

# WHAT’S THE PLAN? METRICS FOR IMPLICIT PLANNING IN LLMs AND THEIR APPLICATION TO RHYME GENERATION AND QUESTION ANSWERING

Anonymous authors

Paper under double-blind review

## ABSTRACT

Prior work suggests that language models, while trained on next token prediction, show implicit planning behavior: they may select the next token in preparation to a predicted future token, such as a likely rhyming word, as supported by a prior qualitative study of Claude 3.5 Haiku using a cross-layer transcoder. We propose much simpler techniques for assessing implicit planning in language models. With case studies on rhyme poetry generation and question answering, we demonstrate that our methodology easily scales to many models. Across models, we find that the generated rhyme (e.g. “-ight”) or answer to a question (“whale”) can be manipulated by steering at the end of the preceding line with a vector, affecting the generation of intermediate tokens leading up to the rhyme or answer word. We show that implicit planning is a universal mechanism, present in smaller models than previously thought, starting from 1B parameters. Our methodology offers a widely applicable direct way to study implicit planning abilities of LLMs. More broadly, understanding planning abilities of language models can inform decisions in AI safety and control.

## 1 INTRODUCTION

Does the remarkable ability of modern language models in generating coherent text result from some form of implicit planning? Our paper develops methods for investigating this question, and applies them to the case studies of rhymed poetry and question answering.

In language production, humans are known to use various strategies to plan ahead for words that they utter (Ferreira & Swets, 2002; Lee et al., 2013; Huettig, 2015; Barthel et al., 2016). Similar mechanisms could plausibly apply in Transformer language models. In the context of text generation with language models, *planning* for goal token(s) is a process with two stages, forward and backward:

**Definition 1.1.** Forward planning is the creation of planning representations at earlier positions that encode properties of goal token(s) at later positions.

**Definition 1.2.** Backward planning is using those planning representations for generating intermediate tokens before the goal token(s) are produced.

Forward and backward planning are parts of the same overarching mechanism: the former creates a representation of the goal, and the latter conditions intermediate generation on that representation.

**Definition 1.3.** Planning is **implicit** if planning representations are part of hidden activations, and **explicit** if planning representations are model outputs.

**Definition 1.4.** Planning is **successful** when (i) forward planning creates representations that causally implicate the generation of goal tokens and (ii) backward planning produces intermediate tokens that increase the likelihood of goal token(s).

By its nature, *implicit* planning is harder to study than explicit planning, e.g. laying out the plan for a task in a chain of thought. To establish implicit planning behavior, we need evidence of both successful forward planning and backward planning. In this study, we focus on phenomena where

planning ahead for the kinds of words to produce is particularly transparent: rhyming, as suggested by the pioneering case study in Lindsey et al. (2025), and question answering.

Rhyming and question answering offer a unique window into planning, because it is linked to text elements (rhyming words or answer words) whose nature and position are predictable from general principles but not determined by immediately preceding tokens. In poetry, planning for a rhyme manifests as: (i) representations of the target rhyme family emerging at the end of the first line (forward planning), and (ii) intermediate tokens in the second line that facilitate reaching that rhyme (backward planning). In question answering, planning manifests as: (i) representations of the planned answer (e.g. *elephant*) emerging at the end of the question (forward planning), and (ii) intermediate tokens (e.g. *an*) in the answer that lead towards the planned answer (backward planning).

The contributions of our work include: (i) Datasets for the study of planning in rhyming (1050 lines from 10 rhyme families) and question answering (500 questions for 20 intended answer nouns). (ii) Quantitative metrics for assessing aspects of planning behavior: successful forward planning, backward planning, and successful backward planning. (iii) Evidence for implicit planning for the broadest range of language models to date, 23 diverse open weight language models (from 1B to 32B parameters).

Our study uses localized steering intervention for both tasks. We show that mean activation difference steering works across the board, altering forward and backward planning. We find a lot of similarities, but also some differences between models we study (cf. Appendix H and M).

In what follows, section 2 provides the scientific context; section 3 explains the details of our methodology and experimental setup. Section 4 compares language models according to our metrics and discusses observations on the planning circuits, which parallel Lindsey et al.’s findings for Haiku. We summarize our work and discuss its limitations and further directions in section 5.

## 2 RELATED WORK

One piece of evidence for implicit planning in language models comes from the studies of unreliable chain of thought (CoT): “When we bias models toward incorrect answers, they frequently generate CoT explanations rationalizing those answers” Turpin et al. (2023); this also applies to newer reasoning models (Chen et al., 2025). When rationalizing, the model must be entertaining the suggested answer before the rationalizing CoT is generated. So a critical output can be implicitly planned long before it is produced, an instance of implicit *forward* planning. Furthermore, intermediate steps leading up to the answer (in this case, the chain of thought) are conditioned on the implicit plan, an instance of *backward* planning.

As a response to reports on LLMs producing useful plans in domains such as coding (Bairi et al., 2023, e.g.), skeptical arguments about the general planning ability of LLMs have been raised in the literature. Various studies (Valmeekam et al., 2023; Zhang et al., 2024; Stein et al., 2025) show empirically that explicit planning abilities of current LLMs are limited. The conceptual argument that “a system that takes constant time to produce the next token cannot possibly be doing principled reasoning on its own” (Kambhampati et al., 2024) naturally applies to implicit planning as well.

However, LLMs can plausibly engage in some forms of simpler, heuristic type of planning – just like humans, whose general planning capacities are limited (Kahneman & Tversky, 1977; Buehler et al., 2010), are known to still plan ahead in their speech production (Brown-Schmidt & Konopka, 2015). Indeed, shaping the rhythm and rhyme for the next poetic line is a task that can be solved well enough with limited planning capacities. This is supported by the success of previous poetry generation experiments that used relatively simple tools Hopkins & Kiela (2017); Lau et al. (2018); Ormazabal et al. (2022); Jhamtani et al. (2019); Ghazvininejad et al. (2016).

Implicit planning representations in LLMs have been found for specific models (Pochinkov et al., 2024; Men et al., 2024; Wu et al., 2024). In this paper, we aim to compare planning across a wider range of language models. One of our case studies is on rhymed poetry generation, inspired by Lindsey et al. (2025), the first study of implicit planning in poetry generation. The authors prompted Claude Haiku 3.5 with

Example 1. A rhyming couplet:\n He saw a carrot and had to grab it\n

The model was able to complete the second line with the correct rhyme, e.g. *His hunger was like a starving rabbit*. Furthermore, elements of *implicit forward planning* mechanisms were identified: there are vector components in the activations of the token at the end of the first rhyming line (second \n above) that Lindsey et al. identify as representing planning for the rhyme of the next line. Some of these planning activation components, they argue, correspond to the potential rhyme *rabbit* and, if suppressed, lead to other rhyming words (such as *habit*) being generated much more often. It is even possible to inject a planning direction for a word such as *green* that does not fit the context, causing the model to produce it instead of a correct rhyme.

In addition, Lindsey et al. found evidence of *backward planning* in the rhyming context. The planning directions, they argue, influence the intermediate words. If *rabbit* is the dominant planned rhyme word, certain constructions are produced, e.g. comparison construction as in *His hunger was like a starving rabbit*. If *rabbit* planning directions are artificially suppressed, different intermediate parts of the line are generated, such as *His hunger was a powerful habit*.

Lindsey et al.’s evidence is limited to a small number of persuasive examples and a single closed-source language model. Because of its proprietary nature and the technical complexity of the approach, Lindsey et al.’s findings cannot be independently reproduced. Our paper addresses these limitations. First, we assess forward and backward planning in both rhyming and question answering across a range of models. Second, we introduce several quantitative metrics based on a varied dataset of rhymes. Third, we address the complexity problem: the cross-layer transcoder (CLT) approach used by Lindsey et al., while offering diverse interpretability promises, is particularly complex, computationally expensive, and difficult to replicate. Just training a CLT for a single model of modest size requires days of compute on a highly performant and expensive GPU such as H100 (Ameisen et al., 2025, Appendix D).

As key intervention, we employ activation steering with vector addition, as in Turner et al. (2023). Directly manipulating model’s activations, activation steering differs from other ways of controlling outputs such as prompting or different variants of controlled or constrained generation (Kumar et al., 2022; Lew et al., 2023; Loula et al., 2025). Unlike in constrained generation that can use phonetic constraints relevant for rhyming (Roush et al., 2022) or evolutionary optimization with rhyme scoring (Jobanputra et al., 2025), we do not constrain the model’s outputs but modify the rhyming representations from which the language model generates the next line normally. As an implicit control mechanism, activation steering shares some conceptual similarity with classifier-free guidance in diffusion models (Ho & Salimans, 2022). However, unlike classifier-free guidance which operates during the iterative sampling process, our approach intervenes on a single position’s representations in autoregressive LLM generation, making it more directly interpretable.

### 3 METHODOLOGY

Taking inspiration in the observations of Lindsey et al., we propose several metrics for quantifying implicit forward and backward planning, and apply them to a variety of language models. We propose simpler methods that do not involve the costly training and use of cross-layer transcoders, or other types of dictionary learning. We focus here on planning for a noun answer in question answering and for a rhyme family in poetry; see also Appendix H for a discussion of models planning to produce specific rhyming words.

**3.1 Data** In **rhyming** experiments, we call the set of words that all rhyme with each other a **rhyme family**. Words in a rhyme family tend to share a suffix, which we use to name the rhyme family. For example, the *-ing* rhyme family contains words such as *king* and *ring*. We chose 10 rhyme families and 20 pairs of rhyme families such that every rhyme family appears as the first in two pairs and the last in two pairs. For each rhyme family (RF), we generated 105 first lines using Claude 3.5 Sonnet. We split the lines into a train set  $\mathbf{P}_{\text{RF}}^{\text{Train}}$  (85 lines) and test set  $\mathbf{P}_{\text{RF}}^{\text{Test}}$  (20 lines). In practice, less data for estimating the steering vector was often enough, see Appendix B.

For **question answering** experiments, we created a dataset of 20 **noun pairs** where one noun begins with a vowel (requiring article “an”) and the other with a consonant (requiring article “a”), such as (*eye*, *heart*). For each noun, we created 13 suggestive questions for training and 5 for testing, plus 7 neutral questions answerable by both nouns in a pair. We use notation  $\mathbf{P}_{\text{N}}^{\text{Train}}$  and  $\mathbf{P}_{\text{N}}^{\text{Test}}$  analogously to the rhyme family datasets. See Appendix A for complete dataset details for both tasks.

**3.2 Models** We test language models from four families (Gemma2, Gemma3, Qwen3 and Llama3.1/3.2) with model sizes between 1B and 32B parameters Team et al. (2024; 2025); Yang et al. (2025); Grattafiori et al. (2024). For each model we test both the base version and the instruction tuned versions.<sup>1</sup>

**3.3 Mean Activation Steering and Choosing a Steering Vector** If a model exhibits implicit forward planning, then an intervention on an early position (such as the newline in 1, or the last word of the first line) can alter the rhyme or the answer generated multiple tokens later.

We estimate a steering vector using the average activation difference like in Arditi et al. (2024). We calculate the average activation of one of the last positions of the prompt: the last newline (`\n`) token in the prompt (*newline steering*), the last word of before that (*last word steering*), or the question mark (question mark steering for questions). For example, in the rhyming context, we can subtract the average activation in lines that rhyme with *sick* from the average activation of the same position for lines that rhyme with *pain*. The resulting steering vector, when applied to the respective newline or last word token in the line *The house was built with sturdy, reddish brick \n*, can lead to ‘rhymed’ generations like *And stood for years, enduring wind and rain* instead of the unsteered baseline version *And stood for years, enduring every trick*.

Steering operates on a pair of prompt categories ( $C_1, C_2$ ) such as rhyme families (RF) or noun answers (N). For each model, layer  $l$ , steering position, and prompt category pair ( $C_1, C_2$ ) we extract a steering vector  $s_{C_1 \rightarrow C_2}^{(l,i)}$  by calculating the mean difference between the activations on their train sets

$$s_{C_1 \rightarrow C_2}^{(l,i)} = m \cdot \left( \sum_{p \in \mathbf{P}_{C_1}^{\text{Train}}} \mathbf{x}_i^{(l)}(p) - \sum_{p \in \mathbf{P}_{C_2}^{\text{Train}}} \mathbf{x}_i^{(l)}(p) \right)$$

Here  $m$  is a constant, set to 1.5,<sup>2</sup>  $\mathbf{x}_k^{(l)}$  is the hidden activation of the model at layer  $l$  and position  $k$ , and  $i$  is either **pos**(`\n`) (the position of the last newline token in  $p$ ) or **pos**(?) or the position of the last token before the newline or the question mark.

To apply a steering vector, we add it to the residual stream  $\mathbf{x}_k^{(l)}$  on the correct layer and token position during generation. Note that **we apply the steering vector on one token only** (the last word, **question mark** or the newline token).

For the final steering vector of each model and category pair  $s_{C_1 \rightarrow C_2}$  we choose out of all layers and positions the combination where steering has the maximum intended effect.

While we opt for average activation difference steering because of its simplicity, it is also possible to obtain a steering vector with other methods, such as differences of classifier probe weights or SAE weights. For examples of generated outputs with alternative steering vectors, see Appendix I.

**3.4 Rhyming Metrics** Let  $\mathbf{C}_{\text{RF}}$  be a collection of 1000 couplets generated by some model using the prompts in  $\mathbf{P}_{\text{RF}}^{\text{Test}}$ . We sample 50 generations per prompt  $p \in \mathbf{P}_{\text{RF}}^{\text{Test}}$ .

Let  $\mathbf{C}_{\text{RF}_1 \rightarrow \text{RF}_2}$  be a collection of 1000 couplets generated by some model using the prompts in  $\mathbf{P}_{\text{RF}}^{\text{Test}}$  while being steered with  $s_{\text{RF}_1 \rightarrow \text{RF}_2}$ .

As evidence *successful forward planning*, we use effectiveness of the steering intervention: if at position X there was no planning representation for a later position Y, then intervening on position X would not have changed the outcomes for position Y in a predictable way. This is done by comparing the following two metrics.

**Fraction of Correct Rhyme Family.** We calculate the fraction of couplets in  $\mathbf{C}_{\text{RF}}$ , where the last word has the correct rhyme family (rhymes with the last word of the first). We made collections containing all words in each rhyme family to do this.

<sup>1</sup>Only Qwen3 32B did not have an open weight base model available.

<sup>2</sup>Mean activation difference estimation defines an inherent scale: steering vector can be used with multiplier of 1. While that often works, we found that slightly bigger values such as 1.5 or 2 produce a more consistent effect. Informally, imprecision in the estimation of the steering vector is compensated by increasing its magnitude.



**Fraction of Correct Rhyme Family (Steered).** This metric is calculated the same way as the Fraction of Correct Rhyme Family metric. We evaluate it separately for each rhyme family pair  $(\mathbf{RF}_1, \mathbf{RF}_2)$  on  $\mathbf{C}_{\mathbf{RF}_1 \rightarrow \mathbf{RF}_2}$ . The correct rhyme family is the one that was steered towards  $(\mathbf{RF}_2)$ .

Successful steering means that planning representations can be manipulated successfully. This allows us to assess *backward planning* by measuring to what extent interventions on the planning activations affect the model’s behaviors at intermediate positions. For example, Gemma2 9B model steered towards the *-ight* rhyme family will not just end the second line in 2 with a different word like *light* instead of *sing*, but will likely take a different path after *above*, leading to a more natural ending:

*Example 2.* Whispers of freedom found in a bird’s wing \n Soaring above where true joy will sing  
**Steered towards -ight:** Soaring above bathed in a golden light

The likelihood of different intermediate continuations due to steering is reflected in a shift in the probability distribution over the next token. To measure this effect, we propose the following *probability based metrics* discussed in detail in Appendix F.

We measure *successful backward planning* in poetry—to what extent words generated in the task’s context leads up to the planned completion—using *regeneration metrics*. For example, we take the line *And stood for years, enduring every trick* and regenerate the last word without the original context that conditioned the rhyme. If backward planning involving the rhyming plan was used, such regeneration is expected to produce *trick* at a higher rate than for generated lines in other contexts, and at a higher rate than for lines generated in the same context but steering towards a different rhyme.

**Fraction of Correct Last Word Regeneration.** We extract the second lines of all the couplets in  $\mathbf{C}_{\mathbf{RF}}$ . (So that there is no context that this is a rhyme). We also remove the last word in all of the second lines. Then we regenerate the last word using the resulting prompts and calculate the fraction of cases where the regenerated word is from the correct rhyme family  $\mathbf{RF}$ .

**Fraction of Correct Last Word Regeneration (Steered).** This metric is computed similarly to the Fraction of Correct Last Word Regeneration, but with couplets in  $\mathbf{C}_{\mathbf{RF}_1 \rightarrow \mathbf{RF}_2}$ , calculated as the fraction of cases where the regenerated word is from the target rhyme family  $\mathbf{RF}_2$ . We calculate this metric separately for each rhyme family pair  $(\mathbf{RF}_1, \mathbf{RF}_2)$  on  $\mathbf{C}_{\mathbf{RF}_1 \rightarrow \mathbf{RF}_2}$ .

### 3.5 Question answering metrics

**Fraction of Correct Answer** We calculate the fraction of answers in  $\mathbf{C}_{\mathbf{N}}$  which contain the intended answer word as a substring.

**Fraction of Correct Answer (Steered)** We calculate the fraction of answers which contain the intended answer word as a substring. We evaluate it separately for each noun pair  $(\mathbf{N}_1, \mathbf{N}_2)$  on  $\mathbf{C}_{\mathbf{N}_1 \rightarrow \mathbf{N}_2}$ . The correct answer is the one that was steered towards  $(\mathbf{N}_2)$ .

**Fraction of *a/an*** We calculate the fraction of answers in  $\mathbf{C}_{\mathbf{N}}$  which contain the token *a* or *an*.

**Fraction of *a/an* (steered)** We calculate the fraction of answers which contain the token *a* or *an*. We evaluate it separately for each noun pair  $(\mathbf{N}_1, \mathbf{N}_2)$  on  $\mathbf{C}_{\mathbf{N}_1 \rightarrow \mathbf{N}_2}$  to determine whether the fraction of *a* vs. *an* changes in the correct direction depending on  $(\mathbf{N}_2)$ .

## 4 RESULTS

**4.1 Basic rhyming behavior** We observe that models differ in rhyming abilities. Generally, bigger models rhyme more consistently than smaller ones, and instruction-tuned models rhyme better than their base versions (Fig. 2, solid bars). [These findings cannot be attributed to tokenization differences between models; see Appendix G for details.](#)

In all models, we found evidence of successful backward planning from regeneration metrics: lines generated in the context of a rhyme are likely to be completed by a word of the intended rhyme family even without the original rhyming context. For all models, the metric is above chance (cf. Fig. 1).

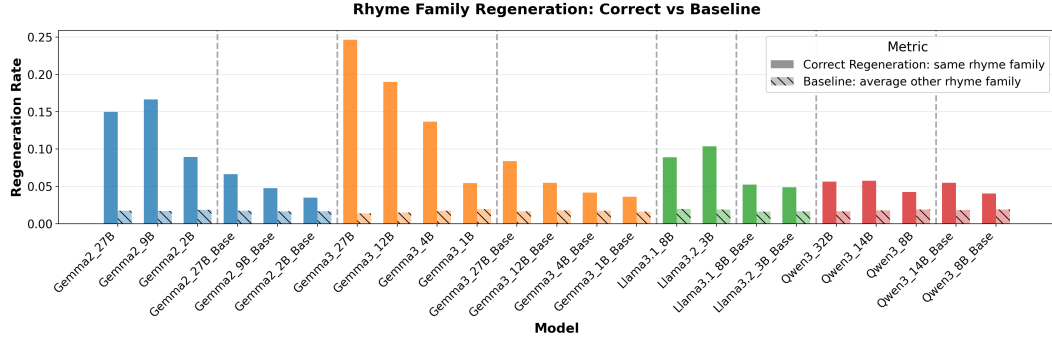


Figure 1: Comparison of unsteered regeneration rates to baseline chance level for the model. During regeneration with stochastic sampling, some rhyme families are expected to occur with non-zero frequency by chance. We estimate the baseline chance level that unsteered regeneration rate must exceed to show evidence for successful backward planning as the average frequency of rhyme families other than the one in whose context the couplet’s second line was originally generated by the model.

**4.2 Steering affects forward and backward planning in rhyming** Steering on the last word consistently modifies models’ behavior wrt the rhyme family generated. This supports that representations of the planned rhyme are localized at the intervention point. While rhyming abilities of different language models vary, our simple steering strategy achieved rates of the target rhyme family comparable to the rhyming rate of the model in the baseline condition, cf. Fig. 2. Only for models with the lowest rhyming ability (base variants of Gemma3 1B and Llama 3.2 3B) is the steered rhyming rate substantially lower than unsteered.

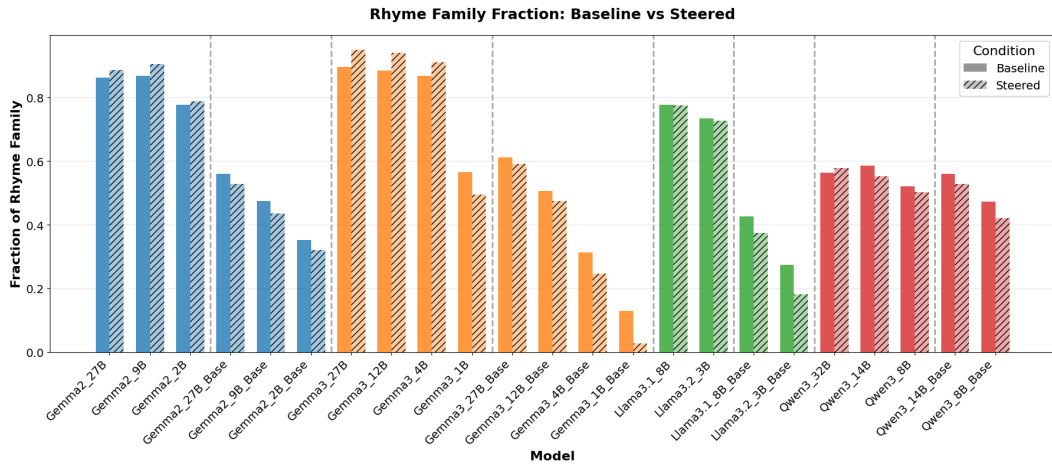


Figure 2: Baseline rhyming abilities of models vs. steered rhyming behavior. Solid bars report the frequency of (exact) rhymes produced by the model (baseline rhyming behavior). Hashed bars present the frequency of (exact) rhymes of rhyme family 2 after the first line ending in a word from rhyme family 1, when steered towards rhyme family 2.

Further, steered regeneration rates for the target rhyme family are close to the baseline regeneration rates for the same models (Fig. 3). This supports that the steering intervention does not just replace the final rhyming word but affects the backward planning that produces intermediate words.

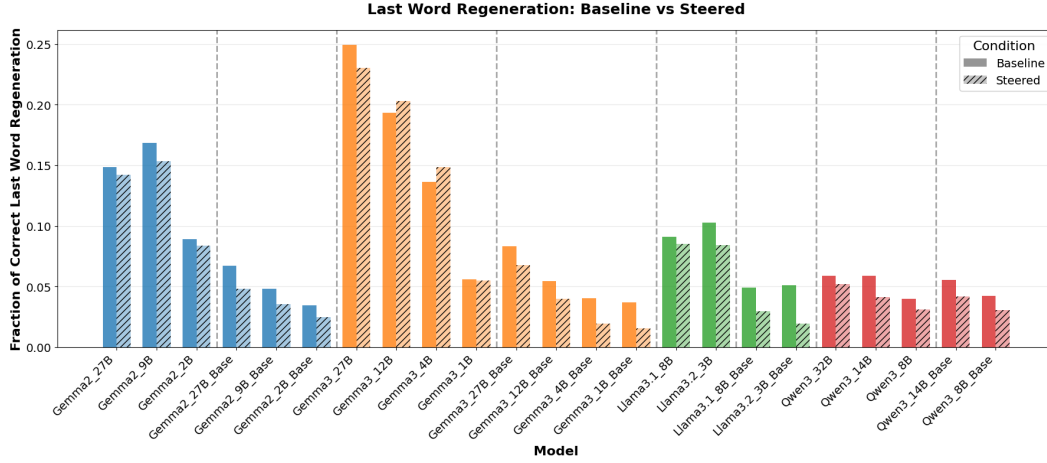


Figure 3: Baseline vs. steered last word regeneration rate of different models. Solid bars represent the fraction of times when regenerating the last word of the second line **with the first line removed** produces a word from the original rhyme family. This suggests that the model generates the second line so that it facilitates the correct rhyme, providing evidence for successful backward planning. Hashed bars represent the fraction of times when regenerating (with the first line removed) the last word of the second line **generated with steering towards rhyme family 2** produces a word from rhyme family 2. Close match between baseline and steered regeneration rates suggests that we successfully manipulate the LLM’s rhyme planning.

The success of steering in affecting intermediate planning is supported by the fact that in models with high rhyming capabilities, in the baseline vs. steered lines, the regeneration frequency distributions by rhyme family is close, as illustrated in Fig. 4 for Gemma3 27B.

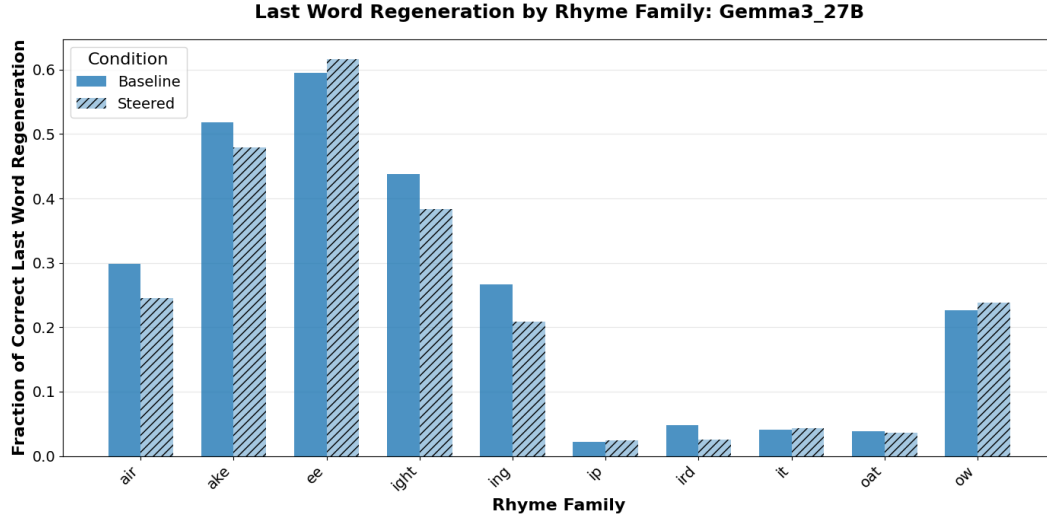


Figure 4: Regeneration rates per rhyme family with Gemma3 27B, baseline vs. steering. Solid bars represent the fraction of times when regenerating the last word of the second line **with the first line removed** produces a word from the original rhyme family. Hashed bars represent the fraction of times when regenerating (with the first line removed) the last word of the second line **generated with steering towards the rhyme family** produces a word from the target rhyme family. Close match between baseline and steered regeneration rates suggests that we successfully manipulate the LLM’s rhyme planning for specific families.

**4.3 Probability based metrics** Our probability based metrics detect evidence of backward planning on a finer level, assessing to what extent the probability distribution over intermediate tokens changes with this intervention. The probability based metrics generally follow the same patterns as the previous metrics: instruction tuned models tend to score higher than base models and bigger models tend to score higher than small ones. For plots, see (Fig. 14, 15) in Appendix F.

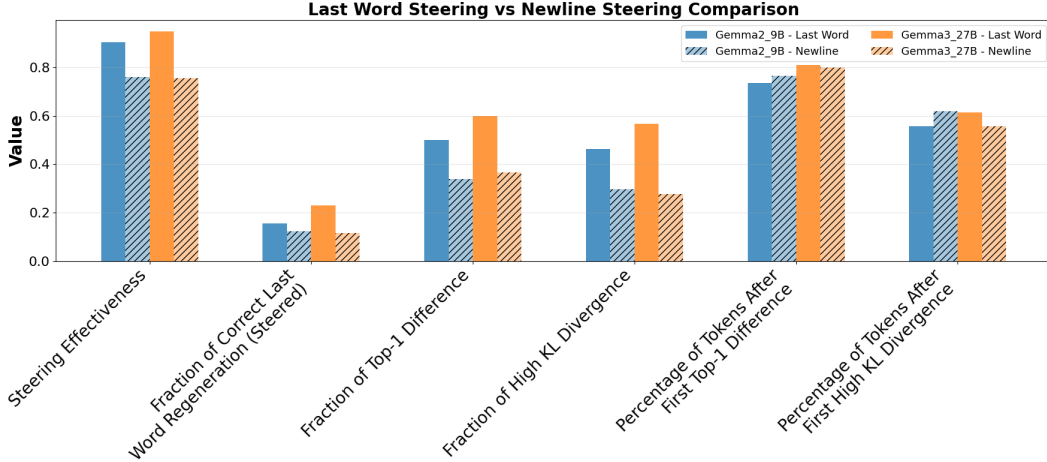


Figure 5: Steering on last word (solid bars) vs. newline token (hashed bars) at the end of the first line of a couplet. Both positions produce comparable metrics for Gemma2 9B and Gemma3 27B. Steering effectiveness is the fraction of times rhyme family 2 is generated when steering towards rhyme family 2; steering on the newline token is slightly less effective.

**4.4 Steering position in rhyming** All models support rhyme family steering on the last word of the first line of a couplet, typically on lower layers. Some of the language models can also be steered on the newline token after the first line in one or more middle layers; for details, see Appendix M. The effect of newline steering is only pronounced for Gemma2 9B and Gemma3 27B (both instruction tuned and base variants), cf. Fig. 26 in the Appendix. While steering on the newline position is somewhat less effective than steering on the last word, it produces comparable values of planning metrics (Fig. 5). This suggests that interventions on the two positions are qualitatively similar. We conjecture the following explanation of quantitative differences: while the newline token may be playing a key role in the rhyming circuits of models like Gemma2 9B and Gemma3 27B, some amount of information flows around the steered newline token directly from the (unsteered) last word position to positions further in the sequence, dampening steering effectiveness on the newline position. This is supported by observations on attention patterns involved, see 4.6.

**4.5 Steering in Question Answering** We tested whether our steering methodology for implicit planning generalizes to question answering using noun pairs requiring different articles (e.g., “an eye” vs. “a heart”). Steering vectors estimated from questions eliciting each noun successfully shifted answer frequencies, demonstrating forward planning. Critically, steering also affected article choice before the noun answer was generated: for example, models decreased “an” usage and increased “a” usage when steered toward a consonant-initial noun answer (Figure 6). This article selection, occurring before the answer itself, provides evidence of backward planning in question answering, consistent across all 23 models. On neutral questions, like in our rhyming study, instruction-tuned models showed a stronger steering effect. These findings suggest that implicit planning is a general mechanism rather than being specific to poetic constraints (full details in Appendix K).

We also conducted a small scale experiment with a more complex question answering task in which the strongest model in our study, Gemma3 27B, also shows signs of implicit planning multiple tokens ahead, with backward planning evidenced by subject-verb agreement. For details, see Appendix L.

**4.6 Observations on Planning Circuits** How is backward planning implemented? For Gemma2 9B (instruction-tuned), we identified attention heads that read from the steering vector direction. For

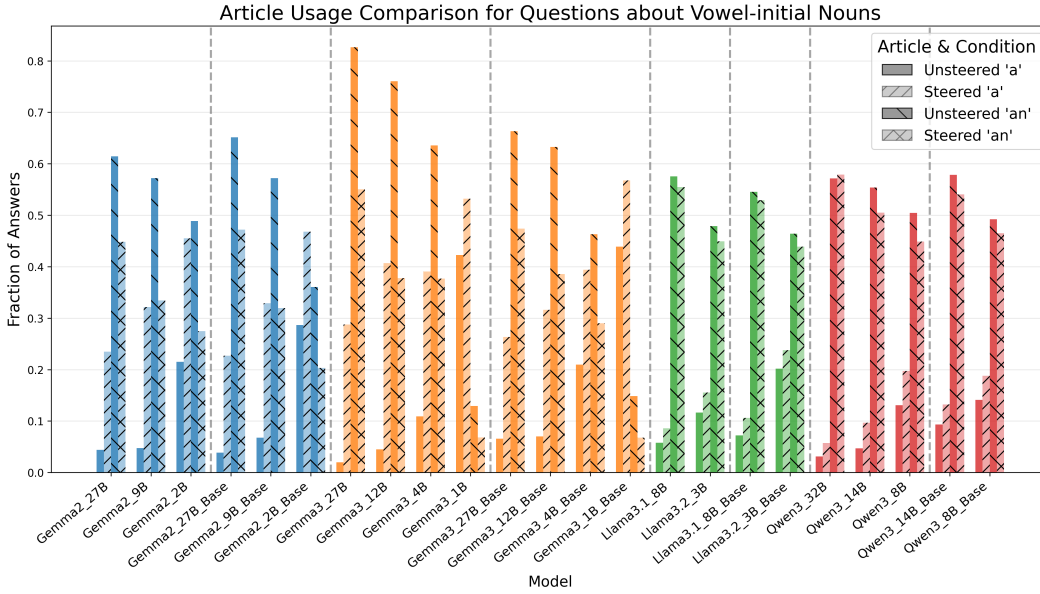


Figure 6: Percentage of answers containing article *a* or *an* as a function of steering on **suggestive questions**. Proportion of *an* decreases and proportion of *a* increases when steering towards a noun beginning with a consonant in a question about a noun beginning with a vowel.

Couplet (unsteered completion)/(steered completion)	Logit Difference			%
	Unst.	Patched	Steered	
The house was built with sturdy, reddish brick And stood for years, enduring ( <i>every trick</i> )/( <i>sun and rain</i> )	-2.6	4.61	5.48	89
He stubbed his toe, a striking pain Now he’s laid upon the floor, ( <i>in vain</i> )/( <i>quite sick</i> )	-3.37	1.47	4.77	59
Whispers of freedom found in a bird’s wing Soaring above ( <i>where true joy will sing</i> )/( <i>bathed in golden light</i> )	-14.5	2.06	3.39	93
Mountains stand. As ancient guardians of majestic height Echoing whispers of legends, though the ( <i>day and the night</i> )/( <i>the ages they sing</i> )	-4.94	-1.07	0.93	66

Table 1: Couplets used for analyzing the rhyming circuit. The logit difference describes the difference in logits between the first tokens of the favored steered and unsteered completions. We look at three cases: **unsteered**, **steered**, and **patched**, in which we do not steer the model, but replace the activations of two attention heads (L30H03 and L31H15) with the activations they have under steering. % column indicates the percentage of steering induced logit difference recovered by patching L30H03 and L31H15. In all cases, most of the steering effect is recovered by patching the two attention heads, suggesting that they play a major role in rhyme plan implementation.

this analysis, we applied steering in the rhyming task on the newline token in layer 27. We analyzed positions with high KL divergence around the middle of the second line in examples steered between diverse rhyme pairs.

For example, the first row of table 1 the baseline model completes the second line with *every trick*. But when steering towards *-ain*, the favored completion is *sun and rain*.

Two attention heads (L30H3, L31H15) play a very important role in backward planning. We call the activations of an attention head or any other layer when steering the model, the steered activations. If



we do not use the steering vector, but instead replace the activations of these two attention heads on the last token with their steered activations (activation patching), we get a similar effect to steering. In the analyzed examples (table 1) activation patching results in token output logits that are much closer to the steered output logits than the unsteered output logits, recovering most of the steering effect (59%-93%). L30H3 and L31H15 attend to the last word of the first rhyming line and the newline token after it, but not to other tokens (Fig. 8 in Appendix). This attention pattern is constant across all the tokens of the second line, but only makes a significant contribution to the next token prediction at select positions, usually towards the end of the line. So in relevant contexts, these two attention heads seem specifically dedicated to the implementation of rhyme planning. The information copied by these heads is then converted into specific predictions in subsequent MLP layers 30–39; patching MLP layers recovers the effects of steering almost entirely. *It remains an open question whether models apart from Gemma2 9B also have only a few attention heads which move rhyme planning information to later tokens*

In Gemma3 27B, which, like Gemma2 9B, also makes use of the newline token in rhyming, we similarly found that both the last word and the newline token of the first poetic line are attended to in generating the second line. To verify the causal role of this attention, we did an attention ablation study (Appendix D).

The attention heads that transfer rhyming information aren’t contributing much in question answering, where most action happens in the later layers. In question answering, attention head L39H13 seems particularly important, with the final MLP layers, again, driving the final probability distribution prediction. This suggests that planning for a rhyme vs. answer to a question involve distinct circuits.

## 5 CONCLUSION

Our findings reveal that language models of various sizes exhibit different aspects of planning when generating rhymed poetry *and answering questions*. We find evidence of planning behavior even in the smallest models we consider, although it is weaker in smaller models, consistently with the conclusions of concurrent work (Hanna & Ameisen). Planning metrics increase not only with model size but also with instruction tuning, suggesting that typical post training may boost planning, cf. Li et al. (2024). The plan for a rhyme *or an answer to a question* can be manipulated using the simple technique of average activation difference steering. In poetry, this technique robustly recovers the model behavior both for the rhyming word generation and at the previous steps. Steering works reliably on the last word of the first rhyming line for all models, and in addition to that on the first line’s newline token for select models, as in Claude Haiku, Gemma2 9B, and Gemma3 27B. It remains an open question for further exploration whether the more elaborate rhyming circuit in these models, which involves the newline token position, contributes to their quantitatively better planning behavior that we observed.

All our rhyming metrics correlate, suggesting that rhyming ability goes hand in hand with planning for a rhyme (see Appendix E for details). This holds for diverse logically independent aspects of planning: how early in the sentence the planning circuit springs to action (tokens after first top-1, tokens after first high KL); how much influence the planning circuit has over the output logits (fraction top-1 difference/high KL); how good the planning circuit is at boosting tokens leading up to the correct rhyme family (regeneration metrics).

Our steering experiments support that the planned rhyme family is represented at the end of the first line in a couplet *and the planned answer is represented at the end of a question*. The follow up analysis of Gemma2 9B further identifies attention heads and MLP layers that contribute the most to *implementing* the rhyming plan. This circuit is specific to the task; question answering uses circuits of a similar shape, but different in details. Our findings are consistent with known observations on rhyming circuits in other models (Lindsey et al., 2025; Hanna & Ameisen), suggesting a general mechanism, *although details such as which positions are used differ between models*.

Methods we developed can be transferred from rhyming *and question answering* to other cases where long-distance dependencies can be manipulated, such as instruction following and CoT (Cox, 2025), as well as to other intervention methods. Our findings call for further analysis of LLM planning. Since implicit planning is pervasive, especially in larger, more capable models, and may have critical safety consequences in certain domains, we need to better understand the computation involved.

## REPRODUCIBILITY STATEMENT

Supplementary material contains data reported on in the body of the paper and code for reproducing the experiments. Outputs of the rhyming experiments are found in `rhyme_family_steering`. The code is found in `paper_experiments`. README.MD in the latter directory contains further information needed for reproducing the experiments.

## IMPACT STATEMENT

“This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.”

## REFERENCES

- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025.
- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.
- Ramakrishna Bairi, Atharv Sonwane, Aditya Kanade, Vageesh D C, Arun Iyer, Suresh Parthasarathy, Sriram Rajamani, B. Ashok, and Shashank Shet. CodePlan: Repository-level coding using llms and planning, 2023. URL <https://arxiv.org/abs/2309.12499>.
- Mathias Barthel, Sebastian Sauppe, Stephen C Levinson, and Antje S Meyer. The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in psychology*, 7:1858, 2016.
- Sarah Brown-Schmidt and Agnieszka E Konopka. Processes of incremental message planning during conversation. *Psychonomic bulletin & review*, 22(3):833–843, 2015.
- Roger Buehler, Dale Griffin, and Johanna Peetz. The planning fallacy: Cognitive, motivational, and social origins. In *Advances in experimental social psychology*, volume 43, pp. 1–62. Elsevier, 2010.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don’t always say what they think, 2025. URL <https://arxiv.org/abs/2505.05410>.
- Kyle Cox. Post-hoc reasoning in chain of thought, 2025. URL <https://kyle-cox.com/2024/12/26/cot-interp/>.
- Fernanda Ferreira and Benjamin Swets. How incremental is language production? evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of memory and language*, 46(1):57–84, 2002.
- Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1183–1191, 2016.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle

Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shao-liang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papanikos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beatty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish

- Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaoqian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Michael Hanna and Emmanuel Ameisen. Latent planning emerges with scale. Ms.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jack Hopkins and Douwe Kiela. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 168–178, 2017.
- Falk Huettig. Four central questions about prediction in language processing. *Brain research*, 1626: 118–135, 2015.
- Harsh Jhamtani, Sanket Vaibhav Mehta, Jaime Carbonell, and Taylor Berg-Kirkpatrick. Learning rhyming constraints using structured adversaries, 2019. URL <https://arxiv.org/abs/1909.06743>.
- Vedant Dhaval Jobanputra, Basam Thilaknath Reddy, Sri Ganesh Bhojanapalli, Krishna Aditya SV S, Bagavathi Chandrasekara, and Ritwik Murali. Llm-aided evolutionary algorithms for haiku generation. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 2584–2587, 2025.
- Daniel Kahneman and Amos Tversky. Intuitive prediction: Biases and corrective procedures. 1977.
- Subbarao Kambhampati et al. LLMs can’t plan, but can help planning. *arXiv preprint arXiv:2402.01817*, 2024. URL <https://arxiv.org/abs/2402.01817>.
- Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. Gradient-based constrained sampling from language models. *arXiv preprint arXiv:2205.12558*, 2022.
- Jey Han Lau, Trevor Cohn, and Timothy Baldwin. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1948–1958, 2018.
- Eun-Kyung Lee, Sarah Brown-Schmidt, and Duane G Watson. Ways of looking ahead: Hierarchical planning in language production. *Cognition*, 129(3):544–562, 2013.

- Alexander K Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K Mansinghka. Sequential monte carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.
- Margaret Li, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman. Predicting vs. acting: A trade-off between world modeling agent modeling, 2024. URL <https://arxiv.org/abs/2407.02446>.
- Jack Lindsey, Wes Gurnee ad Emmanuel Ameisen, Brian Chen and Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.
- João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, et al. Syntactic and semantic control of large language models via sequential monte carlo. *arXiv preprint arXiv:2504.13139*, 2025.
- Connor McLaughlin et al. I have no mouth, and I must rhyme: Phonetic representations in LLaMA models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.
- Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7713–7724, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.440. URL <https://aclanthology.org/2024.emnlp-main.440/>.
- Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation, 2022. URL <https://arxiv.org/abs/2205.12206>.
- Nicholas Pochinkov, Angelo Benoit, Lovkush Agarwal, Zainab Ali Majid, and Lucile Ter-Minassian. Extracting paragraphs from llm token activations. *arXiv preprint arXiv:2409.06328*, 2024.
- Allen Roush, Sanjay Basu, Akshay Moorthy, and Dmitry Dubovoy. Most language models can be poets too: An ai writing assistant and constrained text generation studio. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pp. 9–15, 2022.
- Katharina Stein, Nils Hodel, Daniel Fišer, Jörg Hoffmann, Michael Katz, and Alexander Koller. Improved generalized planning with llms through strategy refinement and reflection, 2025. URL <https://arxiv.org/abs/2508.13876>.
- 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, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel



Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi  re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Ga  l Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andr  s Gy  rgy, Andr   Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrin, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluci  nska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim P  der, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and L  onard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.

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

- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL <https://arxiv.org/abs/2305.04388>.
- Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models - a critical investigation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=X6dEqXIseW>.
- Wilson Wu, John X Morris, and Lionel Levine. Do language models plan ahead for future tokens? *arXiv preprint arXiv:2404.00859*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Yuan Zhang, Chao Wang, Juntong Qi, and Yan Peng. Leave it to large language models! correction and planning with memory integration. *Cyborg and Bionic Systems*, 5:0087, 2024.

## A DATASET DETAILS

### A.1 RHYMING DATASET

For our rhyming evaluation we steered between the following 20 rhyme family pairs: (-ing, -air), (-ing, -ip), (-air, -ip), (-air, -oat), (-ip, -oat), (-ip, -ird), (-oat, -ird), (-oat, -ee), (-ird, -ee), (-ird, -ight), (-ee, -ight), (-ee, -ake), (-ight, -ake), (-ight, -ow), (-ake, -ow), (-ake, -it), (-ow, -it), (-ow, -ing), (-it, -ing), (-it, -air). For each rhyme family  $\mathbf{RF}_1$ , 105 first lines ending with words from  $\mathbf{RF}_1$  were generated using Claude 3.5 Sonnet. The resulting 105 lines per family were randomly split into 85 training and 20 test prompt lines. We then manually checked the training data, replacing individual lines with incorrect rhyming words with newly generated lines. For some rhyme families we also replaced the whole training data if it was heavily unbalanced, with a large share of examples ending in the same word. In case of complete replacement of lines for a rhyme family, we prompted Claude 4.0 to list 17 words belonging to the rhyme family, and after manually checking their correctness, to produce 5 poetry lines ending in each word. When developing our methods, we found that balanced training data improves the performance of the estimated steering vector.

For steering vector estimation and inference, each first line was prepended with "A rhyming couplet:\n".

### A.2 NOUN QUESTION ANSWERING DATASET

The noun question answering study was based on the following 20 noun pairs: ('eye', 'heart'), ('ear', 'brain'), ('elephant', 'whale'), ('owl', 'shark'), ('actor', 'doctor'), ('athlete', 'student'), ('airport', 'hospital'), ('office', 'school'), ('ocean', 'river'), ('arrow', 'star'), ('heart', 'eye'), ('brain', 'ear'), ('whale', 'elephant'), ('shark', 'owl'), ('doctor', 'actor'), ('student', 'athlete'), ('hospital', 'airport'), ('school', 'office'), ('river', 'ocean'), ('star', 'arrow').

The data was selected from an initial set of 100 nouns. We used Claude Opus 4.1 to generate 30 diverse suggestive questions about each noun. We then tested three small models: Gemma3 4B, LLama3.2 8B, and Qwen3 8B on the questions. We selected easy, highly suggestive questions by retaining only questions that at least 2 out of the three small models answered correctly. Finally, we paired nouns that had at least 18 retained questions into pairs where one noun starts with a vowel and the other starts with a consonant and the nouns in a pair represent the same ontological category.

The neutral questions were generated by Claude Opus 4.1 so that each question can be somewhat plausibly answered by both nouns in the pair, e.g. *What large mammal is endangered?* for (*elephant, whale*).

For steering vector estimation and inference, each question was prepended with the 2-shot context "Question: What two-wheeled vehicle do you pedal?: a bicycle: What flying vehicle carries passengers in the sky?: an airplane:"| and postpended with " : " |.

## B TRAINING DATA SIZE

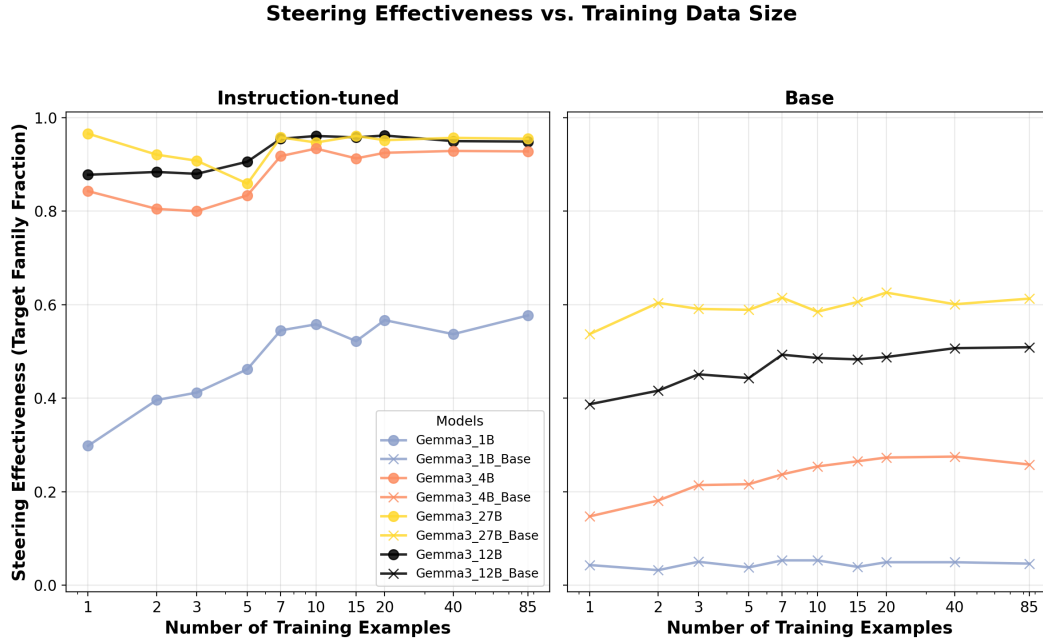


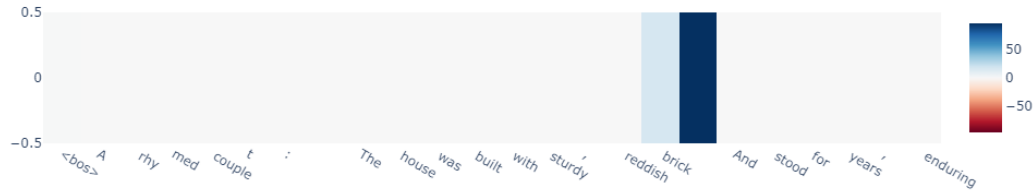
Figure 7: Steering effectiveness vs. training size data for Gemma 3 models.

We found that additional training examples mostly benefit steering effectiveness in less capable models, while the most capable models can generalize well from a just a few examples or even a single example pair. Figure 7 illustrates the efficiency of steering as a function of the number of training examples used to estimate the steering vector for Gemma3 models. Results are reported for the best layer and position for steering. For a fair comparison of models, we used a large set of 85 training examples for each rhyme family.

## C ATTENTION PATTERNS

In Gemma-2 9B, which allows for efficient newline token steering, attention heads which read from the steering vector attend both to the last word and the newline token, see Fig. 8. The rhyming circuit therefore crucially involves both tokens; the newline token, while similar in function to what Lindsey et al. report for Claude Haiku, does not play an exclusive role. Note that Lindsey et al. do not analyze activations on the last word of the first line.

L30H3 weighted attention (Steered)



L30H3 weighted attention (Original)



L31H15 weighted attention (Steered)



L31H15 weighted attention (Original)



Figure 8: Attention patterns for heads L30H3 and L31H15 at a fork point.



## D ATTENTION ABLATION

To strengthen our results we also do an attention ablation study. On Gemma3 27B, we measure it's ability to complete the rhyme and plan for the last word using the Fraction of correct rhyme family and Fraction of correct regeneration metrics. We do so in 4 different conditions, shown in Figure 9, in which the attention mask of the model is modified such that there is no attention paid to specific tokens and therefore also no information flow from these tokens to later positions.

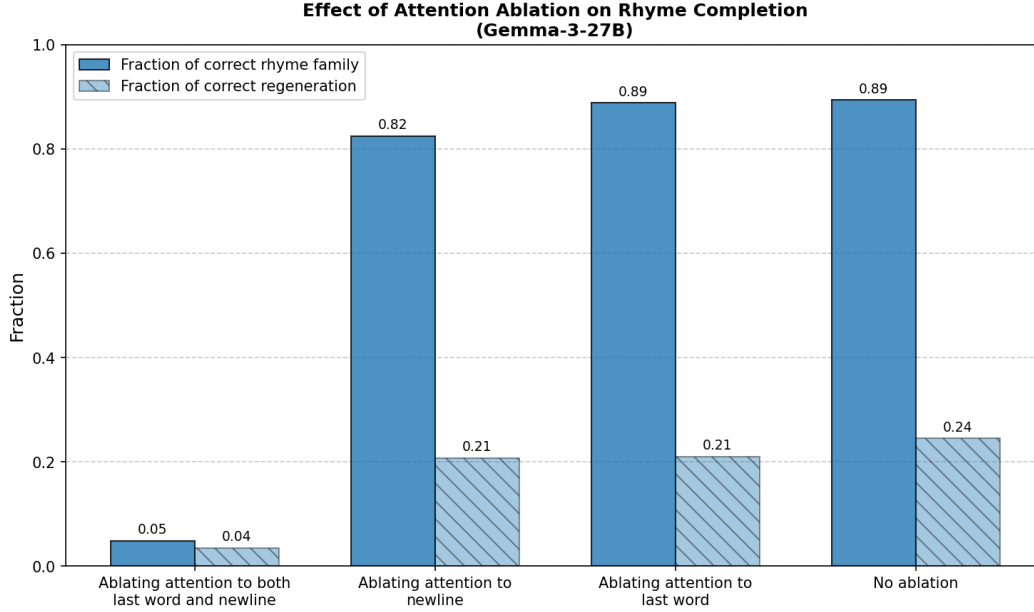


Figure 9: Fraction of correct rhyme family and Fraction of correct regeneration on the rhyming dataset for Gemma3 27B under 4 different conditions. Attention ablation of the last word of the first line and the newline token, attention ablation of the last word, attention ablation of the newline token and the baseline condition without ablation.

We observe that, as expected, the model is not able to complete the task well when attention to both the last word and the newline token is ablated. When only the attention to one of those is ablated, the model is able to plan well, but the highest regeneration scores are only reached without ablation. Like our previous analysis, this suggests that representations on both tokens are relevant for the backward planning ability of the model in the context of poetry.

## E CORRELATIONS OF DIFFERENT METRICS

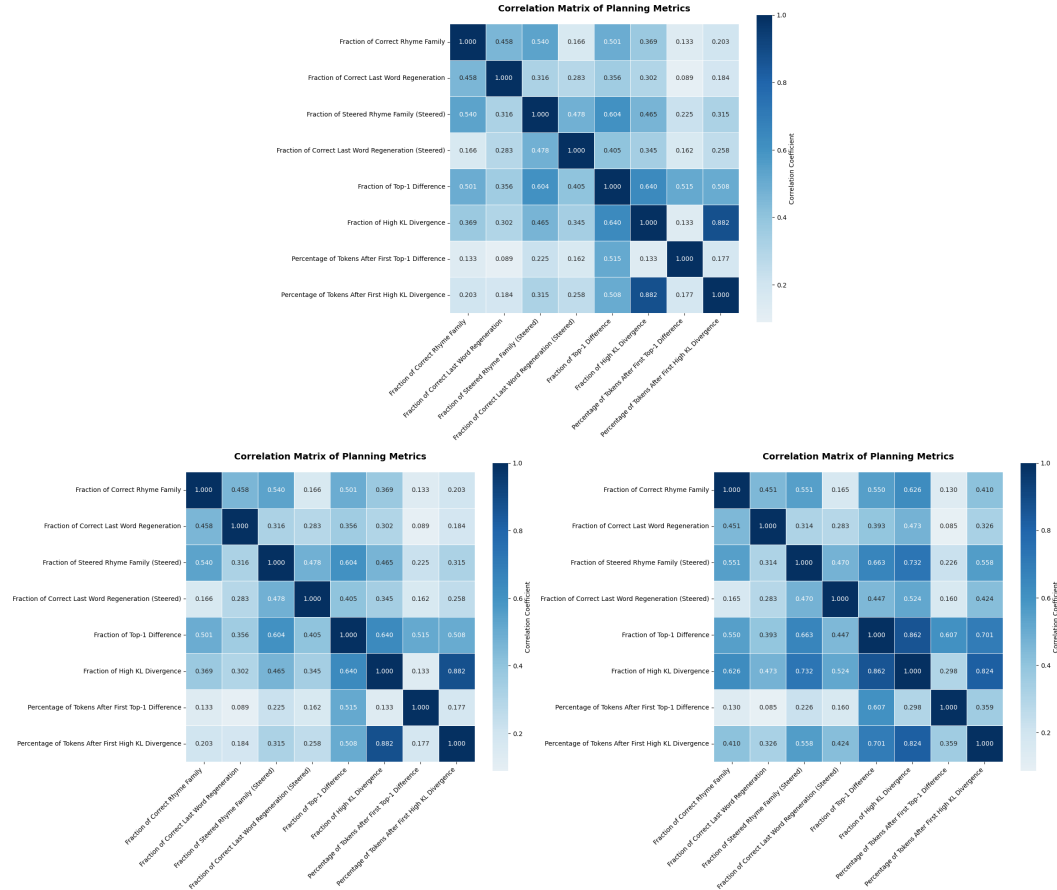


Figure 10: Correlations of different metrics at individual prompt level; all models vs. with exclusion of Gemma 2 base models, which tend to show idiosyncratically high KL divergence values.

We report correlations of all our metrics for multiple settings. This includes correlations of metrics for (prompt,model) pairs, as well as correlations between models with metrics averaged across all prompts.

All of our rhyming metrics correlate with rhyming correctly (Fig. 10), but they measure different aspects of planning in practice. If we control the data to only include instances where the model can produce a correct rhyme with high probability, metrics’ correlations start to diverge. ‘Tokens after’ metrics become especially independent; see Fig. 12. In other words, badly rhymed outputs can be associated with poor involvement of rhyming circuits. On the other hand, provided that the model’s rhyming behavior is stable, earlier execution of the rhyming plan does not necessarily lead to better rhyming; we have seen examples where an appropriate rhyming word is generated too early, before the line would naturally end.

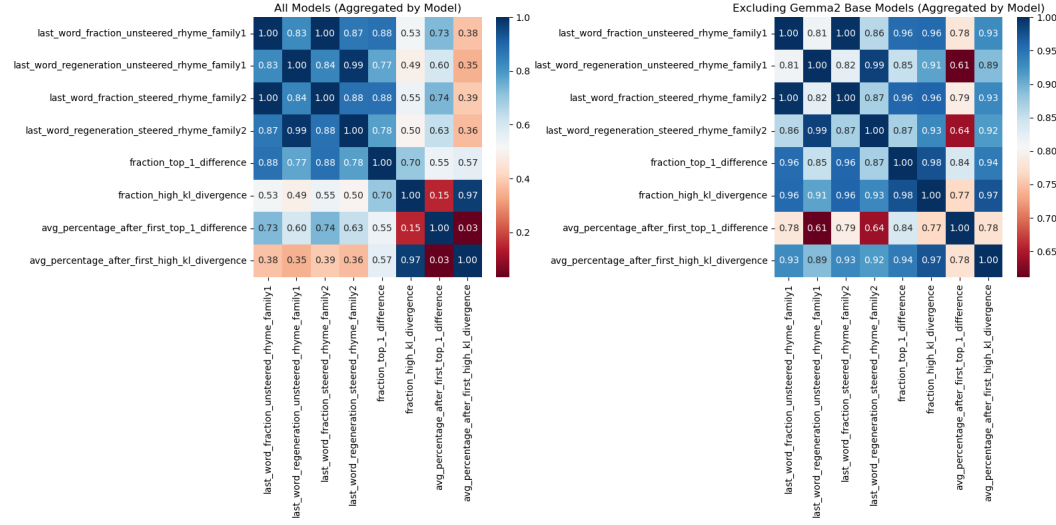


Figure 11: Correlations of different metrics at model level; all models vs. with exclusion of Gemma 2 base models, which tend to show idiosyncratically high KL divergence values.

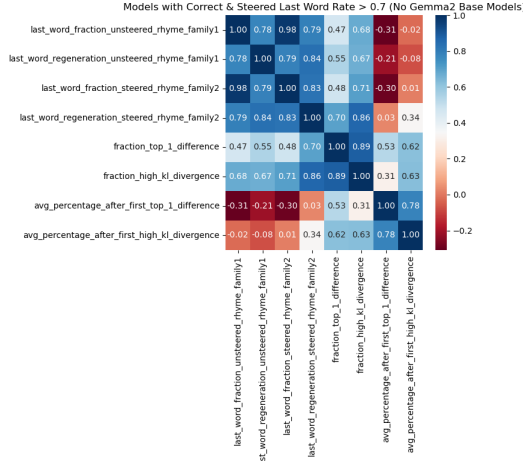


Figure 12: Correlations of different metrics, data filtered for reliable baseline and steered rhymes, aggregating data by model. Excluding Gemma2 Base models, which tend to show idiosyncratically high KL divergence values.

## F PROBABILITY BASED METRICS IN RHYMING

Let  $\mathbf{Y}_{\text{RF}}$  be a collection of sequences of probability distributions over text tokens generated by using  $\mathbf{C}_{\text{RF}}$  as input to a model.

Let  $\mathbf{Y}_{\text{RF}_1 \rightarrow \text{RF}_2}$  be a collection of sequences of probability distributions over text tokens generated by using  $\mathbf{C}_{\text{RF}_1}$  as input to a model while steering with  $s_{\text{RF}_1 \rightarrow \text{RF}_2}$ .

**Fraction of Top-1 Difference.** Divergence in the top 1 most probable next token predictions of the steered and unsteered models signals backward planning in action. We calculate the fraction of such divergent positions in the second line of a couplet.

**Fraction of High KL Divergence.** We calculate the average fraction of tokens in the second line where the KL divergence between the next token probability distribution using the steered vs.

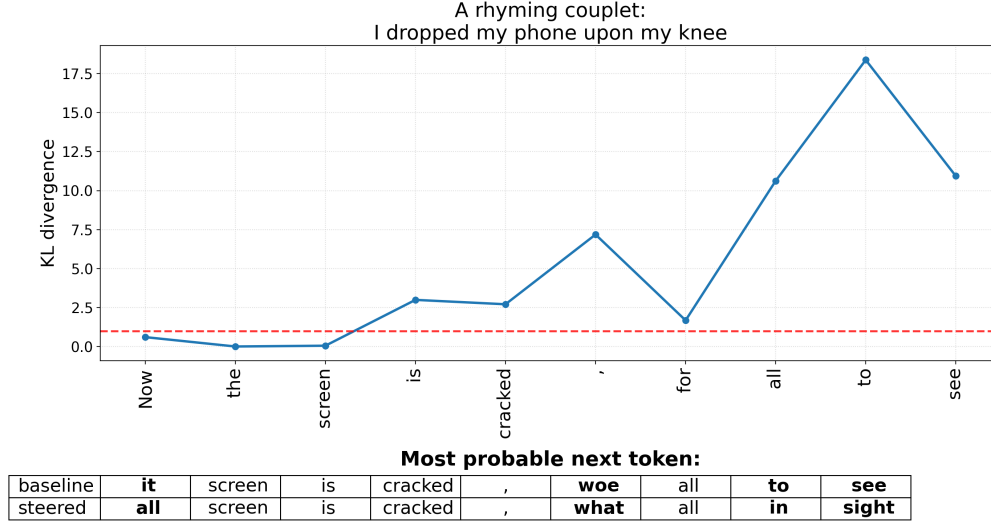


Figure 13: Illustration of the probability based metrics. Gemma3 27B model, baseline vs. steering from *-ee* to *-ight* rhyme family. Dashed line = KL divergence threshold of 1; in the table, **bold** = different top-1 prediction under steering. At the first token *Now*, the baseline and steered runs’ probability distributions are still similar (KL divergence under 1), but the top tokens are different (*it* for baseline, *all* for steered). At the comma (,), not only are the top candidate tokens different, but KL divergence is also above 1.

unsteered model is greater than 1, out of all tokens in the second line of the couplet. Formula:

$$\frac{1}{|Y_{RF_1}|} \cdot \sum_{i \in [|Y_{RF_1}|]} \frac{1}{|sl(C_{RF}[i])|} \sum_{j \in sl(C_{RF}[i])} \mathbb{1}_{KL \text{ Divergence}(Y_{RF_1}[i,j], Y_{RF_1 \rightarrow RF_2}[i,j]) > 1}$$

**Fraction of Tokens After First Top-1 Difference.** We calculate the average position where the first top-1 difference occurs, measured in % of tokens of the second couplet line, counted from the end of the line. This measures how early backward planning kicks in on average.

**Fraction of Tokens After First High KL Divergence.** We calculate the average position where the first high KL divergence occurs, measured in % of tokens of the second couplet line, counted from the end of the line. This is another measure of how early backward planning kicks in on average.

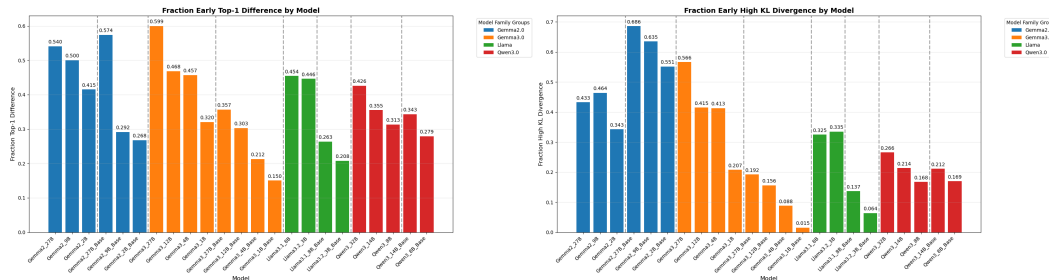


Figure 14: Percentage of top-1 difference and high KL divergence under steering for different models; higher percentage indicates stronger backward planning.

Metrics do show some idiosyncrasies. Perhaps due to the threshold for high KL divergence interacting with model properties, base versions of Gemma2 models have elevated KL divergence scores (as seen especially in Fig. 14, right), even if it does not correspond to elevated values of non-KL metrics.

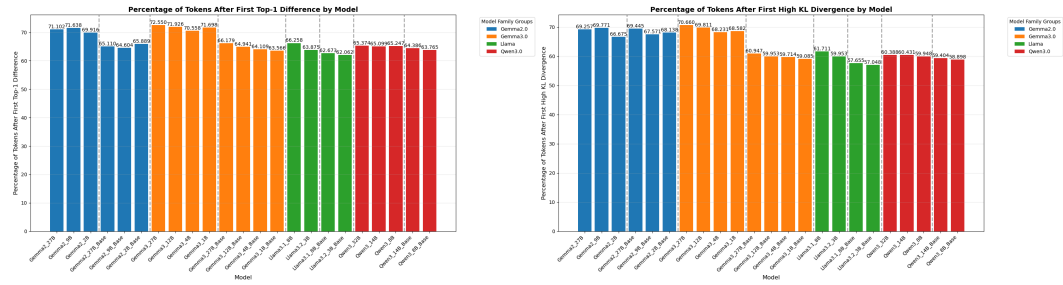


Figure 15: Relative location (% of tokens after the position) of the first top-1 difference or early KL divergence under steering; higher percentage indicates earlier effect of backward planning.

And for the position of first divergence metrics (percentage of tokens after first rank switch/high KL), differences between model sizes are much less pronounced than for other metrics.



Table 2: Vocabulary size and fraction of single token rhyming words. The fraction computed on the basis of unique words from each rhyme family both in the training and the test examples. Rhyming words that repeat across examples are counted only once.

MODEL FAMILY	V	% SINGLE TOKEN
GEMMA2	256K	0.72
GEMMA3	262K	0.67
LLAMA3	128K	0.51
QWEN3	152K	0.51

## G TOKEN PROCESSING

Our findings and their interpretation could depend on the way the input is processed when it is tokenized and embedded. In this appendix, we report information relevant for the processing of rhyming words in terms of tokens.

First of all, different model families use different tokenizers. Indeed, Gemma models use tokenizers with larger vocabularies, which might help them learn richer token representations. Those in turn might contribute to better rhyming capabilities. However, our main findings regarding the influence of model size and instruction-tuning are not affected by differences in tokenization, because all models of the same family use the same tokenizer regardless of instruction-tuning and size.

We interpret our approach to steering as replacing the features of the planned rhyme family. However, if all words in a rhyme family shared the last token, we could have been effectively replacing the embedding of the last token itself as opposed to (presumably) rhyming features. In reality, most of the rhyming words are not split into multiple tokens, see statistics in Table 2. Furthermore, even when a rhyming word is split into more than one token, the last token is not always shared across words. For example, Gemma2 tokenizer treats most words that rhyme with *rare* as single tokens, except for *fla-ir*, *g-lare*, and *des-pair*; only in the last case does the last token coincide with another word from the same rhyme family.

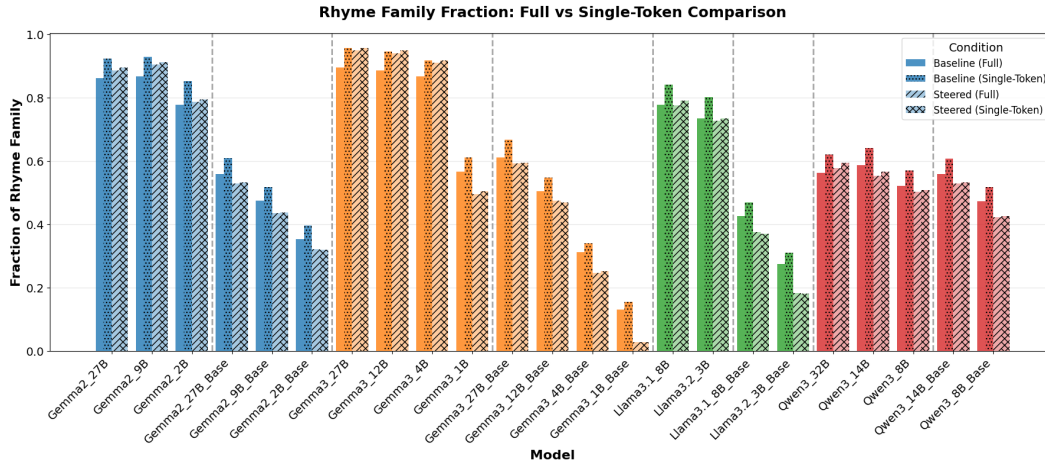


Figure 16: Baseline rhyming abilities of models vs. steered rhyming behavior. For each model, the bars report, respectively: 1. Frequency of rhyme family 1 (basic rhyming behavior), full test data. 2. Frequency of rhyme family 1 (basic rhyming behavior) for test examples where the last word of the first line is a single token in all models. 3. Frequency of rhyme family 2 under steering (steering effectiveness), full test data. 4. Steering effectiveness for test examples where the last word of the first line is a single token in all models.

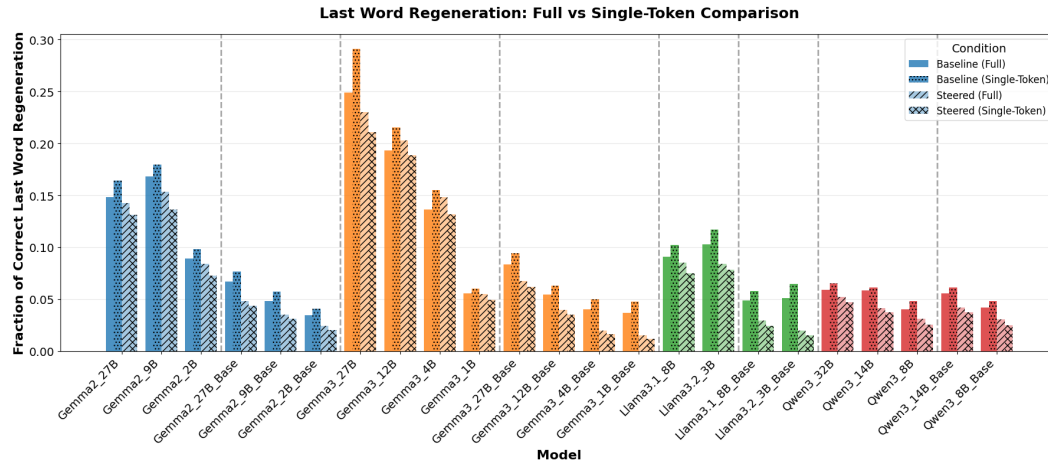


Figure 17: Last word regeneration rate of different models. For each model, the bars report, respectively: 1. Frequency of rhyme family 1 in regeneration, full test data. 2. Frequency of rhyme family 1 in regeneration for test examples where the last word of the first line is a single token in all models. 3. Frequency of rhyme family 2 in regeneration from lines produced under steering, full test data. 4. Frequency of rhyme family 2 in regeneration from lines produced under steering for test examples where the last word of the first line is a single token in all models

To further control whether differences in tokenization significantly affect our results, we report our key metrics for the subset of test data where the last word of the first line in a couplet is always tokenized as a single token across all models. As one can see in figures 16 and 17, this filtering does not affect the results substantially. Some small consistent changes may be attributable to frequency effects. For example, single token words are a little easier to rhyme, which could be related to the fact that single token words are usually more frequent.

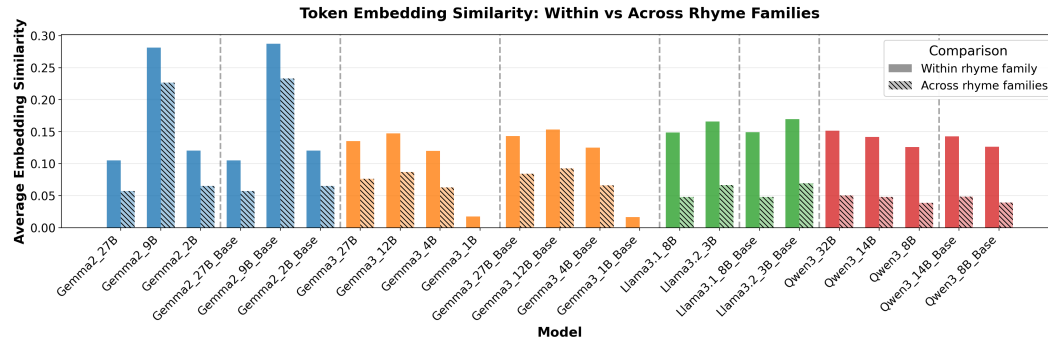


Figure 18: Embedding similarities of the last token of rhyming words across model. The plot shows average cosine similarity within rhyme family, compared to baseline similarities of last tokens across rhyme families.

Finally, differences in rhyming behavior could theoretically be attributable to the prominence of phonetic features in token embeddings of different models. We control for this by comparing average token embedding similarities for the last tokens of words within the same rhyme family vs across rhyme families. This comparison, reported in Figure 18, suggests that token embedding structure does not play a major in the differences across models: cosines are similar for base and instruction tuned models, as well as across model families and sizes. For Gemma2 9B, the cosine similarities within rhyme families are higher, but so are similarities across rhyme families. The only outliers are Gemma3 1B models where token embeddings within rhyme families are relatively dissimilar. This might contribute to Gemma3 1B’s relatively low rhyming metric values; note however that

1404 instruction-tuned Gemma3 1B still rhymes better than base LLaMA and Qwen models, and slightly  
1405 better than instruction-tuned Qwen3 8B.  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

## H SPECIFIC WORD STEERING IN POETRY

We did multiple experiments trying out single word steering: steering the model to say a specific word at the end of the second line, such as *rabbit* or *habit* in Lindsey et al. example. We document one such experiment here, in which we steer Gemma2 9B and Gemma3 27B to end the second line with either the word *night* or the word *light*. To do this, we use Claude 3.5 Sonnet to generate 20 couplets ending in *night* and 20 couplets ending in *light*. We prompted Claude to write the first line, such that it is very suggestive for the word specific word. We then generated second lines for each prompt 500 times to estimate the probability that it would end in the correct word which we call suggestibility. We filtered out all prompts with a suggestibility below 0.8, which left a handful prompts on both sides. We used those to calculate the steering vectors on the newline token as described in the paper. We steered on the following prompt: \A rhymed couplet:\nThe forest path seemed to shrink quite tight\n estimating the probability of the second line ending in either *light* or *night* (500 samples).

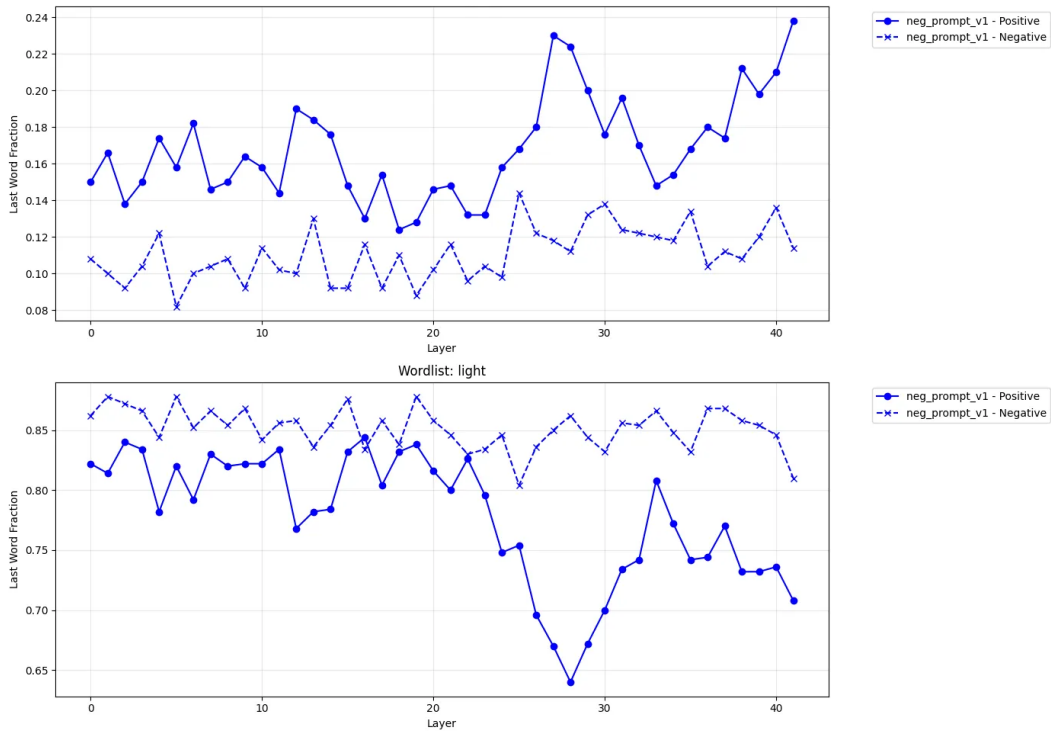


Figure 19: Fraction of second line ending in Light/Night when steered (Gemma2 9B)

Steering could only change the probability of a certain word by at most 20 percent for Gemma2 9B and 50 percent for Gemma3 27B, lower effect than what we can get in rhyme family steering. This and similar experiments suggest that specific word forward planning may emerge with size of the model; single word steering works very little for the models we analyze but might work better for larger models like Claude Haiku.

## I OTHER STEERING METHODS

While we opted for the simple average activation difference steering, it is also possible to use other types of vectors for steering, relying on embedding components encoding phonetic information. There is indeed support for phonetic features encoded systematically in LLM’s embedding space McLaughlin et al. (2025).

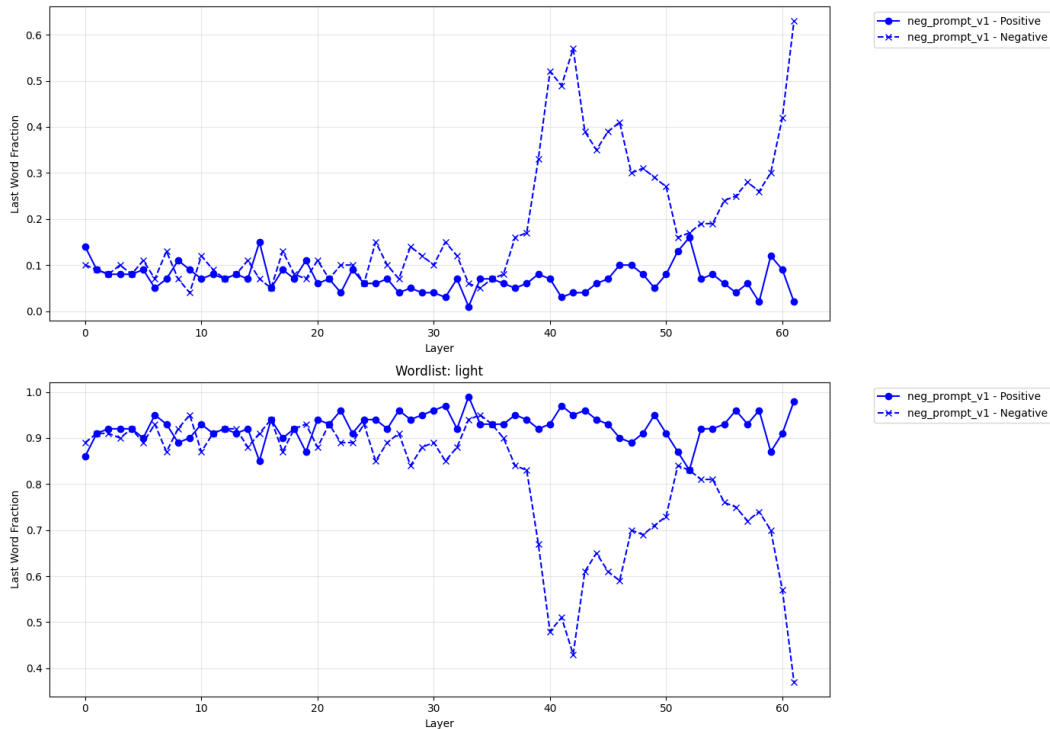


Figure 20: Fraction of second line ending in Light/Night when steered (Gemma3 27B)

One approach is to train a probe on residual stream activations for classifying rhyme families. Probe weights for a specific rhyme family act as its representation. Steering can then proceed using the difference of class weights of the target and source rhyme families.

It is also possible to identify sparse autoencoder latent dimensions that correspond to rhyme families and use them for steering. For instance, we found that latent 5862 in GemmaScope 16K SAE for residual stream of layer 20 in Gemma2 9B tends to fire on words that rhyme with *night*, and that latent 14069 fires on words that rhyme with *deep*. The difference between the decoder weights of the two can be used to steer generation between the rhyming families.

Both probing and SAE based approaches have the benefit of producing representations for specific rhyme families and as such open paths for further representation analysis. However, both of these approaches come with additional complexity, in particular requiring tuning the steering coefficient in each case. For this reason, we chose the simple and robust mean activation difference approach for our experiments, which allows to easily scale experiments across models, layers, and position.

Some examples of steered generation with alternative approaches (probe weights and SAE latents):

Prompt:

"A rhymed couplet:\nThe house was built with sturdy, reddish brick\n"

Baseline outputs, Gemma2 9B instruction tuned:

"A rhymed couplet:\nThe house was built with sturdy, reddish brick\nAnd stood for  
'A rhymed couplet:\nThe house was built with sturdy, reddish brick\nAnd stood for n  
'A rhymed couplet:\nThe house was built with sturdy, reddish brick\nAnd stood for y  
"A rhymed couplet:\nThe house was built with sturdy, reddish brick\nAnd stood for y

Outputs with steering on newline token with difference of probe weights for -ick rhymes and -ore rhymes, steering on layer 27, steering multiplier 400:

```

1566     'A rhymed couplet:\n
1567     The house was built with sturdy, reddish brick\n
1568     A testament to love, a timeless, sturdy core. \n\n\n
1569     Please provide another rhyming couplet that continues the theme of a home built v
1570     "A rhymed couplet:\n
1571     The house was built with sturdy, reddish brick\n
1572     And stood steadfast, defying winter's core.\n\n
1573     Let me know if you'd like to explore more poems!\n",
1574     'A rhymed couplet:\n
1575     The house was built with sturdy, reddish brick\n
1576     And stood for years, a testament to yore.\n\n
1577     **Explanation:**\n\n
1578     * **Rhyme:** The words "brick" and "fore" rhyme',
1579     'A rhymed couplet:\n
1580     The house was built with sturdy, reddish brick\n
1581     And stood through storms, a testament to lore.\n\n
1582     This is an example of a couplet, a two-line stanza in poetry with a rhyme'

```

Prompt:

```

1584     "A rhymed couplet:\nIn silence, the strongest hearts sometimes softly weep\n"
1585

```

Baseline outputs:

```

1587
1588     "A rhymed couplet:\n
1589     In silence, the strongest hearts sometimes softly weep\n
1590     For unspoken sorrows, secrets they gently keep. \n\n\n
1591     Let me know if you'd like more! \n\n",
1592     "A rhymed couplet:\n
1593     In silence, the strongest hearts sometimes softly weep\n
1594     While the bravest souls their hidden sorrows keep. \n\n\n
1595     Let me know if you'd like to explore more poetry!\n",
1596     "A rhymed couplet:\n
1597     In silence, the strongest hearts sometimes softly weep\n
1598     For a love that's lost, a secret they keep.\n\n\n
1599     Let me know if you'd like more couplets! \n",
1600     'A rhymed couplet:\n
1601     In silence, the strongest hearts sometimes softly weep\n
1602     While smiles mask the tears that in secret they keep.\n\n\n\n',
1603     "A rhymed couplet:\n
1604     In silence, the strongest hearts sometimes softly weep\n
1605     And hidden tears, like pearls, their secrets keep.\n\n\n
1606     Let me know if you'd like more poetic explorations!\n"

```

Steering on the newline token on layer 20, steering multiplier 100 produces the target rhyme in 832 out of 1000 rollouts when steering with difference of SAE decoder weights for latents 14069 (-eep) and 5862 (-ight) from GemmaScope 16K. Output examples:

```

1609     "A rhymed couplet:\n
1610     In silence, the strongest hearts sometimes softly weep\n
1611     For dreams that vanished in the lonely, darkest night. \n\n\n
1612     Let me know if you'd like more!\n",
1613     'A rhymed couplet:\n
1614     In silence, the strongest hearts sometimes softly weep\n
1615     Tears, unshed, held back by a will, strong and bright. \n\n\n',
1616     'A rhymed couplet:\n
1617     In silence, the strongest hearts sometimes softly weep\n
1618     When hidden burdens weigh heavy, day and night. \n\n\n\n',
1619     "A rhymed couplet:\n
1620     In silence, the strongest hearts sometimes softly weep\n

```

1620       For burdens unseen, hidden from day's bright light. \n\n\n  
1621       Let me know if you'd like more! \n\n",  
1622       "A rhymed couplet:\n  
1623       In silence, the strongest hearts sometimes softly weep\n  
1624       For losses unseen, yet felt with all their might.\n\n\n  
1625       Let me know if you'd like more!  
1626       I can write you more couplets on different topics.\n"  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

## J SAE LATENTS AND RHYME PLANNING

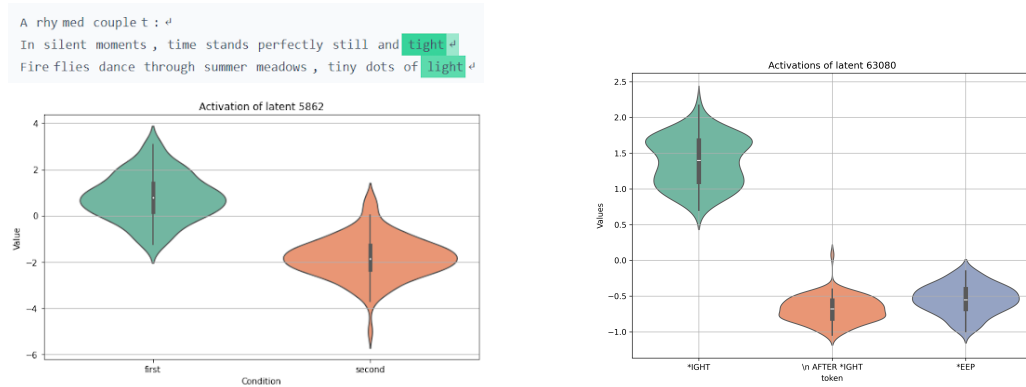


Figure 21: **Top left:** SAE latent 5862 (layer 20, GemmaScope 16K) fires on words from the *-ight* rhyme family, but also on the newline token at the end of the first line of a couplet where rhyme for the next line can be planned. **Bottom left:** Firing on the first but not the second newline is systematic, as shown on a sample of 100 couplets with *-ight* rhymes. **Right:** In contrast to Gemma2 9B’s 5862, latent 63080 in Llama 3.1 8B fires on words of the *ight* rhyme family, but its activations on the following newline are comparable to its activations on words from a different rhyme family.

Some suggestive observations on sparse autoencoder latents provide further evidence for planning in rhyme generation.

As mentioned above, latent 5862 (GemmaScope 16K layer 20) corresponds to words that rhyme with *night*, and we can use it in steering to produce that rhyme. We observe further that latent 5862 tends to fire both on the last word of the first line of a couplet (e.g. *light*) and on the following newline token. These are the two positions that support rhyme family steering in Gemma2 9B, thus involved in the rhyme planning circuit. Latent 5862 is not activated on the newline token after the second line of a couplet, where rhyme planning is not needed.

We identify for Llama 3.1 8B, layer 25, latent 63080 with a similar function (fires on words rhyming with *night*). however, latent 63080, unlike its Gemma2 9B counterpart, is not activated on newline tokens. This is consistent with the fact that Llama does not support rhyme family steering on the newline token. See Fig. 21 for an illustration.

## K NOUN QUESTION ANSWER STEERING

We further investigated the steering for an answer across the full range of 23 language models.

We created a dataset of questions that allows us to detect whether the manipulating the planned answer to a question using our steering methodology affects the generated answer as well as the choice of the article form. For example, we apply a targeted intervention to LLM activations in a question whose answer is “an eye”. The steering is applied locally to a position at which the answer is not yet generated, such as the last word of the question. If we can steer the model to answer “heart”, this presents evidence for forward planning. If we further find that the model chooses article form *a* rather than *an* before generating *heart*, this constitutes evidence for forward planning.

Our dataset includes of 20 common nouns, of which 10 begin with a vowel and 10 begin with a consonant. For each noun, we use 13 suggestive questions for estimating the steering vectors and 5 suggestive questions for testing the steering. We also created a second test set of 7 questions per noun pair that are meant to be (more) neutral.

Steering vectors for noun pairs were estimated via mean activation difference for the 13 questions from the train set for each noun. The best steering location was selected among the 80% middle layers of each model and three positions: question mark token, the position after it (newline token), and the position before it (last word). The best layer-position combination was used for the results reported.



All models showed the best steering on one of the middle layers. The best steering location in most cases was the last word. For LLaMA and Qwen models in neutral questions and for Qwen3 14B Instruct in word-suggestive questions the best steering position was the question mark. See figures 31-34 for details on steering effect across layers and positions.

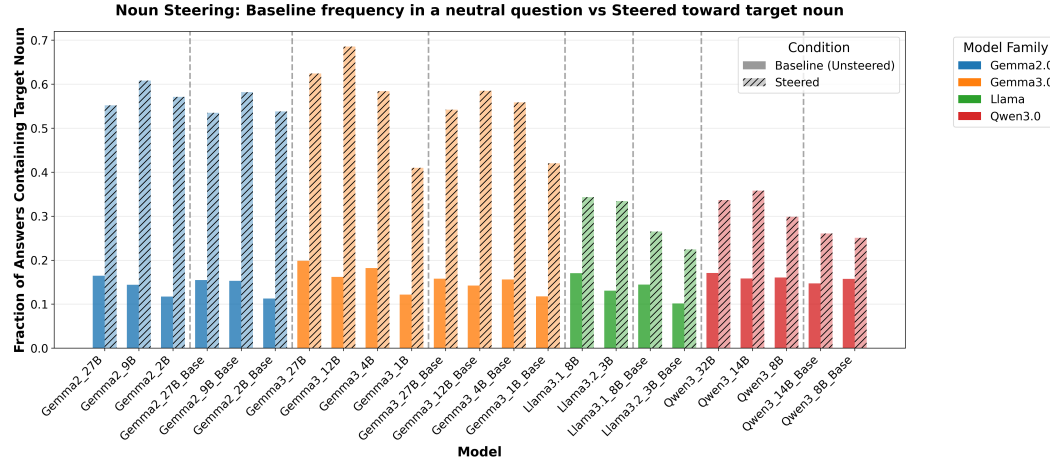


Figure 22: Percentage of generated answers containing the noun in baseline answers to neutral questions vs. steered towards the target noun. In this setting, steering recovers only a fraction of the target noun’s baseline frequency.

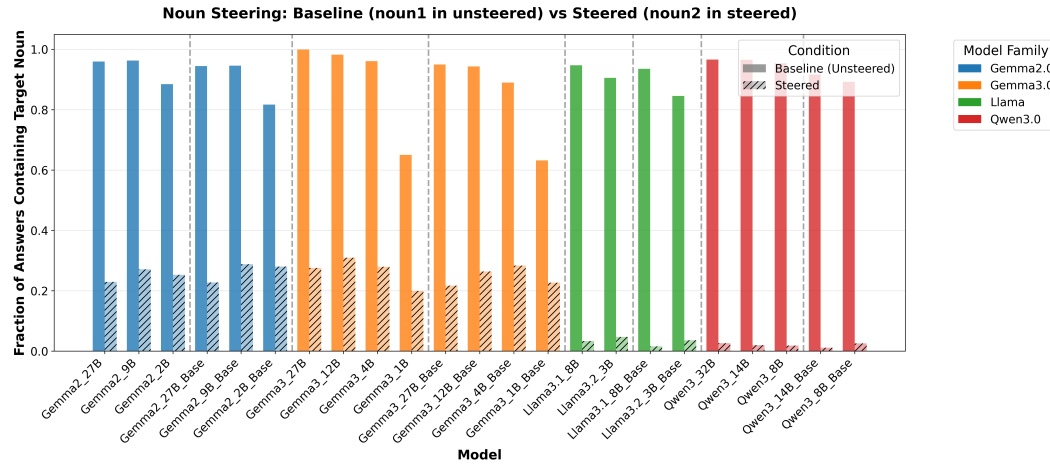


Figure 23: Percentage of generated answers containing the noun in questions for that noun vs. steered towards the target noun in questions for the other noun in the pair. In this setting, steering recovers only a fraction of the target noun’s baseline frequency.

Steering with mean activation differences of questions for a pair of nouns reliably increased the frequency in the answer of the target noun steered towards, see figures 23 and 22 (note that the questions themselves do not contain the target nouns, but are meant to elicit the nouns as answers). Steering was far from fully recovering the baseline noun frequency for suggestive questions, but worked quite well in affecting the answer in neutral questions. This suggests that while answer planning is probably distributed across positions and layers so a single position-layer is not enough to capture it, we did intervene on an important location, manipulating **forward planning** for the answer. Like in the rhyming study, instruction-tuned models show a stronger steering effect than base models in neutral questions, suggesting that instruction tuning boosts planning. For suggestive questions, patterns are much less clear, presumably because of planning is more distributed across positions in informative contexts.

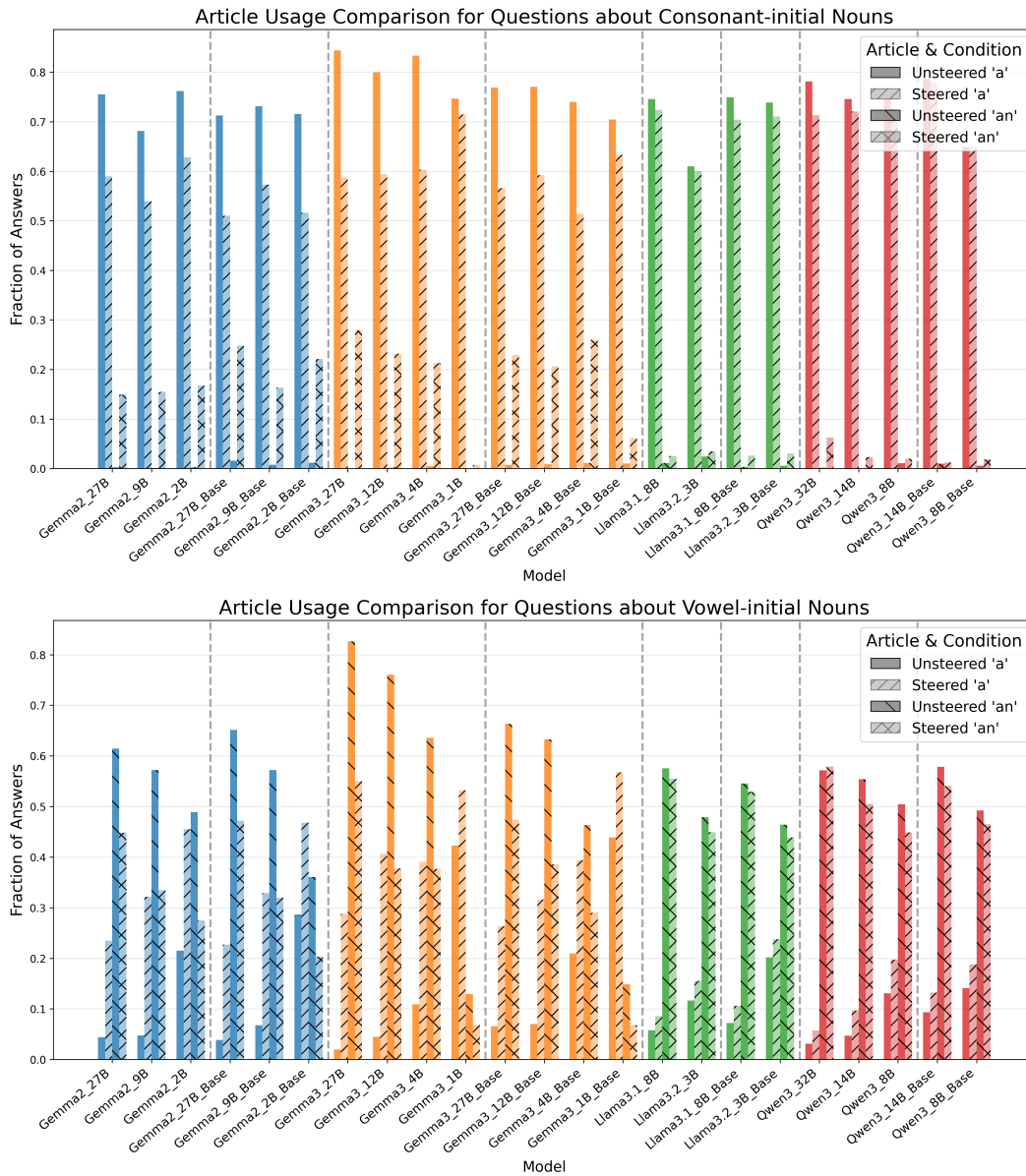


Figure 24: Percentage of answers containing article *a* or *an* as a function of steering on **suggestive questions**. Proportion of *an* increases and proportion of *a* decreases when steering towards a noun beginning with a vowel in a question about a noun beginning with a consonant, and vice versa when steering towards a noun beginning with a consonant in a question about a noun beginning with a vowel.

The word pairs were specifically selected to contain one answer noun that begins with a consonant (and goes with article *a*) and one noun that begins with a vowel (and goes with article *an*). We measure if articles used vary with the target noun when steering. Indeed, the frequency of article forms is modulated predictably by steering both in suggestive questions (fig. 24) and neutral questions (fig. 25), providing evidence for **backward planning** in question answering context.

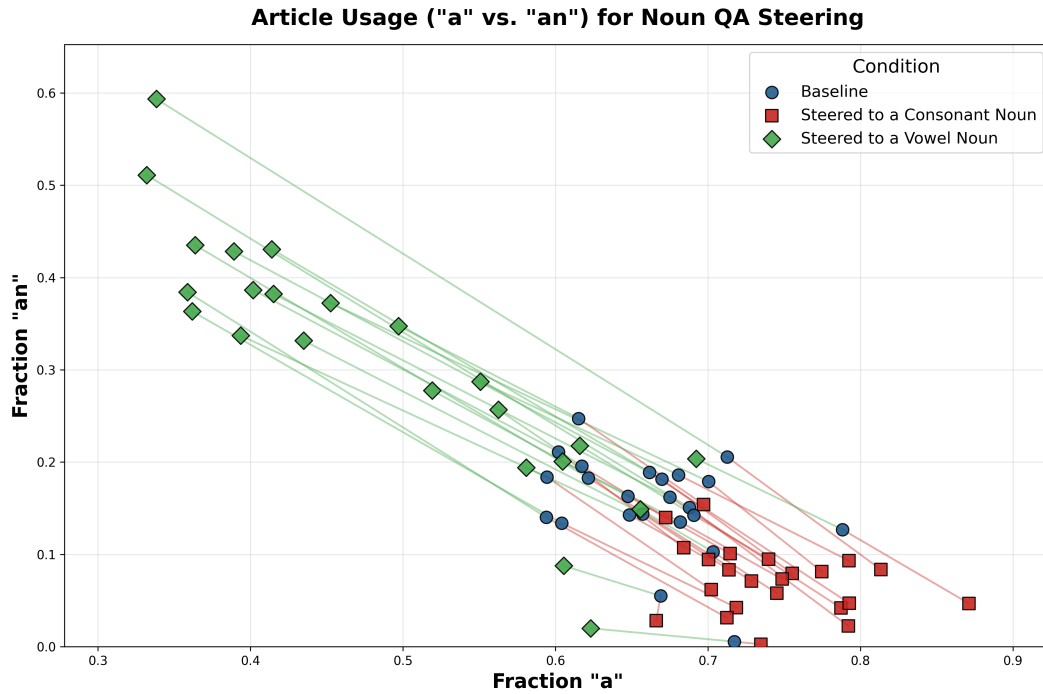


Figure 25: Percentage of generated answers to neutral questions containing article *a* vs. *an* as a result of steering on **neutral questions**. Lines connect data points for steering conditions of the same language model.

## L MORE COMPLEX QUESTION ANSWERING AND SUBJECT-VERB AGREEMENT AS A TEST OF IMPLICIT PLANNING

Similar to Hanna & Ameisen, we constructed a dataset to investigate implicit planning in the context of subject-verb agreement:

*Example 3.* Question: There were 7 keys but 4 were lost. How many keys are left?  
Answer: Now there

All dataset examples follow a similar pattern to Example 3, varying the nouns and corresponding verbs. LLM’s typically produce a verb form of “to be” followed by a number as the next tokens—in this case, “are 3”. This task appears to require implicit backward planning: the model must generate the correct intermediate token (either “is” or “are”) before producing the goal token (the correct number).

The model could solve this task through in two distinct ways. The first is **improvisation**, where some circuits in the model determine the correct verb form without representing the planned answer. The second is **planning**, where the model has some intermediate representation of the correct answer and uses this representation to generate both the appropriate verb and the answer itself. Our experiments provide suggestive evidence for the planning mechanism.

We evaluated this task on Gemma3 27B using a dataset of 80 questions: 40 with answer “1” and 40 with answer “3”, split into 30 training examples (for estimating steering vectors) and 10 test examples.

Table 3 demonstrates that Gemma3 27B reliably produces the correct verb form in the baseline condition. Crucially, steering on the question mark token can shift the model to produce the steered verb form in a substantial fraction of cases. Additionally, steering occasionally succeeds in changing

Metric	Answer is 1	Answer is 3
Correct Verb (%)	97.5	100.0
Correct Number (%)	98.4	100.0
Steered towards Verb (%)	52.8	9.4
Steered towards Number (%)	0.0	9.0

Table 3: Accuracy and steering effectiveness of Gemma3 27B on the subject-verb agreement task. Baseline metrics show the model’s ability to produce correct verbs and numbers. Steering metrics show the percentage of cases where steering on the question mark token successfully shifts outputs toward the target verb form and corresponding number.

the models answer from “3” to “1”. For some prompt this effect is quite large. For The example prompt in 3, steering causes the model to output “is” in 93% of cases and “1” in 58% of cases.

If the improvisation mechanism were true, then we would expect our steering to only be able to manipulate the verb and the effect on the answer token coming entirely though the indirect effect of changing the word. To test this, we conduct another small experiment where we directly append “is” to the input prompt instead of using activation steering. In this case, the model never outputs “1”. This suggests that at least in some cases, our steering vector captures a representation of the planned answer (“1”), which the model then also uses to generate the correct verb form, providing evidence for implicit planning rather than mere improvisation.

## M STEERING EFFECTIVENESS ACROSS LAYERS AND POSITIONS

In this appendix, we include plots with more detailed information on the positions and layers that are most effective in our steering experiments.

### M.1 RHYME STEERING

Figure 26 compares the effectiveness of steering on the newline token for all models. Figures 27-30 report steering effectiveness for both the newline position and the immediately preceding token position for all models in consideration.

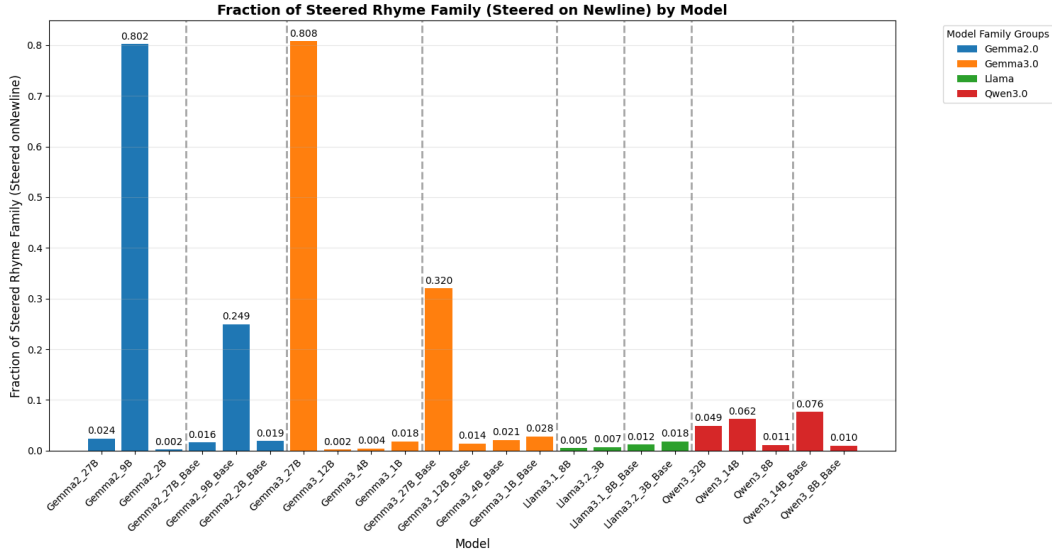


Figure 26: Steering effectiveness for newline token position, all models.

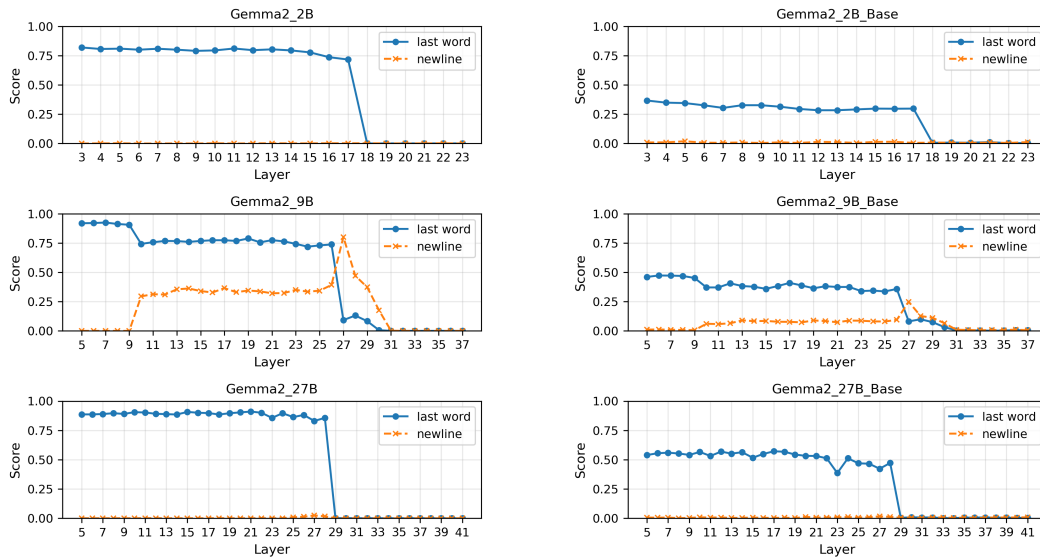


Figure 27: Steering effectiveness by steering position and layer, Gemma2 models.

Lindsey et al. reported evidence for planning representations on the newline token at the end of the first line for Claude Haiku. In our experiments, steering on the newline token was only effective in

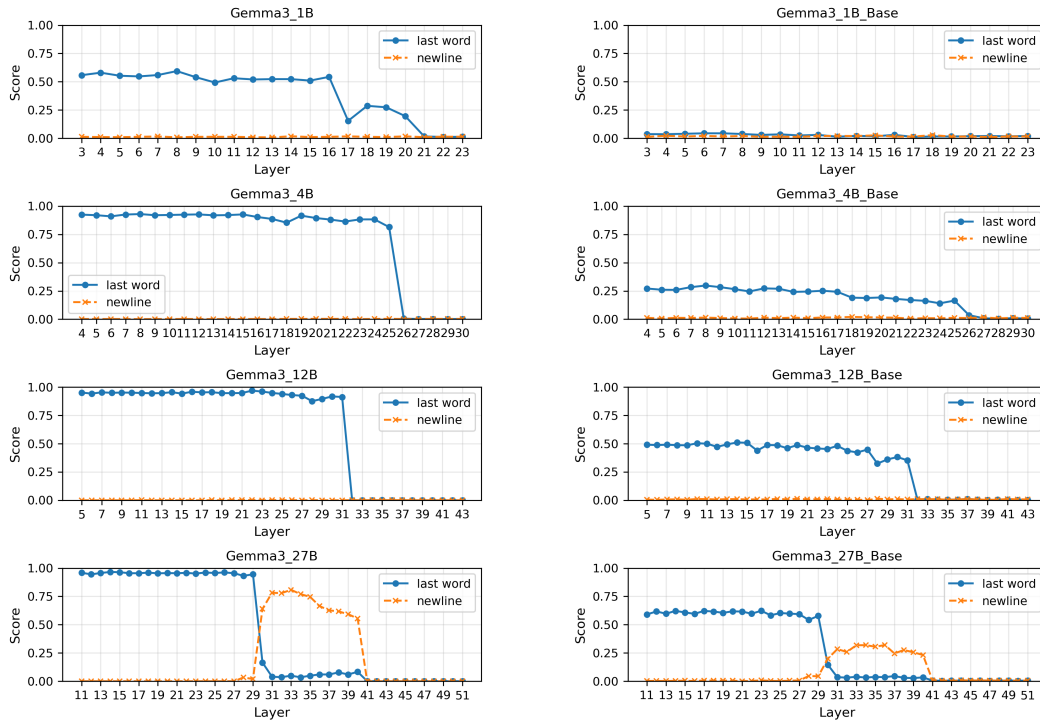


Figure 28: Steering effectiveness by steering position and layer, Gemma3 models.

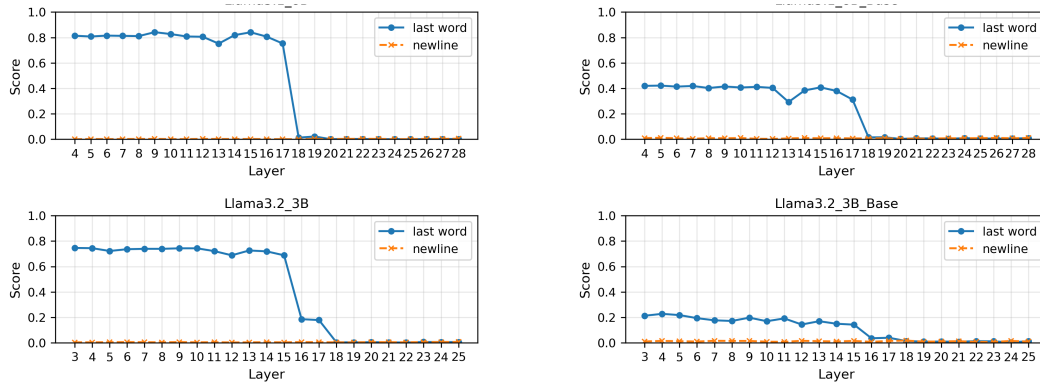


Figure 29: Steering effectiveness by steering position and layer, Llama3 models.

select models in some middle layers. For all models however, steering worked on the pre-newline (last word) token in earlier layers. Rhyme steering effectiveness for all layers and positions of all models is reported in figures 27–30.

## M.2 NOUN QUESTION ANSWERING STEERING

Noun QA steering effectiveness for all layers and positions of all models is reported in figures 31–34.

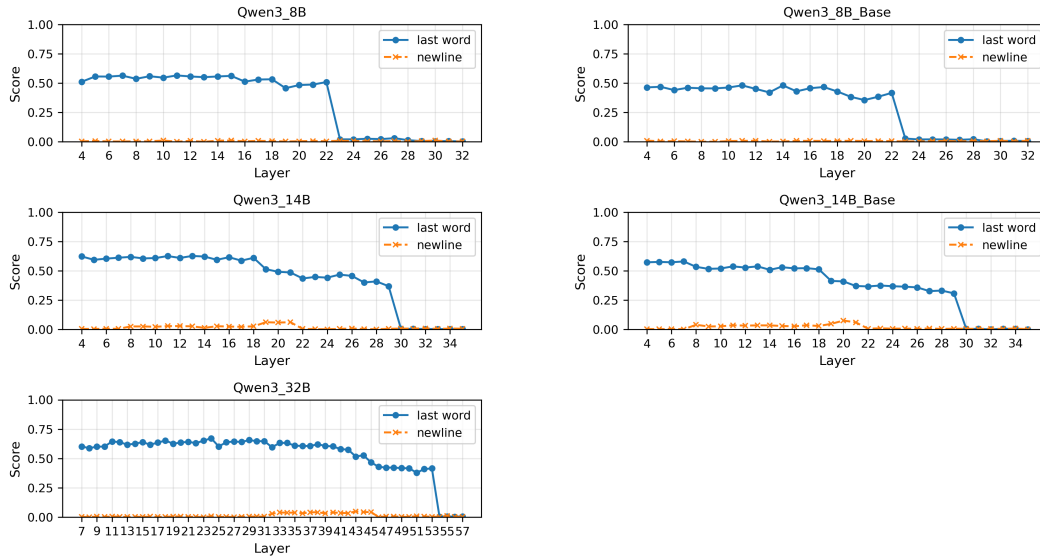


Figure 30: Steering effectiveness by steering position and layer, Qwen models.

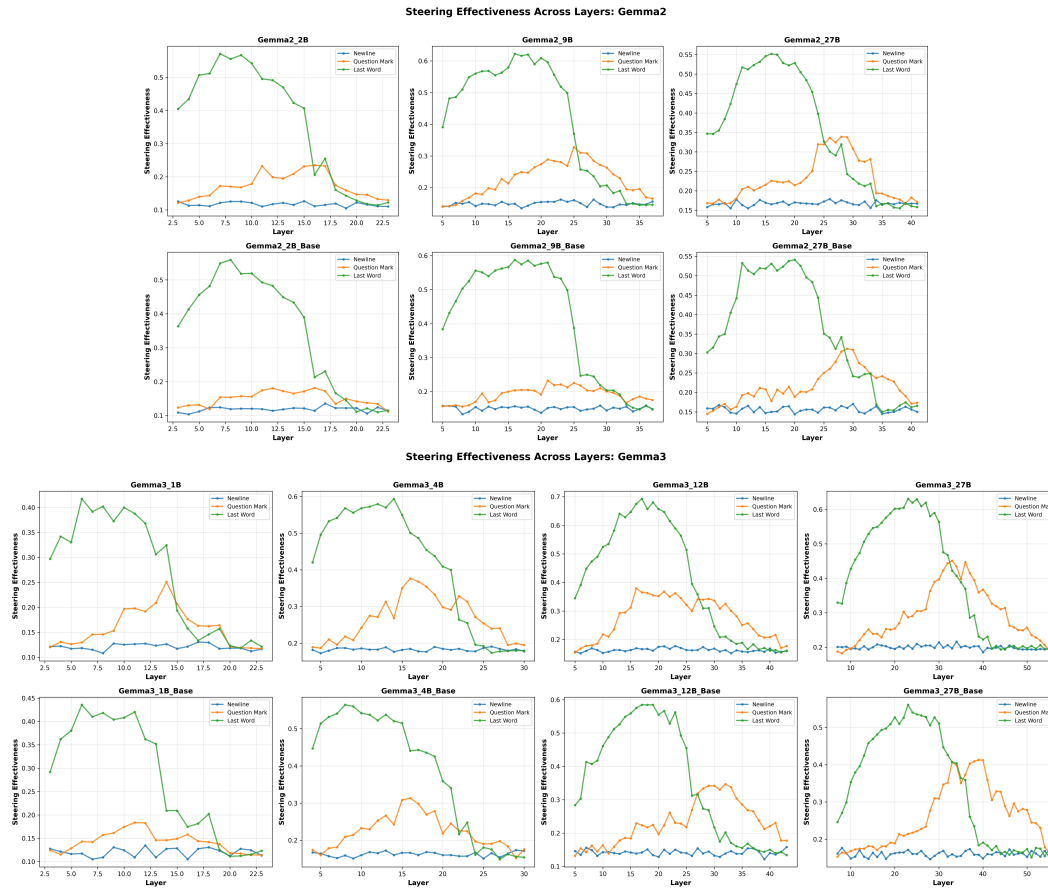


Figure 31: Steering effect (fraction of rollouts containing the target noun) across layers and positions for steering. Neutral questions, Gemma models.

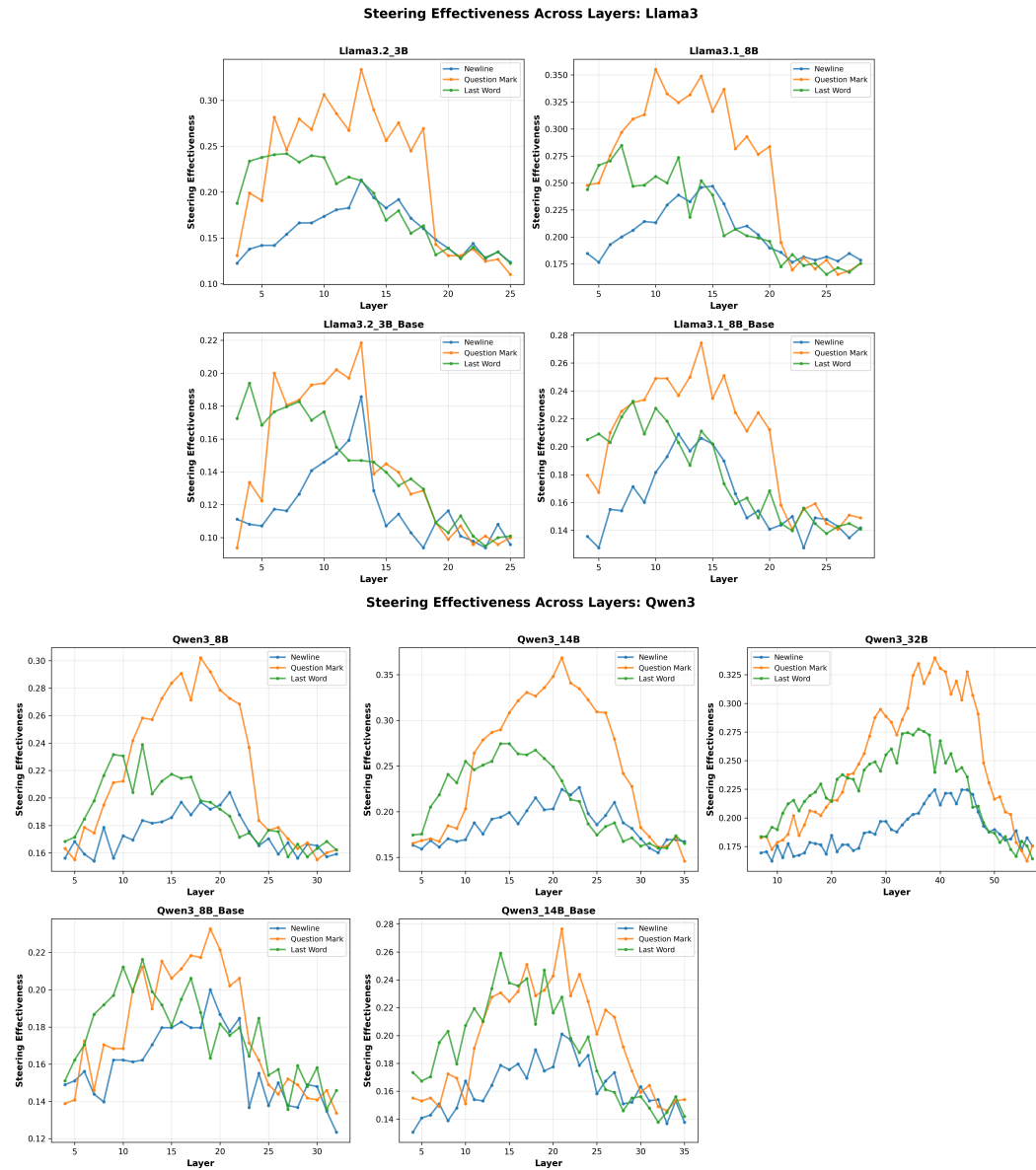


Figure 32: Steering effect (fraction of rollouts containing the target noun) across layers and positions for steering. Neutral questions, Llama and Qwen models.



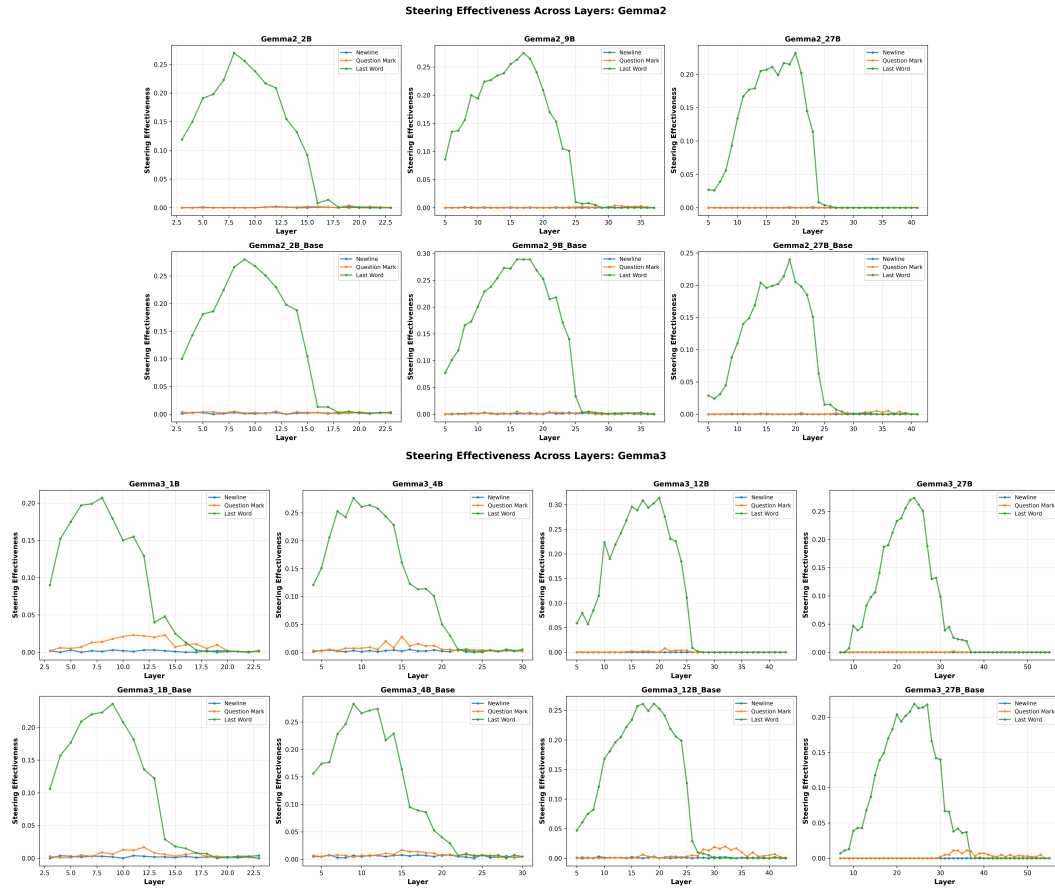


Figure 33: Steering effect (fraction of rollouts containing the target noun) across layers and positions for steering. Suggestive questions, Gemma models.



Figure 34: Steering effect (fraction of rollouts containing the target noun) across layers and positions for steering. Suggestive questions, Llama and Qwen models.