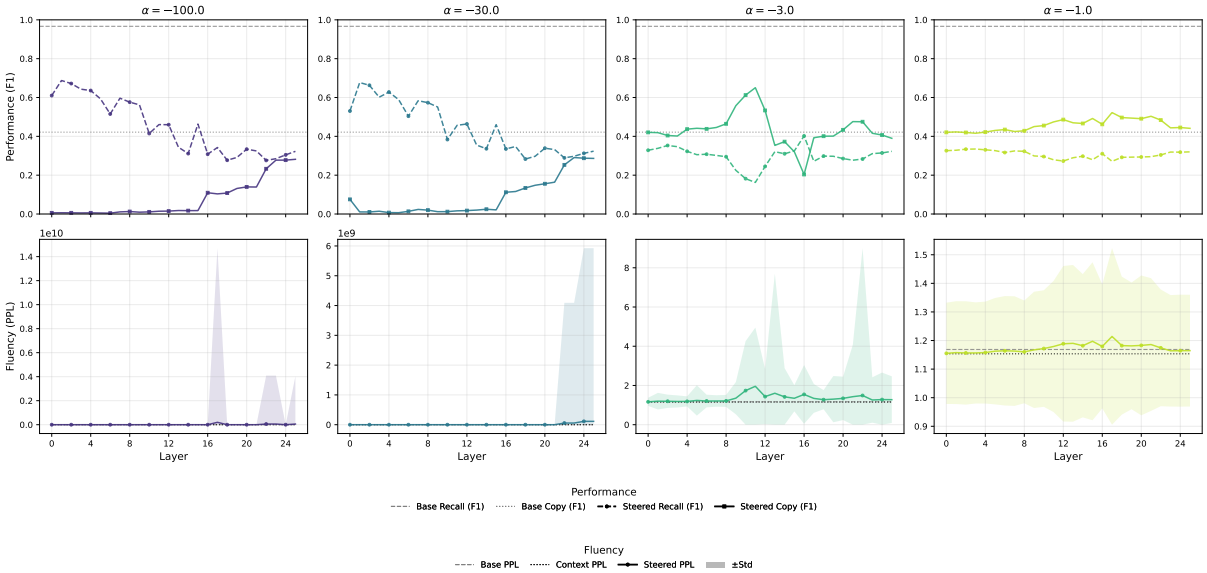


Figure 74: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.7.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

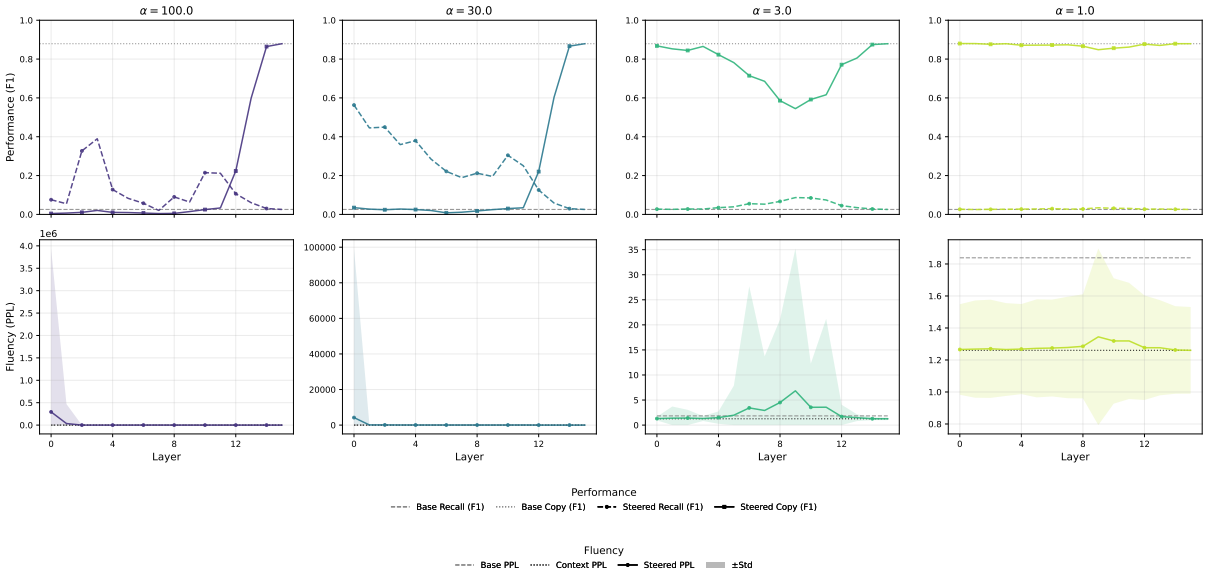


Figure 75: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.7.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

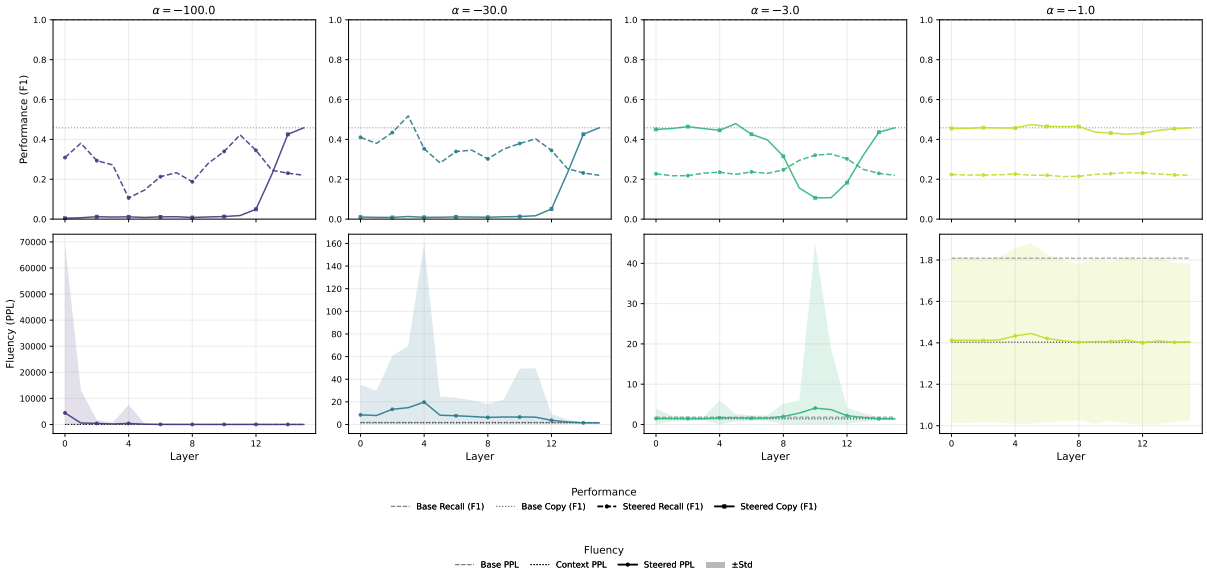


Figure 76: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.8 PEQ: Subject Token (Context-First)

E.8.1 Parametric Steering (Copy→Recall) - Gemma2-2B

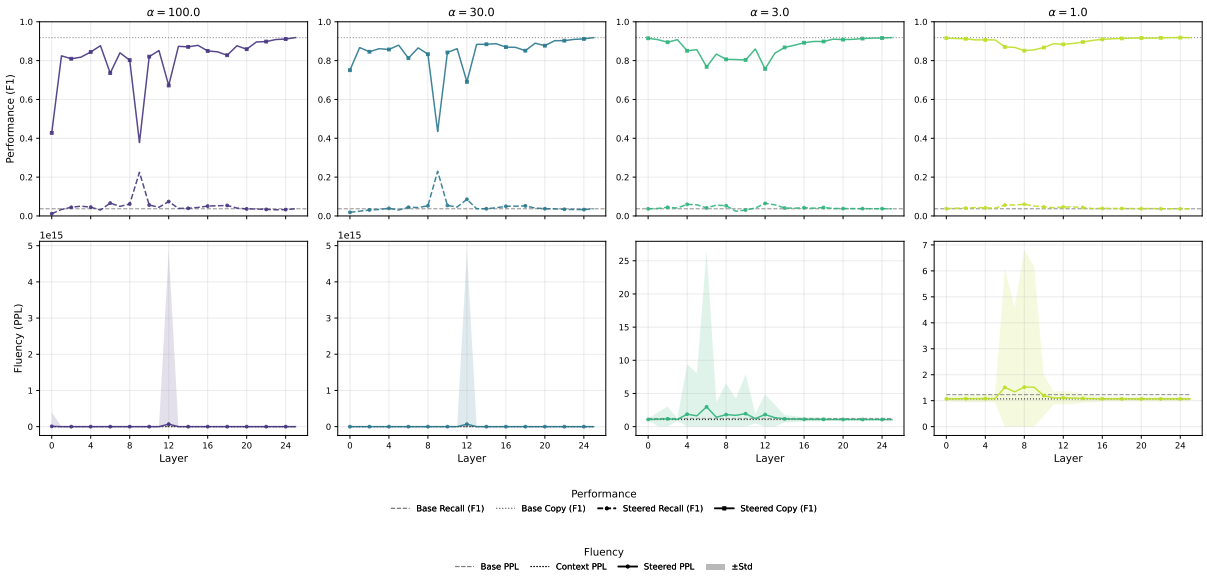


Figure 77: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.8.2 Contextual Steering (Recall→Copy) - Gemma2-2B

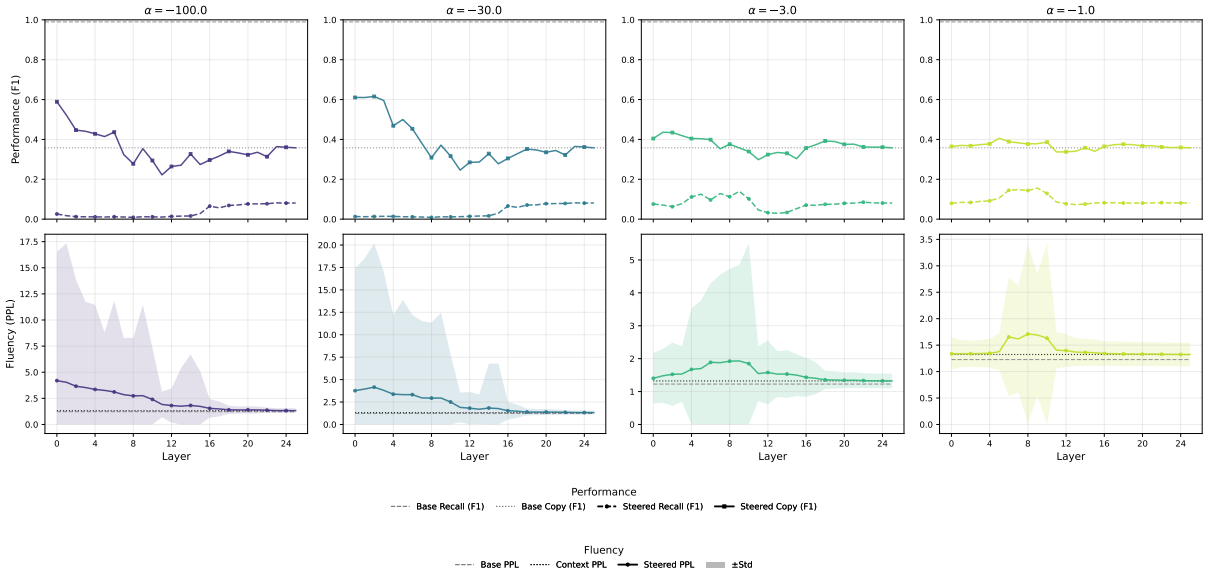


Figure 78: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.8.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

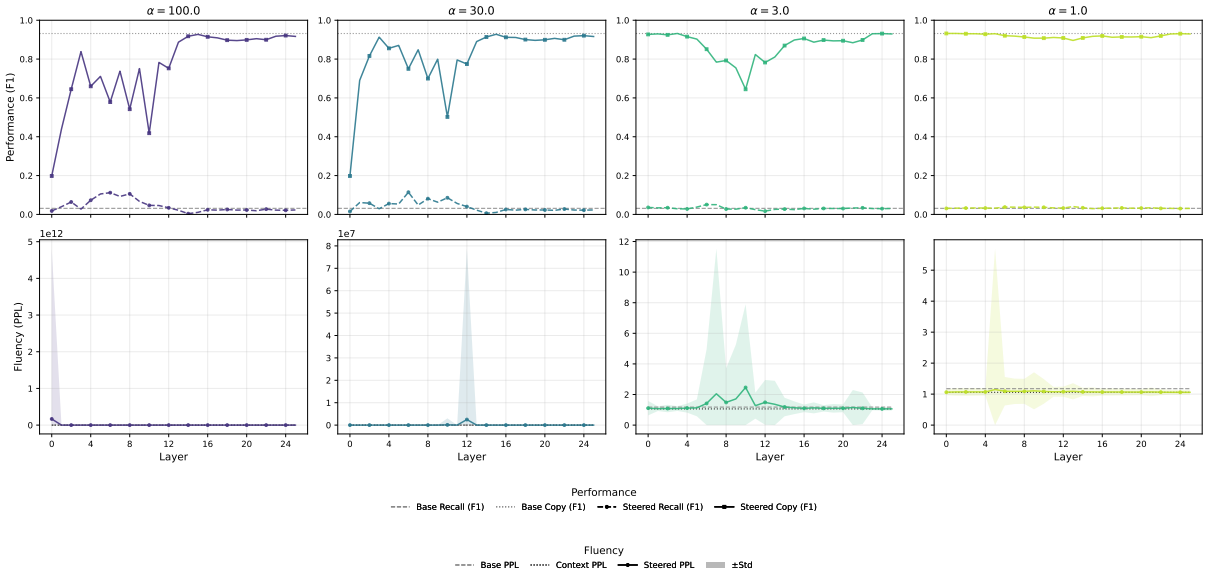


Figure 79: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.8.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

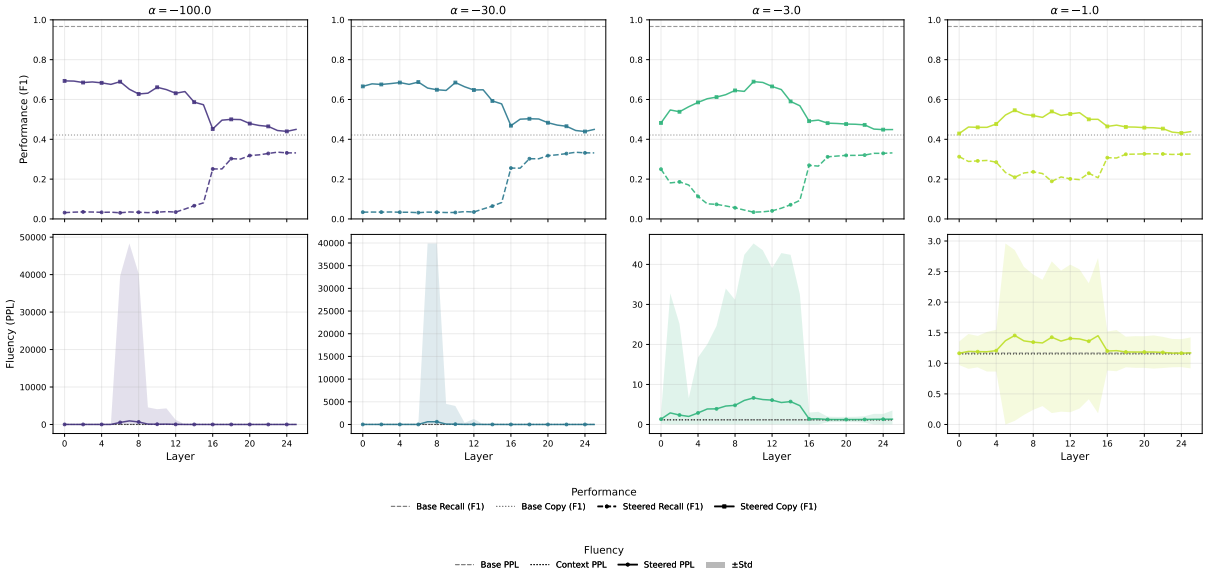


Figure 80: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.8.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

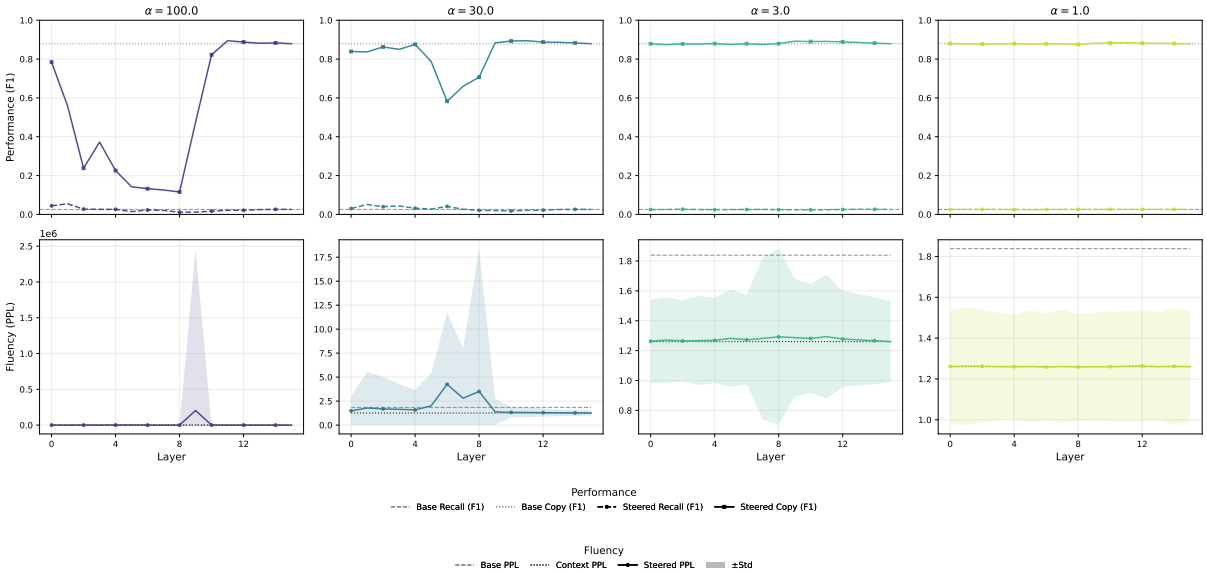


Figure 81: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.8.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

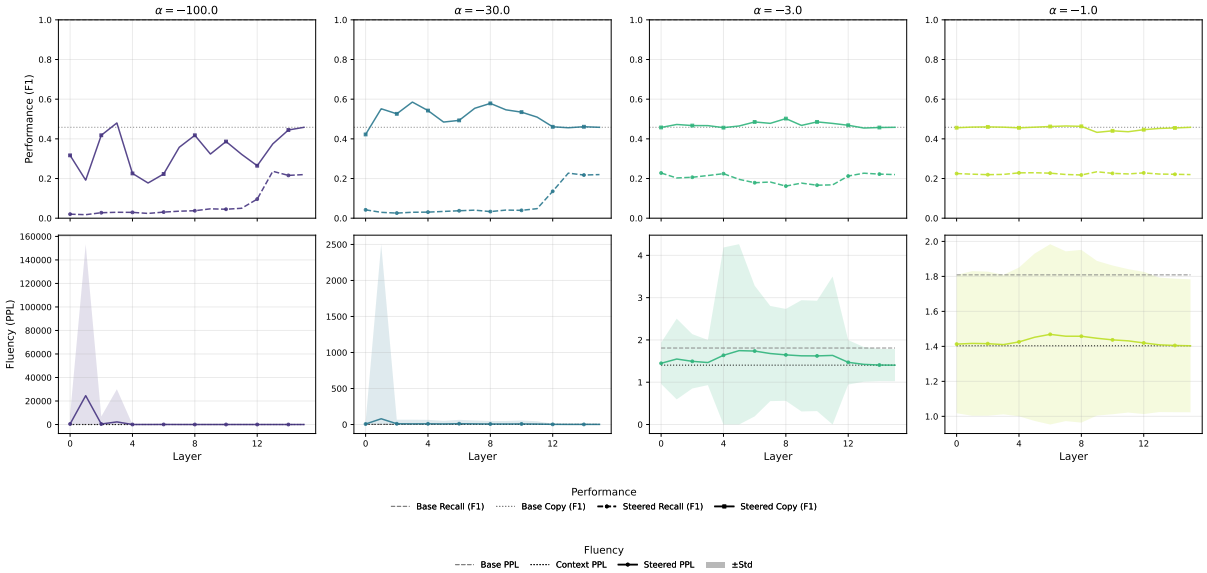


Figure 82: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.9 PEQ: Last Token (Query-First)

E.9.1 Parametric Steering (Copy→Recall) - Gemma2-2B

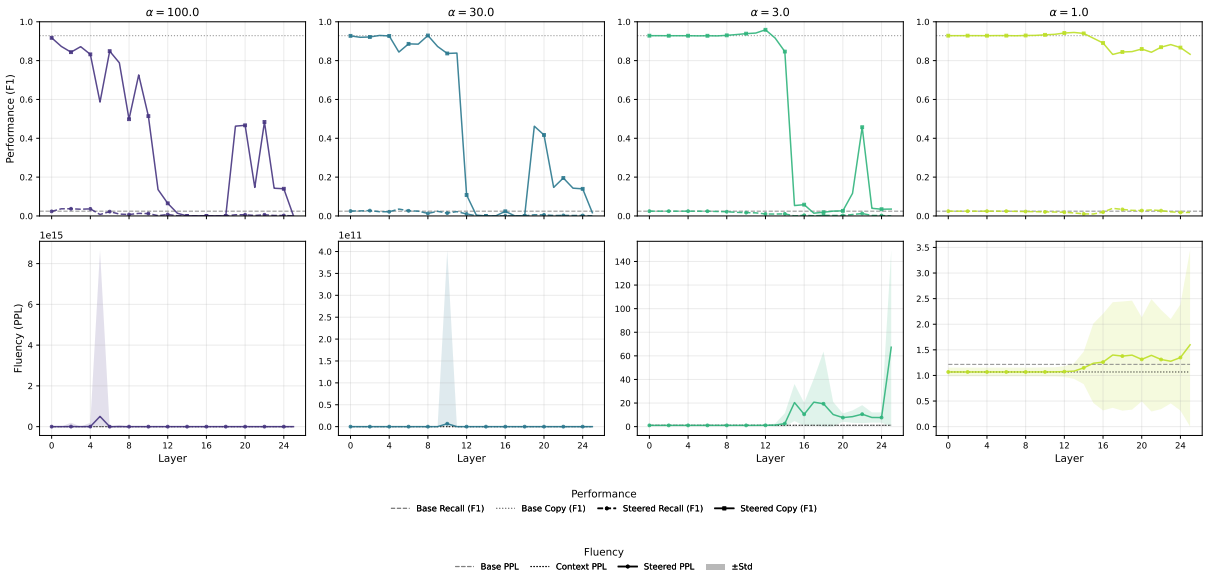


Figure 83: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.9.2 Contextual Steering (Recall→Copy) - Gemma2-2B

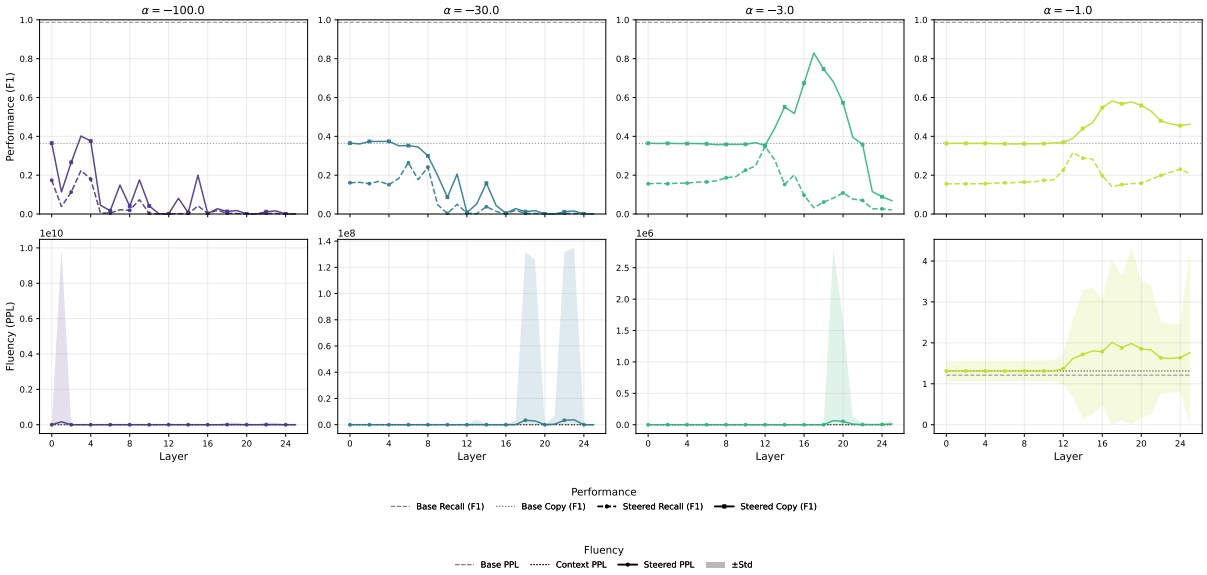


Figure 84: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.9.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

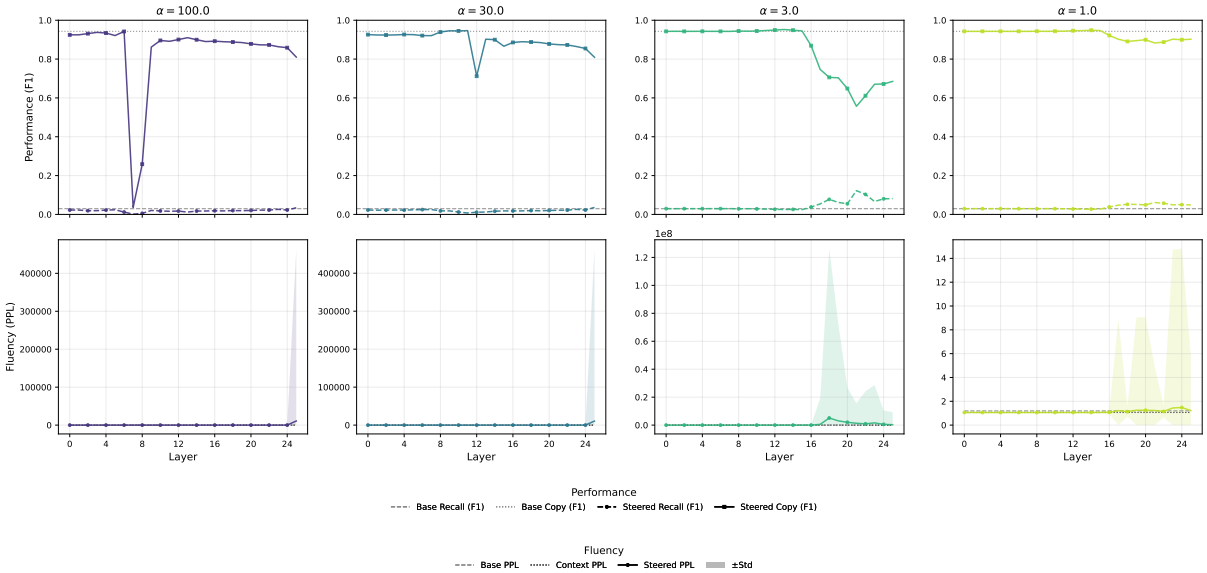


Figure 85: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.9.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

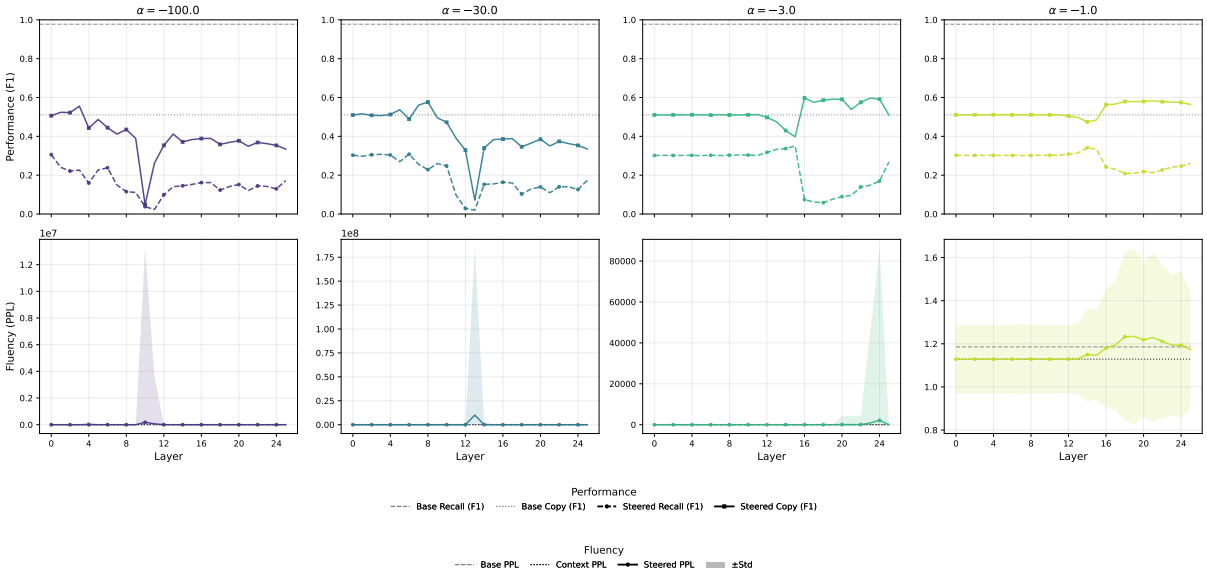


Figure 86: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.9.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

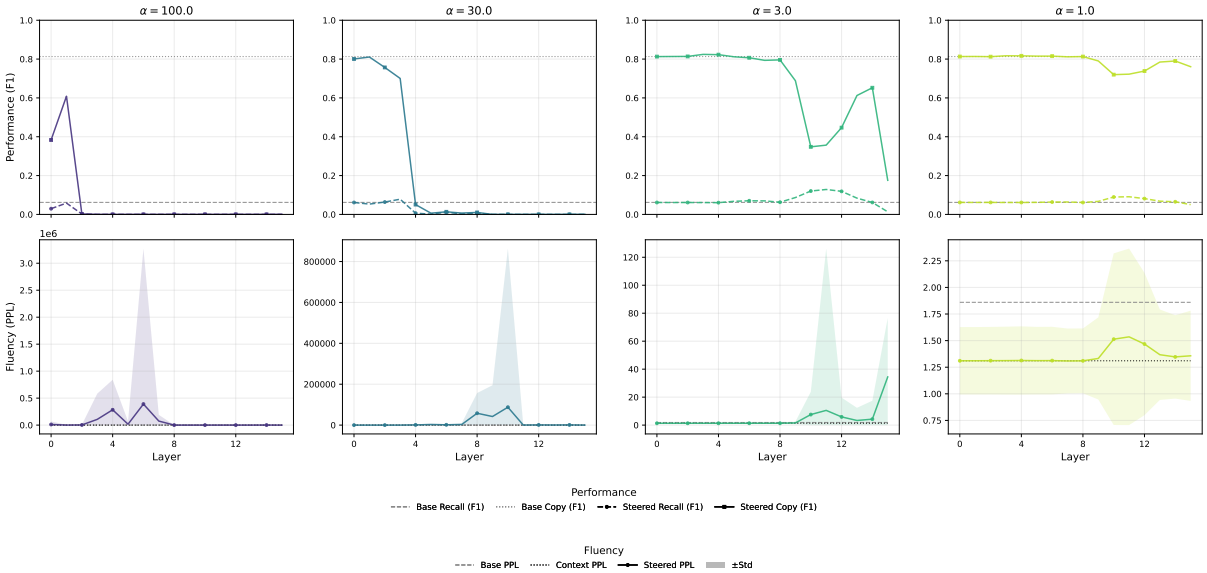


Figure 87: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.9.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

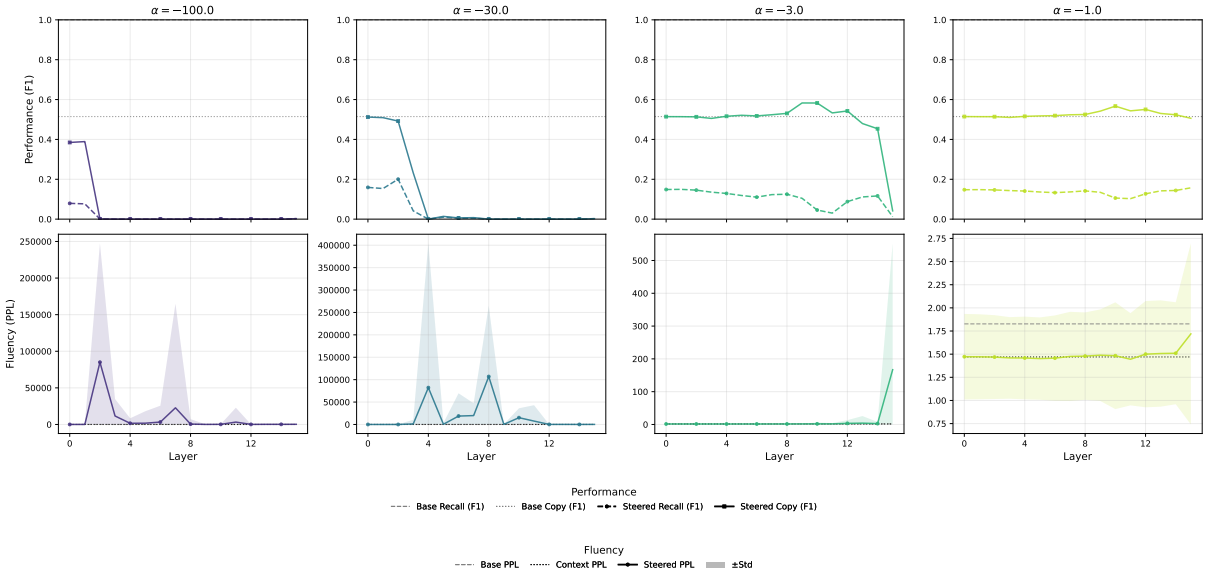


Figure 88: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.10 PEQ: Last Token (Context-First)

E.10.1 Parametric Steering (Copy→Recall) - Gemma2-2B

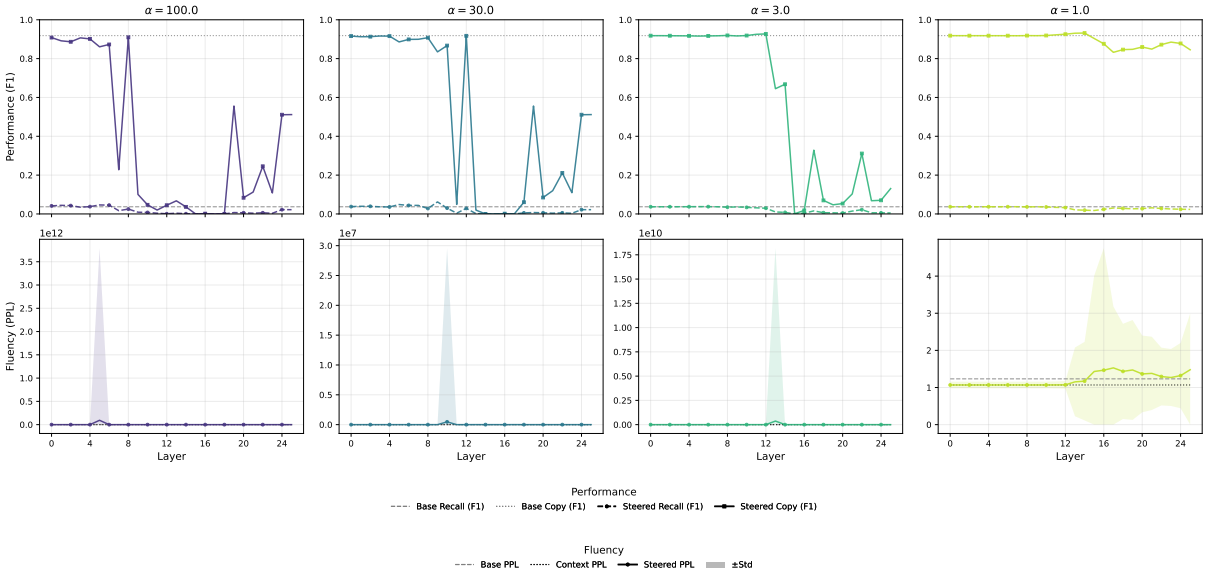


Figure 89: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.10.2 Contextual Steering (Recall→Copy) - Gemma2-2B

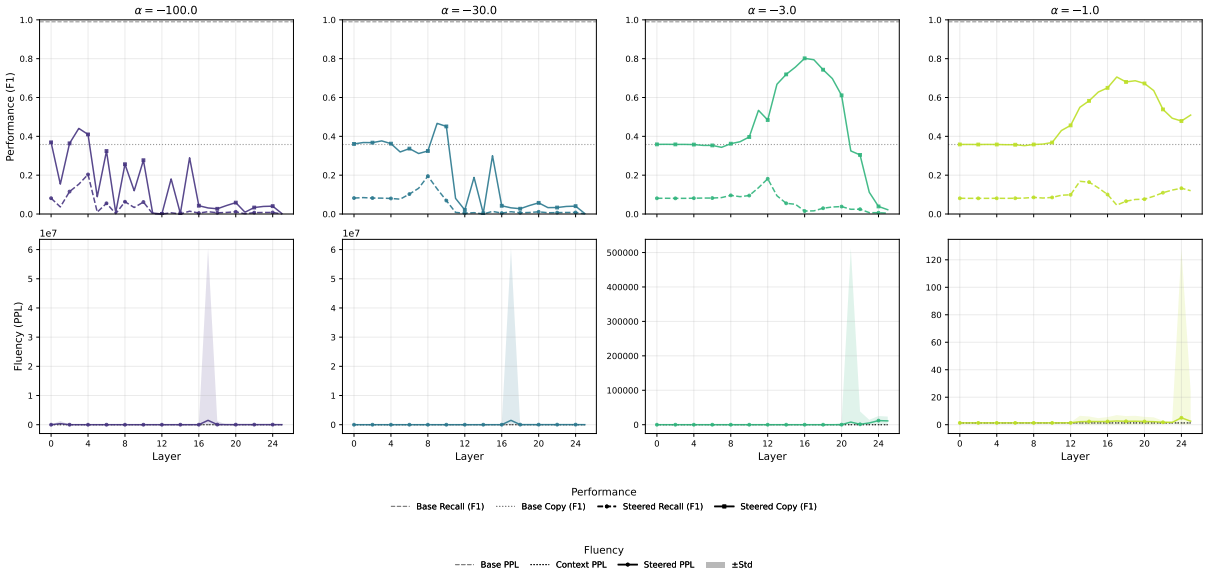


Figure 90: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.10.3 Parametric Steering (Copy→Recall) - T5Gemma-2B

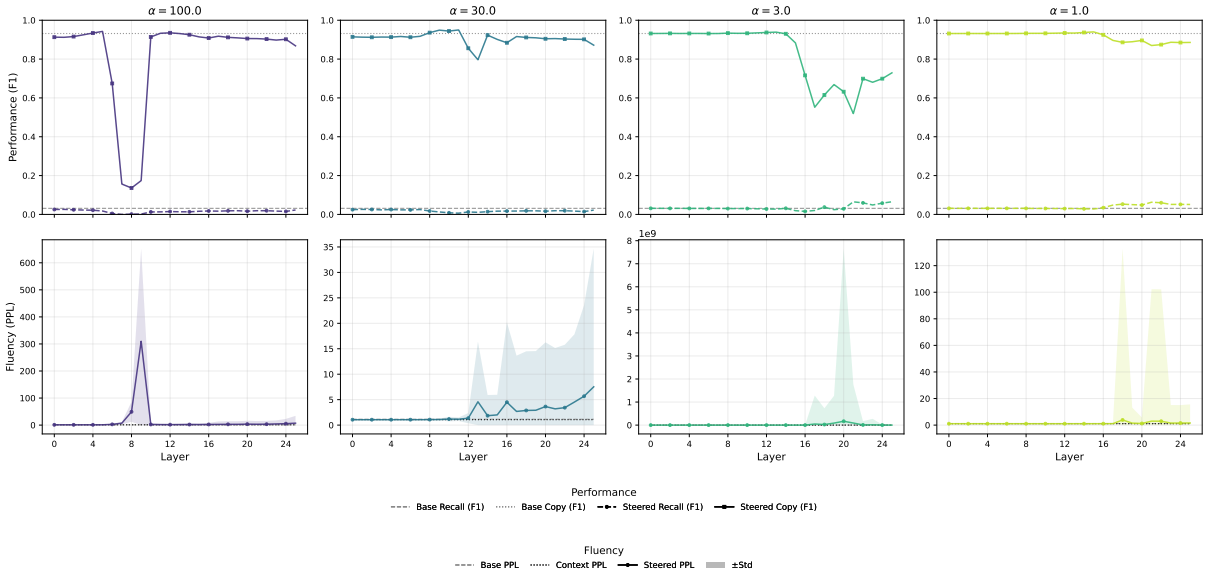


Figure 91: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.10.4 Contextual Steering (Recall→Copy) - T5Gemma-2B

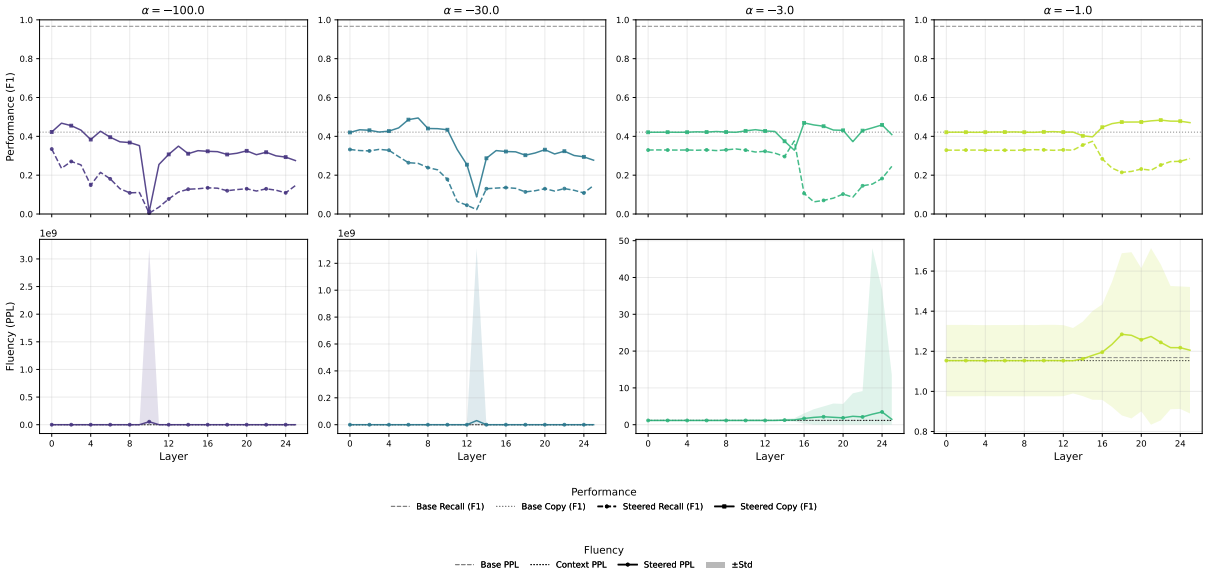


Figure 92: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.10.5 Parametric Steering (Copy→Recall) - OLMo-2-0425-1B

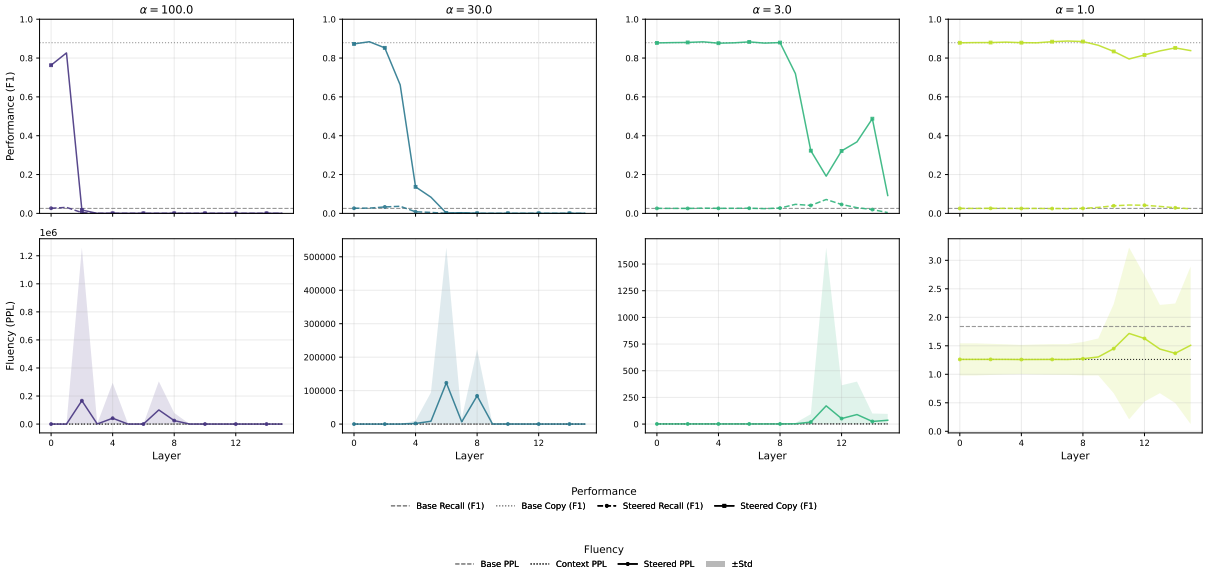


Figure 93: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.

E.10.6 Contextual Steering (Recall→Copy) - OLMo-2-0425-1B

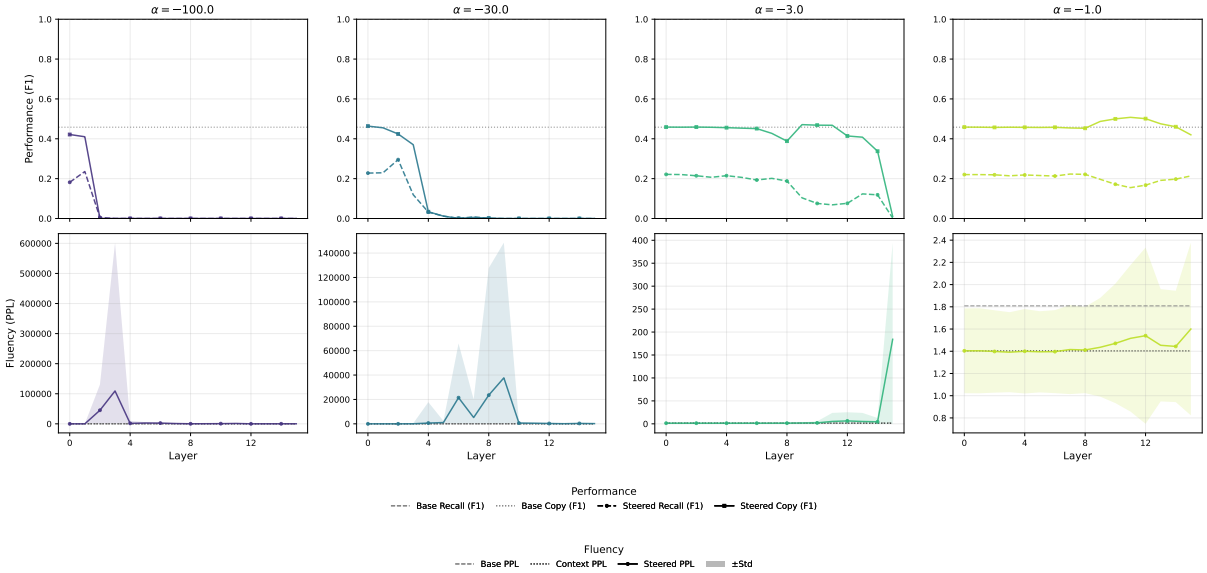


Figure 94: Layer-wise effects of steering across four α. Rows show performance (F1) and fluency (PPL). Columns correspond to α.