Steering Language Models in Multi-Token Generation: A Case Study on Tense and Aspect

Alina Klerings[1], Jannik Brinkmann[2], Daniel Ruffinelli[1], Simone Paolo Ponzetto[1]

[1]University of Mannheim, [2]Technical University Clausthal

alina.klerings@uni-mannheim.de

Abstract

Large language models (LLMs) are able to generate grammatically well-formed text, but how do they encode their syntactic knowledge internally? While prior work has focused largely on binary grammatical contrasts, in this work, we study the representation and control of two multidimensional hierarchical grammar phenomena—verb tense and aspect—and for each, identify distinct, orthogonal directions in residual space using linear discriminant analysis. Next, we demonstrate causal control over both grammatical features through concept steering across three generation tasks. Then, we use these identified features in a case study to investigate factors influencing effective steering in multi-token generation. We find that steering strength, location, and duration are crucial parameters for reducing undesirable side effects such as topic shift and degeneration. Our findings suggest that models encode tense and aspect in structurally organized, human-like ways, but effective control of such features during generation is sensitive to multiple factors and requires manual tuning or automated optimization.[1]

1 Introduction

Growing evidence on the generative capabilities of large language models (LLMs) suggests that they encode structural properties of language—such as syntax trees—within their hidden representations (Hewitt and Manning, 2019; Diego Simon et al., 2024). Studying the representation of these properties can reveal how syntactic structures in models compare to those in human language. It may also help develop more linguistically aware AI systems, extend model capabilities to low-resource languages, improve the interpretability of generation, and assist in diagnosing systematic errors in tasks like machine translation (López-Otal et al., 2025).

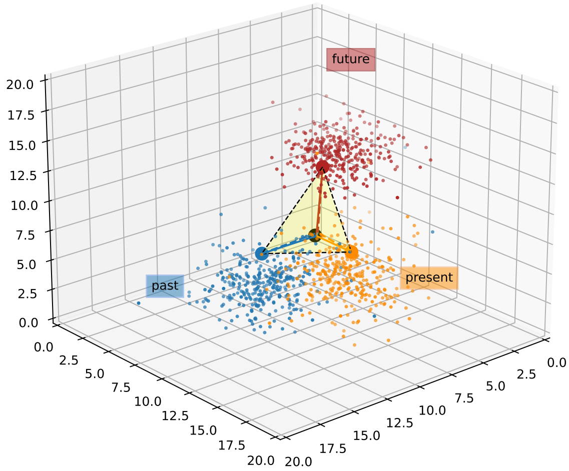

Figure 1: Activations from Qwen-7B projected along three identified feature directions that represent the categorical values of **tense**: present, past and future.

Prior work has assessed grammatical competence with behavioral evaluations, using black-box measures such as perplexity (Chen et al., 2024) and grammatical acceptability benchmarks (Warstadt et al., 2020; Hu et al., 2020). However, these methods do not reveal how grammatical concepts are internally represented. To address this, recent interpretability studies have examined the mechanisms underlying syntactic competence, such as indirect object identification (Wang et al., 2023) and subject-verb agreement (Ferrando and Costa-jussà, 2024), providing detailed insights into the role of specific architectural components. Yet, they largely focus on binary grammatical contrasts (e.g., singular vs. plural agreement), and often evaluate on single-token continuations, which limit their ability to capture more compositional or distributed grammatical phenomena.

In this work, we take a complementary approach focused on grammatical **tense** and **aspect**, two core verb properties that encode temporal relations and often span multiple tokens. While tense positions

[1] https://github.com/klerings/tense-aspect

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8621–8639
November 4-9, 2025 ©2025 Association for Computational Linguistics

an event in different time periods relative to the moment of speech (i.e., past, present or future),

a) She *drove* her car.
b) She *drives* her car.
c) She *will drive* her car.

aspect introduces an additional layer of temporal structure—describing whether an event is completed, ongoing or repeated (e.g., simple, perfect, progressive or perfect progressive) (Klein, 2009).

d) She *has driven* her car.
e) She *is driving* her car.
f) She *has been driving* her car.

Unlike binary syntactic features, tense and aspect involve multiple discrete categories and can be combined (e.g., past perfect). Theoretically, they are independent grammatical features: any aspect should, in principle, be expressible in any tense. This separability is useful for a mechanistic analysis, as it allows for a systematic exploration of how these categories interact. We locate and analyze the representation of both properties in two models via linear probes and Linear Discriminant Analysis (LDA) (Park et al., 2024a), and test their causal relevance (Vig et al., 2020; Mueller et al., 2024) by using them to steer LLM outputs in multi-token generation.

While prior work has demonstrated representational steering for single-token continuations and has focused on semantic properties, our approach applies steering to full sentences by intervening on every generated token individually, and focuses on grammatical phenomena. Since steering multiple tokens increases the risk for degenerate outputs and unintended side effects (Bricken et al., 2023; Stickland et al., 2024), we analyze when and how strongly to intervene. We find that steering effectiveness depends on activation norms, model architecture, and the nature of the task.

Contributions. Our key contributions are as follows:

- We identify distinct, near-orthogonal directions in LLM residual space encoding tense and aspect, using probes and LDA.
- We demonstrate causal control over these grammatical features through concept steering, thereby mediating the LLM output during multi-token generation.
- We investigate conditions for effective steering, analyzing the influence of steering strength and location, activation norm, model type and task on side effects.

2 Background and Related Work

Linear Representation Hypothesis and Concept Directions. The linear representation hypothesis is the assumption that abstract features are encoded as linear directions in LLM residual space (Elhage et al., 2022; Nanda et al., 2023; Park et al., 2024b; Marks and Tegmark, 2024; Park et al., 2024a). This has led to the notion of **feature directions**—unit vectors in activation space corresponding to specific properties—and, more generally, **feature subspaces**, which span sets of related directions (e.g., different tense directions in shared tense subspace) (Geiger et al., 2025; Mueller et al., 2024).

Such representations have been extracted using sparse autoencoders (Bricken et al., 2023; Huben et al., 2024), dimensionality reduction (Gurnee and Tegmark, 2024), (Heinzerling and Inui, 2024), and supervised linear probes (Marks and Tegmark, 2024). Recently, Park et al. (2024a) introduced an LDA-based method to identify scaled directions that capture categorical structure. Unlike prior approaches focused on concepts with natural opposites (e.g. MALE VS. FEMALE), this framework models categorical concepts (e.g., TENSE) as sets of **binary features** (e.g., FUTURE, PRESENT, PAST), independently of predefined class structures.

Steering Language Models. Feature steering has mainly targeted semantic features like sentiment (Rimsky et al., 2024; Lee et al., 2025) or numbers (Heinzerling and Inui, 2024), while categorical linguistic properties remain underexplored. Earlier work has focused on single-token interventions, with recent studies beginning to explore steering full-sentence generation (Lee et al., 2025; Rimsky et al., 2024; Wu et al., 2025). Systematic evaluations, e.g., of steering strength, side effects, or degeneration, are still sparse (Pres et al., 2024).

Our work addresses these gaps by steering categorical grammatical features—specifically, tense and aspect—during sentence-level generation with controlled intervention strength and position-aware application. We extend the LDA-based framework from Park et al. (2024a) by broadening its application from lexical to sentence-level grammatical categories.

Grammatical Knowledge in LMs. The grammatical competence of language models has been an ongoing research topic since before the emergence of LLMs, with subfields such as "Bertology" (Rogers et al., 2020) focusing on earlier architectures like BERT. Tense has been explored in behavioral evaluations as well as causal analyses (Merullo et al., 2024; Zhang et al., 2025), but typically via binary distinctions (e.g., past vs. present) and single-token interventions—aside from Brinkmann et al. (2025), who consider open-ended generation. However, tense and aspect have not been jointly analyzed in a unified framework. We address this by studying these features in multi-token generation, combining probing, representation space analysis, and causal steering to examine how grammatical concepts are encoded and can be controlled. For a more comprehensive review of related work, see App. B.

3 Locating Tense and Aspect

We focus on the interlinked grammatical concepts of verb tense and aspect by localizing and visualizing their subcategories via exploratory methods.

3.1 Experimental Setup

Tense and aspect are core grammatical categories that encode temporal structure. Tense refers to the time at which an event occurs (present, past, future), while aspect characterizes its temporal structure (simple, progressive, perfect, perfect progressive). In theory, each tense can occur in combination with each aspect, yielding a grid of 12 distinct tense-aspect forms (see App. C). This system makes these features ideal for studying multidimensional hierarchical structure in LLMs. In practice, such regular and compositional combinations are rare across languages. The comparatively structured tense-aspect system in English is an exception (Klein, 2009). Therefore, we focus our study on English rather than languages with less regularity.

Data. We analyze sentences annotated with grammatical tense and aspect, requiring each sentence to contain a single, unambiguous tense-aspect combination to isolate a clear signal for each target variable. We use sentences from the Penn Treebank (Marcus et al., 1993) annotated with PropBank (Kingsbury and Palmer, 2002), which provides verb-specific tense and aspect labels. After filtering out ambiguous sentences, the class distribution is highly skewed, with most examples in the simple aspect and some targets having fewer than ten instances. To address this imbalance, we augment rare classes with synthetic examples generated by GPT-4o (Hurst et al., 2024), resulting in 1,543 labeled sentences. Augmentation improves balance, but some categories remain overrepresented (e.g., simple past), therefore we downsample them for training the classifiers. For evaluation, we use the verb tense subset of 281 sentences from BIG-bench (Srivastava et al., 2023; Logeswaran et al., 2018) (see App. D for dataset details).

Models. We use the models Llama-3.1-8B-Instruct (Dubey et al., 2024), primarily trained on English data, and Qwen-2.5-7B-Instruct (Qwen et al., 2025), with a focus on English and Chinese, that are commonly studied for similar analyses. For brevity, we omit version numbers and the 'Instruct' suffix throughout the remainder of the paper.

Localization via Probing. To localize the target concepts—tense, aspect and their combination—we train linear probing classifiers (Belinkov, 2022) on the hidden representations of the pre-trained LLMs. Let the model be defined as $f_\theta : x \to h$, where θ are learnable parameters, $x = (x_1, ..., x_N)$ the input tokens, and $h^l = (h^l_1, ..., h^l_N)$ the hidden representations at layer l. At each layer, we train a probing model $p^l : h^l_{\text{agg}} \to y$ that maps aggregated hidden states h^l_{agg} to the corresponding tense, aspect or tense-aspect labels y using multinomial logistic regression. We compute h_{agg} per layer as follows,

$$h_{\text{agg}} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_i, \qquad (1)$$

summing token-level representations and normalizing by the square root of sequence length N, a strategy that outperforms other aggregation methods, see App. E for more details. Finally, we mean-center the aggregated hidden representations before classification.

Representation Geometry. To analyze how tense and aspect are represented in model activations, we use the framework proposed by Park et al. (2024a) for handling categorical features. In this approach, each categorical concept (e.g., TENSE) consists of a set of subordinate feature values (e.g., PAST, PRESENT, FUTURE). To model each categorical value as a direction, they are represented as **binary features**: {IS_PAST, IS_NOT_PAST}, {IS_PRESENT, IS_NOT_PRESENT}, {IS_FUTURE, IS_NOT_FUTURE}. Following this framework, we estimate vector representations for each feature value using the variant of LDA proposed by Park et al. (2024a). The objective is to find a vector that reduces within-class variation and highlights differences to all other classes. Unlike standard LDA, which relies on both within-class and between-class covariance, this variant omits the latter, enabling the computation of each vector independently of the others.

Formally, for each binary feature w, we compute a **normalized class direction** $\tilde{h}_w$ from the empirical mean of the class-specific activations $\mathbb{E}(h_w)$, adjusted by the pseudo-inverse of the class covariance $\mathrm{Cov}(h_w)^\dagger$:

$$\tilde{h}_w = \frac{\mathrm{Cov}(h_w)^\dagger \mathbb{E}(h_w)}{\|\mathrm{Cov}(h_w)^\dagger \mathbb{E}(h_w)\|_2}. \tag{2}$$

This unit vector captures the direction of the class in residual space. To incorporate the strength of the signal, we scale the direction by the projection of the class mean onto it:

$$\bar{\ell}_w = (\tilde{h}_w^\top \mathbb{E}(h_w))\tilde{h}_w. \tag{3}$$

The resulting vector $\bar{\ell}_w$ encodes the orientation and intensity of the concept in activation space.

Importantly, because each vector is computed independently, the method avoids enforcing any pre-defined class structure of tense and aspect. Instead, the representational geometry that emerges reflects the structure learned by the model itself.

Another key concept introduced by Park et al. (2024a) is **binary contrast**, which captures the distinction between two categorical values within the same parent category and is computed as the vector difference between their feature vectors. In our analysis, we use binary contrasts to model categories in a lower-dimensional space. This allows us to (i) compare subordinate features within a category where their number exceeds the available representational dimensions (i.e., for aspectual values), and (ii) approximate latent dimensions for cross-category comparisons between tense and aspect.

3.2 Results

Representations of tense and aspect emerge early in the model. Our probing classifiers predict tense and aspect from the embedding layer with f1-scores above 90% and improve further with model depth, especially for the fine-grained tense-aspect combinations, see Table 1.

	Tense	Aspect	Tense-Aspect
Llama-8B	1.0	0.98	0.93
Qwen-7B	1.0	0.98	0.92

Table 1: Target-wise F1-scores from the best-performing probe across layers.

This demonstrates that contextualization is beneficial, as many tenses and aspects are expressed across multiple tokens. For detailed results across all layers and more aggregation strategies, see App. E. The findings are consistent across targets and models and similar to earlier work that suggests syntactic processing happens before more complex semantic processing (He et al., 2024).

Grammatical properties form subspaces in representation space. To analyze the LDA results, we use the 2D and 3D visualizations of Park et al. (2024a). To assess whether hidden representations encode a separation between categories of a single grammatical feature, we project test set embeddings from Qwen-7B onto selected directions.

For tense, we use $\bar{\ell}_{\text{PRESENT}}$, $\bar{\ell}_{\text{PAST}}$ and $\bar{\ell}_{\text{FUTURE}}$ as projection axes. For aspect, we use binary contrasts retrieved from vector differences to represent the four categories: $\bar{\ell}_{\text{PROGRESSIVE}} - \bar{\ell}_{\text{SIMPLE}}$, $\bar{\ell}_{\text{PERFECT}} - \bar{\ell}_{\text{SIMPLE}}$ and $\bar{\ell}_{\text{PERFECT PROGRESSIVE}} - \bar{\ell}_{\text{SIMPLE}}$.

In both cases, 3D projections reveal distinct clusters corresponding to the underlying grammatical categories: three well-separated clusters for tense (Figure 1, explained variance: 0.72) and four for aspect (Figure 2, explained variance: 0.70). They span a convex region in both cases, suggesting that the feature vectors define structured subspaces. See App. F for additional cluster quality metrics of both models.

Tense and aspect exhibit representational independence. Next, we examine the relationship between the two grammatical categories by projecting representations onto their respective latent dimensions. Specifically, we use the binary contrasts $\bar{\ell}_{\text{TENSE}} = \bar{\ell}_{\text{FUTURE}} - \bar{\ell}_{\text{PAST}}$ and $\bar{\ell}_{\text{ASPECT}} = \bar{\ell}_{\text{PROGRESSIVE}} - \bar{\ell}_{\text{PERFECT}}$ as proxies for the tense and aspect dimensions[2].

Figure 3 shows the projection of data points onto

[2] Any binary contrast among subcategories (e.g., FUTURE-PAST, PRESENT-PAST, FUTURE-PRESENT) lies in the same one-dimensional parent-contrast subspace. The vector differences in Fig. 3 were chosen as examples.

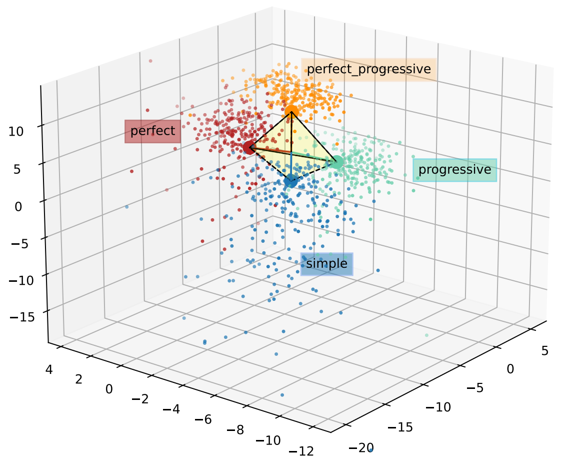

Figure 2: Projection of Qwen-7B hidden states (L0) along LDA-based **aspect** directions.

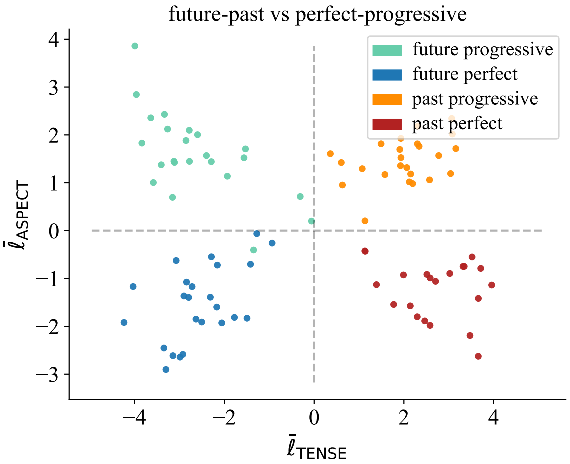

Figure 3: Projection of Llama-8B hidden states (L7) along $\bar{\ell}_{\text{TENSE}}$ and $\bar{\ell}_{\text{ASPECT}}$ clearly separate the data points according to their grammatical property.

the inferred tense (x-axis) and aspect (y-axis) dimensions. Similar to earlier feature-wise visualizations, the points cluster according to their grammatical categories. The groups are organized near-orthogonally in latent space, reflecting a clear separation between tense and aspect. This is further supported by the near-zero cosine similarity of 0.02 between the vectors $\bar{\ell}_{\text{TENSE}}$ and $\bar{\ell}_{\text{ASPECT}}$ (see App. F).

4 Multi-Token Steering

After identifying correlational evidence for tense and aspect directions in model representations, we test their causal impact on model behavior via targeted interventions during text generation. We evaluate functional selectivity by checking whether the manipulated outputs express the steered verb property while preserving other verb features and the original meaning. This quantitative analysis is followed by a qualitative study in which we manually examine model outputs and investigate the impact of steering location, strength and duration.

4.1 Experimental Setup

Tasks. We consider three complementary tasks (prompt details in App. G):

1. **Random Sentence Task:** The model is prompted to generate an open-ended sentence, testing whether grammatical concepts can be induced in semantically unconstrained settings.

2. **Repetition Task:** In a few-shot setup, the model must repeat a sentence after two example repetitions. This copying task requires interventions to override contextual information. It evaluates the ability to steer generation when the model's default behavior follows pattern induction.

3. **Temporal Translation Task:** We use a similar few-shot setup where the model must "translate" a sentence into a different tense or aspect. Unlike repetition, this requires internal transformation, allowing us to test whether interventions can influence more complex linguistic transformations.

Steering Methods. We perform steering at a single transformer layer l on the final position $i = -1$ of the input sequence at every generation step. Concretely, we update the residual stream activation vector $h_i^l \in \mathbb{R}^d$ by adding and/or subtracting the normalized LDA-derived concept directions from earlier. These unit vectors correspond to specific tense and aspect values. A scalar steering factor $\alpha \in \mathbb{R}$ scales the strength of the modification. We evaluate three distinct steering strategies across all layers and different α values:

1. **Target-Addition Only (TA):** The standard steering approach which is commonly used in related work such as Rimsky et al. (2024), directly adds the normalized target direction $\bar{\ell}_T$ to the current activation:

$$h^{\text{steered}} = h_i^l + \alpha \bar{\ell}_T. \tag{4}$$

2. **Target-Addition with Source-Subtraction (TA+SS):** To simultaneously steer the target concept and suppress a known source concept

$\bar{\ell}_S$, we introduce source-subtraction, which subtracts the source direction with equal weight. This is particularly useful for non-binary features, where source and target are not simply inverses:

$$h^{\text{steered}} = h_i^l + \alpha \bar{\ell}_T - \alpha \bar{\ell}_S. \tag{5}$$

3. **Target-Addition with Projection Subtraction (TA+Proj-SS):** Instead of subtracting the full source vector, this method removes only the component of the activation that lies along $\bar{\ell}_S$. This is achieved by computing and subtracting the projection of h_i^l onto $\bar{\ell}_S$:

$$h^{\text{steered}} = h_i^l + \alpha \bar{\ell}_T - (h_i^l \cdot \bar{\ell}_S)\bar{\ell}_S. \tag{6}$$

To ensure comparability between steered and original generations, we use greedy decoding, where the most likely token is selected at each step.

Evaluation Metrics. We evaluate steering success by giving the generated outputs to the model in a new forward pass without any interventions, extracting their representations and applying the trained probing classifiers, following Brinkmann et al. (2025). We probe not only for the steering target but also for the respective other property. To quantify the effect of our interventions, we define the following four performance metrics:

$$\begin{aligned}\text{Steering Success} &= \frac{|S|}{N},\\ \text{Degenerate Rate} &= \frac{|D|}{N},\\ \text{Efficacy} &= \frac{|S \setminus D|}{N},\\ \text{Selectivity} &= \frac{|S_F \setminus D|}{N}.\end{aligned}$$

Here, N is the number of test samples[3], $S \subseteq \{1, \ldots, N\}$ is the set of successfully steered samples and $D \subseteq \{1, \ldots, N\}$ the set of degenerate outputs. An output is considered degenerate, if it either (i) forms an incomplete sentence by missing a verb, as detected by a part-of-speech (POS) tagger, or (ii) exhibits excessive n-gram repetition or low n-gram diversity (see App. I for thresholds). S_F is the subset of S for which the probe's label for the not steered property stays constant. Finally, we report the relative change in perplexity to measure the impact of steering on fluency and coherence.

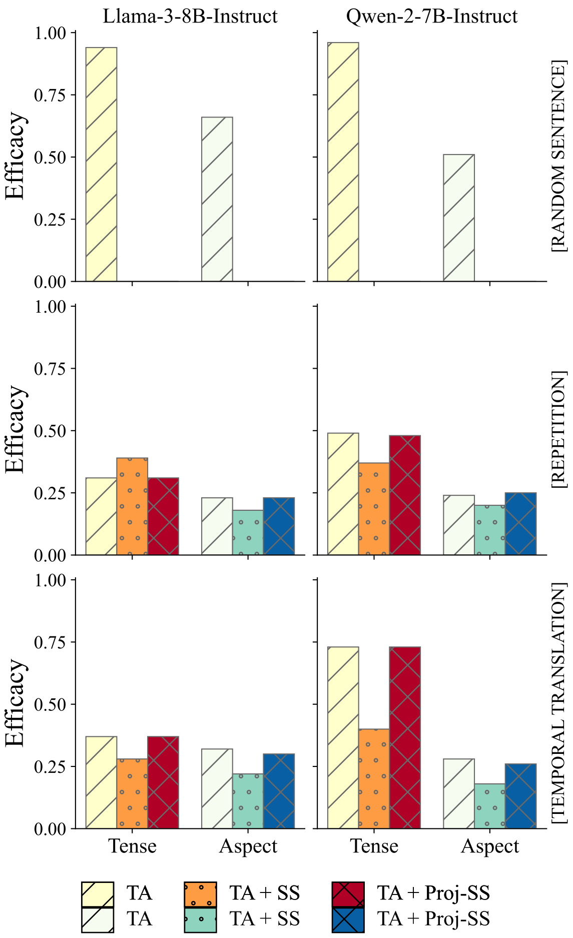

Figure 4: Efficacy of different steering methods. TA+SS and TA+Proj-SS are not applied for the Random Sentence Task because there is no source feature direction to be subtracted. Both models show similar trends with tense being easier to steer than aspect, and random sentences easier than few-shot tasks.

4.2 Quantitative Results

For each task and steering method, we perform a grid search over all layers and selected α values (see App. H) and report the configuration that yields the highest efficacy (Figure 4).

Steering success varies widely by task and target. Overall, steering tense achieves substantially higher success than aspect. For example, on the random sentence task, efficacy is near-perfect for tense (94% for Llama-8B, 96% for Qwen-7B) but noticeably lower for aspect (66% for Llama-8B, 51% for Qwen-7B). Steering in the few-shot settings reduces performance for both targets, with the best scenario reaching 73% (Qwen-7B, tense) and the worst just 18% (aspect for both models). Steering tense or aspect is significantly more difficult when the task requires conflicting verb properties to

[3] N is task dependent, see App. G.

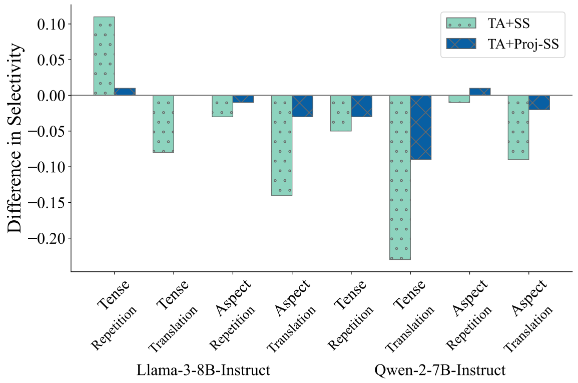

Figure 5: Subtracting the source concept vector (TA+SS) reduces selectivity. Negative bars denote selectivity drop with respect to simple target addition, but this effect can be partially mitigated through more targeted subtraction (TA+Proj-SS).

the steering target[4]. This trend is largely consistent across both model architectures.

Although efficacy implicitly captures output quality through the degenerate rate, we additionally report relative perplexity changes between steered and unsteered generations to evaluate the general impact of steering on text coherence and fluency. Results show low perplexity increases across most scenarios, with few outliers, demonstrating that our steering vectors maintain output quality and perform targeted interventions (App. J.2).

Projection subtraction improves selectivity. Another interesting finding is that subtracting the source concept vector often harms steering performance (-4% and -12.5% on average for Llama-8B and Qwen-7B respectively), potentially because it introduces too much additional change to the residual stream. It also reduces selectivity (Figure 5). However, replacing full vector subtraction with projection subtraction mostly “mitigates” this issue, both for efficacy and selectivity, indicating that more targeted interventions—removing only the component aligned with the source direction—are more effective. We explore this finding further in § 4.4 and provide details on the correlation between efficacy and selectivity in App. J.1.

Activation norm determines steering factor. We find that Llama-8B requires significantly lower α values (5-25) for effective steering compared to Qwen-7B (100-250), a pattern consistent across tasks. Moreover, the optimal α tends to increase with depth for both models (App. J.3). One explanation lies in the activation norm, which increases similarly across layers and is generally higher for Qwen-7B. Kobayashi et al. (2020) have shown that layers with larger activation norms carry more information, suggesting that stronger interventions are required to overwrite pre-existing signals. To test this hypothesis, we examine the projection magnitude onto the source feature direction, that is, the strength with which the original tense and aspect are encoded, and observe a similar growth with depth. Thus, even though the feature’s strength stays roughly constant relative to the activation norm, its absolute magnitude grows across layers, requiring proprotionally larger α values for effective steering.

4.3 Qualitative Behavior and Failure Modes

We present example outputs for the random sentence task in Table 3 and for the few-shot tasks in Table 4 and App. J.5. They demonstrate that while steering grammatical properties is possible, it can also lead to unintended changes in content—from slight alterations (e.g., in the repetition task) to complete topic shifts (e.g., in the random sentence task). This undesired behavior is not captured by the four evaluation metrics, but is crucial to identify when aiming for targeted and selective steering.

To assess topic shift across tasks, we focus on the most effective steering method for each model and target. We compute the semantic similarity metric BERTScore (Zhang et al., 2020) between unsteered and steered outputs, considering only successfully steered samples. On average, steering during the few-shot tasks causes less topic shift (i.e., higher semantic similarity in Table 2), likely because guided generation constrains their output and maintains consistency. In contrast, open-ended generation leaves more room for the steering intervention to reshape the output, leading to higher inconsistency.

Random Sentence	Repetition	Temporal Translation
0.56 ± 0.06	0.69 ± 0.09	0.77 ± 0.07

Table 2: BERTScore between unsteered and steered output, averaged across models and targets, reveals higher consistency for few-shot tasks.

4.4 Exploring the Effect of Steering Location

Building on the finding that optimal steering strength is model- and layer-dependent, we now

[4] E.g., for the repetition task, the sentence "I jumped into the lake." needs to be repeated in its original tense (past), but when steering "future", the intervention is contradicting to the tense the task would require.

TARGET	PROMPT Generate a single sentence:	 Output a random sentence:	 Generate one sentence:
-	The sun was setting over the ocean, casting a warm golden light over the waves.	The sun is shining brightly in the sky.	The new policy will require all employees to undergo a background check before being hired.
PAST	The sun was setting over the vast, open plain, casting a warm glow over the scattered wildflowers that swayed gently in the breeze.	The sun was shining brightly on the day of the wedding.	The sun was shining brightly on the day of the wedding, casting a warm glow over the happy couple as they exchanged their vows.
PRESENT	I am a 25-year-old woman who has been a vegetarian for 10 years and has been a vegan for 5 years.	I am a 25-year-old man who is a student at a university in the United States.	I am a 25-year-old woman who is a student at a university in the United States.
FUTURE	The new policy will be implemented in all schools by the start of the next school year.	I'm going to the store to buy some milk.	The new policy will be implemented in all schools by the end of the 2023-2024 school year.

Table 3: Examples of steering Llama-8B. Steering enforces the target tense regardless of the model's original choice, but often shifts the topic. Colors indicate tense, identified via probing classifiers.

Prompt	Maya was writing a story. \\ Maya was writing a story. She accepted that offer.\\ She accepted that offer. He has thought about this. \\
Output (unsteered)	He has thought about this.
Output (steered *past*)	He had not thought about anything else.

Table 4: Example of Llama-8B on repetition task. *Orange*: tense of unsteered outputs, *blue*: target.

explore where and for how long to apply steering to maximize efficacy and reduce side effects. We conduct a qualitative case study, targeting two verb tenses across three examples per few-shot task using TA and TA+SS. Due to the need for token-level POS-annotations, we limit our analysis to a small set of examples.

We differentiate two steering locations: the *prompt* (i.e., modifying input representations during the initial forward pass), and the *generated tokens* (i.e., intervening during subsequent generation steps). For each, we compare single versus multi-token interventions and both index-based and POS-informed steering. We find that the same steering vector can lead to successful modification, no effect or even degeneration on the same sample, depending on the steering location. Representative outputs are shown in Tables 5 and App. J.5.

Generation-time steering is more effective. Steering during generation is consistently more effective across both tasks, while prompt-based interventions succeed only for the repetition task. A plausible explanation is that repetition relies on pattern induction, whereas temporal translation involves more complex grammatical reasoning, making earlier interventions harder.

Steering before the verb works best. The optimal steering duration depends on the location. For effective prompt interventions, all verb tokens need to be steered—suggesting that, while multi-token expressions may aggregate meaning at their final token (Feucht et al., 2024), a single impulse at the end of a verb phrase is insufficient to overwrite all previous verb information. In contrast, generation-time steering is sensitive to timing and duration: late or extended steering can cause topic shift and degeneration. We find that verb properties are steered most effectively just before the generated verb.

Target addition does not require source subtraction. Interestingly, steering positions that succeed with TA+SS also succeed with TA, which aligns with the findings of our quantitative method comparison in § 4.2. If the source and target directions were truely orthogonal, then removing the source

Steering Position	**Prompt Tokens**	**Generated Tokens**	**Output (TA)**	**Output (TA-SS)**
all verb tokens in prompt	...It is snow ing . \\	It is snow ing .	It was snowing.	It was snowing.
last verb token in prompt	...It is snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
sentence end in prompt	...It is snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
final token in prompt	...It is snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
final tokens during generation	...It is snow ing . \\	It is snow ing .	It was a day...	It was a long...
generated token before verb	...It is snow ing . \\	It is snow ing .	It was snowing.	It was snowing.
first generated verb token	...It is snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
all generated verb token	...It is snow ing . \\	It is snow ing .	It is snowing.	It is snowing.

Table 5: Llama-8B: Steering PAST on the repetition task for the prompt: "He is crying. \\ He is crying.\n\n We were having dinner. \\ We were having dinner. \n\n It is snowing. \\". *Generated Tokens*: unsteered output.

should improve steering by eliminating conflicting information about the feature of interest. The comparable results between TA and TA+SS, however, suggest that the directions may not be fully independent, or that the effect of the target direction dominates in practice. Understanding how these directions interact—and whether source subtraction helps or hinders—remains an open question for future work.

While our analysis is qualitative and grounded in grammatical insights into sentence structure of prompt and output, it emphasizes the critical role of steering position and duration, in addition to steering strength. We encourage future research to develop automated methods for identifying task-specific optimal steering positions, building on early work such as Lee et al. (2025).

5 Discussion and Conclusion

In this work, we studied the representation and controllability of two categorical grammar features in LLMs, this section summarizes our findings.

Syntactic categories are represented orthogonally in latent space. Our findings show that language models obtain structural representations of tense and aspect that go beyond surface-level pattern recognition. They can be probed and visualized, showing structural organization of individual categories similar to those humans use to differentiate these verb properties. Values within the same grammatical category (e.g., past, present for tense) form approximately orthogonal directions in latent space. Similarly, broader tense and aspect vectors appear orthogonal to each other, highlighting their representational independence. We find that this encoding of syntactic structure has causal relevance and can be used to steer multi-token generation across different tasks.

Steering works for causal verification but is not a perfect method for model control yet. We use these verb properties to study factors influencing successful steering. While our results show that steering grammatical features can work, there are pitfalls such as topic shift and output degeneration that need to be monitored. Simple metrics such as n-gram statistics, perplexity, POS-tagging and BERTScore help to track side effects, while more expensive methods (e.g., LLM-as-a-judge) can be applied for final evaluations after tuning hyperparameters. Our results suggest that activation norm can be a useful heuristic when adjusting scaling factors, with higher norms requiring stronger steering. Further, the question of where and how long to intervene is task-dependent but can significantly affect success. We find interventions during generation to be generally more effective than steering via updating the prompt representation, particularly in cases where the target property conflicts with the task context. For categorical target values, we find that adding the desired property vector is sufficient, removing the currently present category vector yields no additional benefit due to the approximate orthogonality of the categorical vectors. We encourage future work to more systematically monitor side effects of steering, and to explore automated methods for optimizing steering conditions.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

Limitations

Temporal Expression in Language. Our study focuses on the two verbal properties tense and aspect in English, which has a very regular inflection system, to investigate categorical and combinatorial grammatical structures in language models. While we find evidence that model representations reflect human-like grammar organization, our study has some limitations. First, we restrict our analysis to sentences containing a single unique tense-aspect combination. This is a simplification, as natural language frequently expresses complex temporal relations and event succession through multiple verb phrases with different tense-aspect combinations. Second, as noted by Klein (2009), tense and aspect are only two of six known strategies for expressing temporal information in language. Other mechanisms such as temporal adverbs are used especially in tenseless languages. We leave it to future work to investigate mechanism-independent time representations (e.g., consistent representation of "past" in different surface forms, such as "he was" and "yesterday"), as well as language-independent tense representations (e.g., cross-lingual past-present contrast, see Brinkmann et al. (2025)).

Scope of Steering Analysis. Our analysis focuses on steering grammatical properties in three tasks, which provide a controlled testbed for comparing different types of interventions. However, side effects such as topic shift may manifest differently when steering other concepts like sentiment. Expanding the range of tasks and steering targets could help identify more general conditions for effective steering, which provides a promising direction for future work.

References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Pablo J Diego Simon, Stéphane d'Ascoli, Emmanuel Chemla, Yair Lakretz, and Jean-Rémi King. 2024. A polar coordinate system represents syntax in large language models. *Advances in Neural Information Processing Systems*, 37:105375–105396.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Michael Elhadad. 2010. Book review: Natural language processing with python by steven bird, ewan Klein, and edward loper. *Computational Linguistics*, 36(4).

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy_model/index.html.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics:*

EMNLP 2024, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Sheridan Feucht, David Atkinson, Byron C Wallace, and David Bau. 2024. Token erasure as a footprint of implicit vocabulary items in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9727–9739, Miami, Florida, USA. Association for Computational Linguistics.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Bangkok, Thailand. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Anisia Katinskaia and Roman Yangarber. 2024. Probing the category of verbal aspect in transformer language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3347–3366, Mexico City, Mexico. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Wolfgang Klein. 2009. *How time is encoded*. Mouton de Gruyter.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems*, 31.

Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. Linguistic interpretability of transformer-based language models: a systematic review. *arXiv preprint arXiv:2504.08001*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Language models implement simple Word2Vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Mueller, Jannik Brinkmann, Millicent L. Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. 2024. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *CoRR*, abs/2408.01416.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024a. The geometry of categorical and hierarchical concepts in large language models. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024b. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. 2024. Towards reliable evaluation of behavior steering interventions in LLMs. In *MINT: Foundation Model Interventions*.

James Pustejovsky, Jessica Littman, Roser Saurí, and Marc Verhagen. 2006. Timebank 1.2 documentation. *Event London, no. April*, pages 6–11.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R. Bowman. 2024. Steering without side effects: Improving post-deployment control of language models. In *Neurips Safe Generative AI Workshop 2024*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: A benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. *Preprint*, arXiv:2501.17148.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Implementation Details

A.1 Infrastructure

The experiments involved the 8 billion parameter model Llama-3.1-8B-Instruct and the 7 billion parameter model Qwen-2.5-7B-instruct. They were run on a single server with 8 NVIDIA RTX A6000 48 GB GPUs with CUDA Version 12.4 and an AMD EPYC 7413 24-Core Processor. The total runtime for training the probes and performing the grid search for steering was less than two weeks.

A.2 Libraries

See Table 6.

B Related Work on Grammatical Competence of LMs

Behavioral Evaluations of Grammatical Knowledge. Various benchmarks assess the linguistic knowledge of LLMs. BLiMP (Warstadt et al., 2020) and MultiBLiMP (Jumelet et al., 2025) evaluate syntactic acceptability via paired sentence probabilities, while HOLMES (Waldis et al., 2024) consolidates linguistic probing datasets across a range of syntactic phenomena, including tense classification tasks from Conneau et al. (2018) and Klafka and Ettinger (2020). Additional studies use probing methods to analyze specific grammatical categories, such as structural syntax (Hewitt and Manning, 2019; Diego Simon et al., 2024) and aspect in morphologically rich languages (Katinskaia and Yangarber, 2024). However, these studies typically model tense as a binary past–present distinction and do not address the full range of categorical tense and aspect distinctions.

Causal and Representational Analyses of Syntax in LMs. A complementary line of work investigates how grammatical information is encoded within LMs and how it can be manipulated. Studies have examined internal representations of syntax through attention patterns (Vig and Belinkov, 2019) and circuit-level structures (Wang et al., 2023; Ferrando and Costa-jussà, 2024), revealing how syntactic features are distributed across components of the model. Building on this, causal intervention methods have been used to identify which internal features are functionally relevant to grammatical behavior. CausalGym (Arora et al., 2024) tests whether linear representations influence syntactic decisions, while other work targets tense specifically, manipulating feed-forward layers (Merullo et al., 2024), attention heads (Zhang et al., 2025), or sparse autoencoder features (Brinkmann et al., 2025) to steer generation. These studies, like the behavioral evaluations, focus primarily on binary tense distinctions and are typically limited to single-token evaluations—with the exception of Brinkmann et al. (2025), who consider open-ended generation. Prior work has not jointly analyzed tense and aspect within a unified framework. Our work advances this area by studying categorical features in multi-token generation, combining probing, representation space analysis, and causal steering to examine how these concepts are encoded and can be controlled.

C Tense and Aspect Overview

We provide an overview of all possible tense-aspect combinations in the English language in Table 7.

Usage	Library	Model	Reference
Training Linear Probes	`scikit-learn`		Pedregosa et al. (2011)
Linear Discriminant Analysis	`scikit-learn`		Pedregosa et al. (2011)
POS-Tagging	`stanza`		Qi et al. (2020)
Propbank Annotations	`nltk`		Elhadad (2010)
	`spacy`	*en_core_web_lg*	Honnibal et al. (2020)
BERTScore	`bert_score`	*microsoft/deberta-xlarge-mnli*	Zhang et al. (2020)

Table 6: Libraries used for experiments.

	present	**past**	**future**
simple	She *drives* her car.	She *drove* her car.	She *will drive* her car.
progressive	She *is driving* her car.	She *was driving* her car.	She *will be driving* her car.
perfect	She *has driven* her car.	She *had driven* her car.	She *will have driven* her car.
perfect progressive	She *has been driving* her car.	She *had been driving* her car.	She *will have been driving* her car.

Table 7: Example sentence conjugated across different tense-aspect combinations.

D Dataset Composition

To ensure high-quality tense-aspect annotations, we prioritized careful dataset selection and manual validation. Although existing resources such as TimeML (Pustejovsky et al., 2006) and Universal Dependencies offer valuable linguistic annotations, they were unsuitable due to annotation differences or incomplete coverage of tense-aspect information. Therefore, we curated a dataset with 692 sentences from PropBank (Palmer et al., 2005), each containing exactly one verb, verified through dependency parsing. To improve the overall data coverage for underrepresented tense-aspect combinations, we generated additional synthetic examples using ChatGPT-4o (Hurst et al., 2024), prompted with templates shown in the text box below.

Prompts for Synthetic Data Generation

> Generate 100 diverse sentences in the [PAST] tense. Each sentence should contain only one verb and should vary in structure, subject, and length.

> Generate 100 diverse sentences in [PAST] tense. Each sentence should not contain more than one verb and should vary in structure, subject, and length.

> Generate a list of 100 random sentences that are in active or passive voice, declarative or interrogative, singular or plural. Each sentence should contain only one single verb phrase of the tense "[PAST]".

After downsampling to address any remaining imbalance, the train set has 348 samples per tense and 261 per aspect. The BIG-bench test set (Srivastava et al., 2023; Logeswaran et al., 2018) consists of 90 samples per tense and 70 samples per aspect.

E Additional Probe Results

Besides length-normalized sum pooling, we explore sum pooling, mean pooling, and the final token as alternative extraction methods:

$$h_{\text{sum pooling}} = \sum_{i=1}^{N} h_i, \tag{7}$$

$$h_{\text{mean pooling}} = \frac{1}{N} \sum_{i=1}^{N} h_i, \tag{8}$$

$$h_{\text{final token}} = h_{-1}. \tag{9}$$

We provide layer-wise probing results for all strategies in Fig. 6 and 7. Across models and targets sum pooling and its length normalized version yield the highest f1-scores. Using the final token only is slightly less informative and limited to the earlier middle layers.

F Cluster Quality of LDA Projections

To quantify the separation effectiveness of our LDA directions, we compute

- **Explained Variance:** ratio of between-class variance and total variance
- **Fisher Discriminant Ratio:** between-class vs within-class variance
- **Silhouette score:** similarity of a point to its own class vs others.

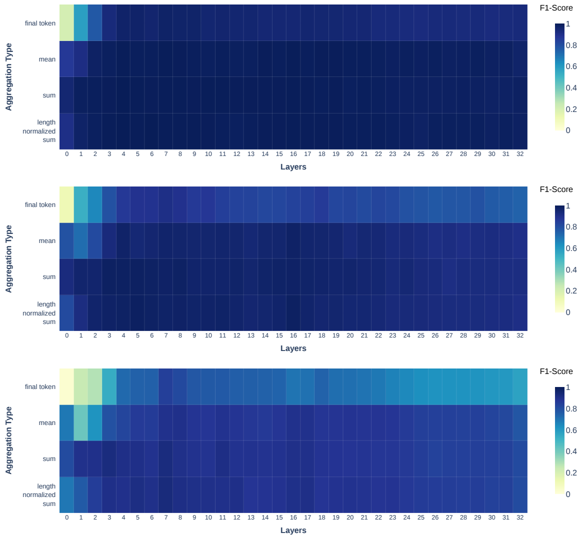

Figure 6: Llama-8B probing f1-scores for **tense**, **aspect** and **tense-aspect** across layers (L0: embedding layer).

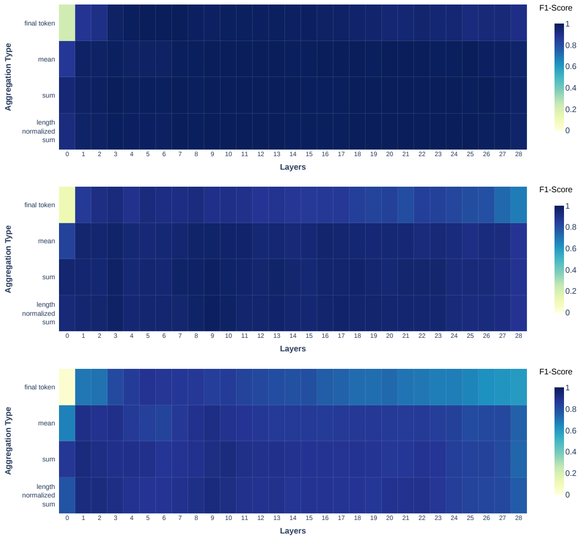

Figure 7: Qwen-7B probing f1-scores for **tense**, **aspect** and **tense-aspect** across layers.

for the L0 projected hidden states, see Table 8. Further, we measure the orthogonality of the parent feature directions $\bar{\ell}_{\text{TENSE}}$ and $\bar{\ell}_{\text{ASPECT}}$, as cosine similarity between all possible vector differences for tense and aspect, considering the layers with the best steering efficacy across tasks and methods. This results in a mean similarity of 0.045 for Llama-8B and 0.118 for Qwen-7B. Cosine values this small imply that the two contrast directions are almost orthogonal, so the models largely encode tense and aspect in independent subspaces.

G Tasks for Multi-Token Steering

We consider three generative tasks to evaluate steering, their prompt formats are detailed below.

Random Sentence Task

Using the template "`<Imperative Verb> <Sentence Description>:`", we form a test set of $N = 83$ distinct prompts.

Imperative Verbs
Generate, Create, Produce, Write, Output, Provide, Construct, Make up, Formulate, Come up, Print, Return, Craft

Sentence Descriptions
a single sentence, one sentence, a random sentence, a sentence using any verb tense, an arbitrary sentence, one grammatically correct sentence

For repetition and temporal translation, we only include samples where the unsteered output is a valid answer.

Few-Shot Tasks

We create a prompt for each sentence in our test set and use other sentences from the test set as few-shot examples. For each steering target, we exclude those samples, where the source feature value is equal to the steering target, resulting in a test set size of $N = 211$ for aspect and $N = 191$ for tense.

Repetition Task
I am writing a story. \\ I am writing a story.

I have finished. \\ I have finished.

The dog is barking. \\

Temporal translation Task
I have been walking through the park. \\ I have walked through the park.

Paul has been visiting the school. \\ Paul has visited the school.

He has been earning a six figure salary. \\

	QWEN-7B		LLAMA-8B	
	Tense	**Aspect**	**Tense**	**Aspect**
Explained Variance (↑)	0.72	0.70	0.63	0.62
Fisher Discriminant Ratio ↑)	2.44	2.46	1.53	1.78
Silhouette score (↑)	0.39	0.24	0.27	0.22

Table 8: Cluster quality scores for the L0 projected hidden states indicate that at least 70% of total variance is explained through the retrieved tense and aspect classes for Qwen-7B and at least 62% for Llama-8B. The Fisher ratios demonstrate that between-class variance exceeds within-class variance by 1.5 - 2.5, confirming that our LDA directions successfully capture the target linguistic distinctions. While these scores indicate moderate rather than perfect separation, this is reasonable since tense and aspect categories can have fuzzy boundaries (e.g., "Tomorrow I leave for Paris." uses present form but future reference).

H Grid search over Steering Factor

We use the random sentence task to perform the initial grid search across the α values listed in Table 9, and use the findings to adjust the search space for the few-shot tasks accordingly. The best steering configurations for each model, task, method and target are listed in Table 10.

I Measuring Degenerates

To measure the rate of degenerate outputs during steering experiments, we track different n-gram statistics. We label an output as degenerate, if it does not pass all the filters in Table 11 and/or does not contain a verb phrase (i.e., AUX / VERB), as detected by stanza's POS-tagger. N-gram diversity is computed as the product of one minus the repetition rates of 2-, 3-, and 4-grams in the text. These metrics to measure text diversity are based on Li et al. (2023).

J Additional Steering Results

J.1 Selectivity.

The most effective steering methods also show the highest functional selectivity (Figure 8), indicating that it is possible to steer one verb property—such as tense—without necessarily affecting the other, like aspect. However, selectivity remains below 50% on average. Despite the orthogonal representations of tense and aspect, steering one target often influences both. This suggests that the intervention may still be too large, modifying more of the activation than intended.

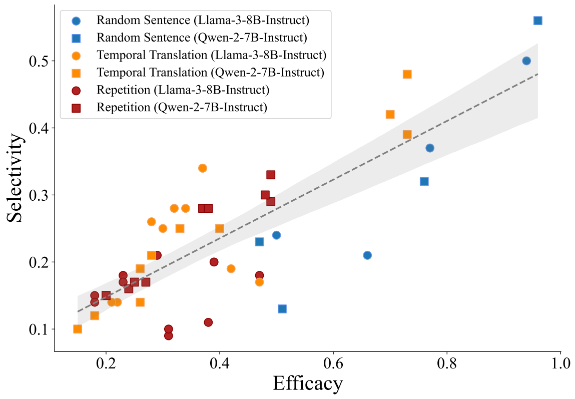

Figure 8: Correlation between efficacy and selectivity is apparent.

J.2 Relative Perplexity Change

In addition to the degenerate rate which is implicitly measured through efficacy, we report the relative change in perplexity between steered and unsteered generation. Results for the best setups per model, task, target and method are reported in Figure 9. For both models, the majority of setups (8/14 for Llama-8B and 11/14 for Qwen-7B) leads to a minor increase in perplexity with a relative change of < 10.

J.3 Activation Norm

There are differences in activation norms across models as well as across layers of the same model that affect the required steering strength for successful interventions. For an overview of average activation norms and feature projection magnitudes per model, see Figure 10. The results of a comprehensive grid search across layers and different steering factors is visualized in Figure 11.

J.4 Nucleus Sampling

Due to the increased risk of degeneration introduced through steering (Stickland et al., 2024), we additionally evaluate model behavior under stochastic decoding using nucleus sampling with a temperature of 0.7, but find the effects on efficacy and

Task	**Model**	**α values**
Random Sentence	both	0.1, 0.5, 1, 1.5, 2, 3, 4, 5, 7, 10, 15, 20, 25, 30, 50
Random Sentence	Qwen-2-7B-Instruct	100, 150, 200, 250, 300, 400, 500
Few-Shot	Llama-3-8B-Instruct	5, 7, 10, 15, 20, 25, 30, 35, 40
Few-Shot	Qwen-2-7B-Instruct	200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 550, 700, 800

Table 9: α values searched across in grid search for steering experiments.

Model	**Task**	**Method**	**Target**	**Layer**	**Alpha**
Llama-3-8B-Instruct	Random	TA	Tense	13	5
Llama-3-8B-Instruct	Repetition	TA	Tense	11	15
Llama-3-8B-Instruct	Repetition	TA+SS	Tense	12	10
Llama-3-8B-Instruct	Repetition	TA+Proj-SS	Tense	11	15
Llama-3-8B-Instruct	Temporal Translation	TA	Tense	13	5
Llama-3-8B-Instruct	Temporal Translation	TA+SS	Tense	12	7
Llama-3-8B-Instruct	Temporal Translation	TA	Tense	12	10
Llama-3-8B-Instruct	Random	TA	Aspect	19	15
Llama-3-8B-Instruct	Repetition	TA	Aspect	14	15
Llama-3-8B-Instruct	Repetition	TA+SS	Aspect	19	15
Llama-3-8B-Instruct	Repetition	TA+Proj-SS	Aspect	14	15
Llama-3-8B-Instruct	Temporal Translation	TA	Aspect	18	20
Llama-3-8B-Instruct	Temporal Translation	TA+SS	Aspect	16	15
Llama-3-8B-Instruct	Temporal Translation	TA	Aspect	18	25
Qwen-2-7B-Instruct	Random	TA	Tense	20	100
Qwen-2-7B-Instruct	Repetition	TA	Tense	22	200
Qwen-2-7B-Instruct	Repetition	TA+SS	Tense	24	200
Qwen-2-7B-Instruct	Repetition	TA+Proj-SS	Tense	22	225
Qwen-2-7B-Instruct	Temporal Translation	TA	Tense	22	200
Qwen-2-7B-Instruct	Temporal Translation	TA+SS	Tense	24	200
Qwen-2-7B-Instruct	Temporal Translation	TA	Tense	22	200
Qwen-2-7B-Instruct	Random	TA	Aspect	21	150
Qwen-2-7B-Instruct	Repetition	TA	Aspect	21	200
Qwen-2-7B-Instruct	Repetition	TA+SS	Aspect	23	200
Qwen-2-7B-Instruct	Repetition	TA+Proj-SS	Aspect	21	200
Qwen-2-7B-Instruct	Temporal Translation	TA	Aspect	22	225
Qwen-2-7B-Instruct	Temporal Translation	TA+SS	Aspect	24	200
Qwen-2-7B-Instruct	Temporal Translation	TA	Aspect	22	225

Table 10: Best steering configuration with regard to efficacy.

Filter	**Threshold**
Unigram Repetition Rate	< 0.25
2-gram Repetition Rate	< 0.3
4-gram Repetition Rate	< 0.2
N-gram Repetition Diversity	> 0.5

Table 11: Thresholds for degeneration-filter.

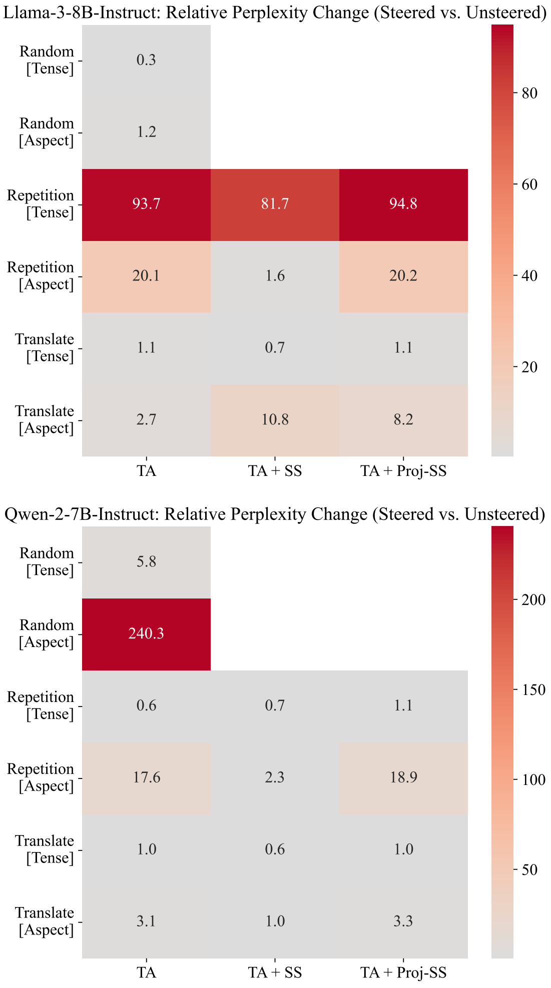

Figure 9: Relative change in perplexity for both models. Overall, the different steering methods lead to minor perplexity increases across tasks, targets and models. However, steering the repetition task appears to be less stable, with outliers across all methods. Similarly, generation quality collapses for steering aspect during random generation of Qwen-7B. This shows that tasks with different internal mechanisms react differently when steered with the same feature vectors and highlights the need for diverse testbeds when developing steering approaches.

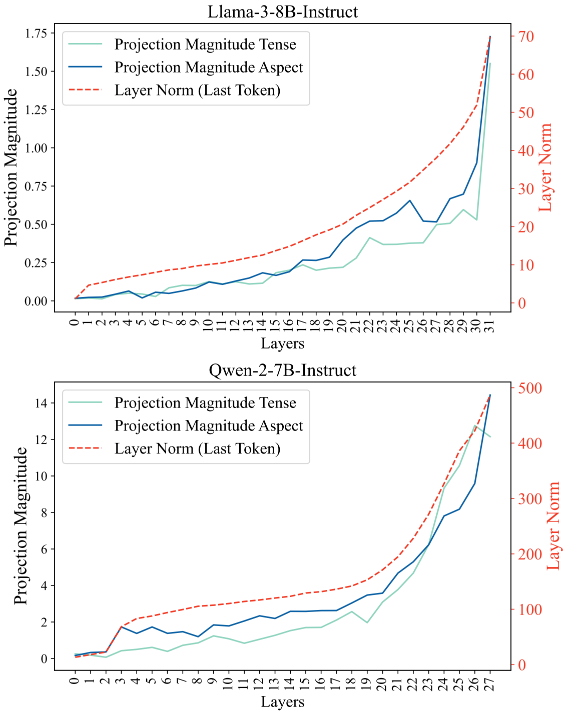

Figure 10: Average activation norm on final token compared to projection magnitude of both features, tense and aspect, averaged across tasks. The graph shows an increase of all three across layers, indicating that the strength of a feature signal roughly scales with the general activation norm.

degenerates to be not consistent, suggesting that the decoding strategy alone cannot prevent degeneration caused by steering.

J.5 Qualitative Results for Temporal Translation

Example outputs for temporal translation task are provided in Table 12, demonstrating successful steering of aspect, while keeping tense and the general topic of the sentence fixed. Ablations results for position-wise steering are given in Table 13.

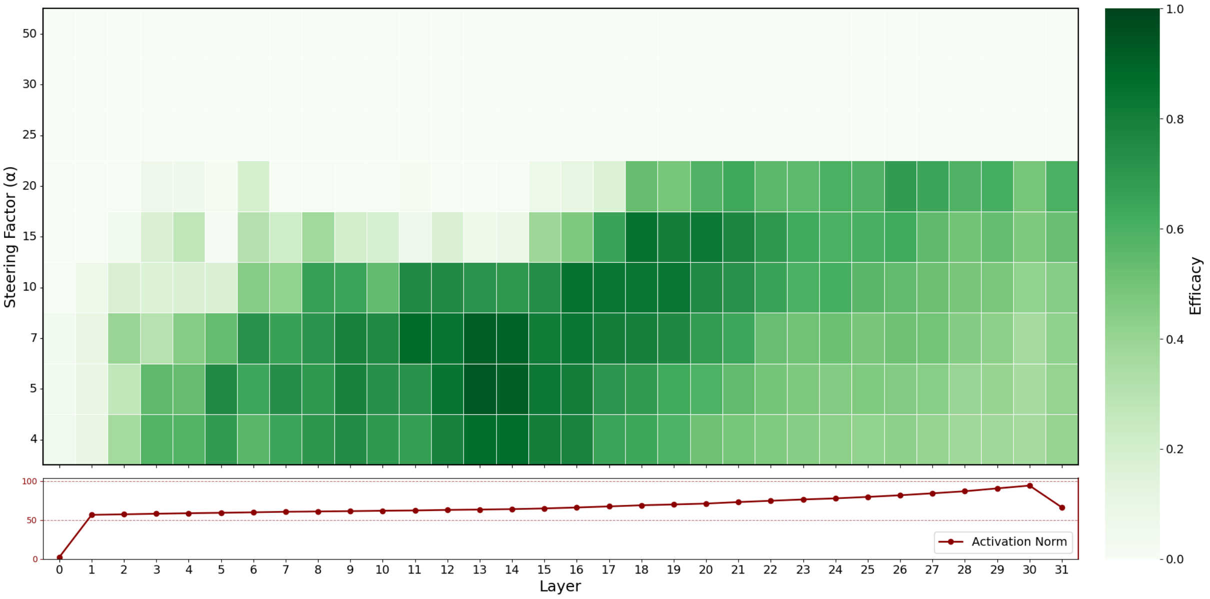

Figure 11: Grid search over best layer and alpha for steering success in Llama-8B. Heatmap displays efficacy with darker colors indicating higher steering success. Bottom plot indicates the average norm of activations per layer.

Temporal Translation	Prompt	Target	Output
Perfect Progressive → Perfect	I have been walking through the park. \\ I have walked through the park. Paul has been visiting the school. \\ Paul has visited the school. He has been earning a six figure salary. \\	Progressive	**Unsteered:** He has earned a six figure salary. **Steered:** He is earning a six figure salary.
Perfect → Progressive	I have walked through the park. \\ I am walking through the park. Paul has visited the school. \\ Paul is visiting the school. He will not have passed the test. \\	Simple	**Unsteered:** He will not be passing the test. **Steered:** He will not pass the test.

Table 12: Qualitative examples of steering aspect in Qwen-2-7B-Instruct on the temporal transformation task. Red indicates the aspect of the original sentence, orange the aspect in the unsteered translation and blue marks the aspect that is expected after steering.

Steering Position	Prompt Tokens	Generated Tokens	Output (TA)	Output (TA-SS)
all verb tokens in prompt	...It was snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
last verb token in prompt	...It was snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
sentence end in prompt	...It was snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
final token in prompt	...It was snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
final tokens during generation	...It was snow ing . \\	It is snow ing .	It was a year...	It was raining.
generated token before verb	...It was snow ing . \\	It is snow ing .	It was snowing.	It was snowing.
first generated verb token	...It was snow ing . \\	It is snow ing .	It is snowing.	It is snowing.
all generated verb token	...It was snow ing . \\	It is snow ing .	It is snowing.	It is snowing.

Table 13: Steering Llama-3-8B-Instruct on the temporal translation task towards past tense for the prompt: "He was crying. \\ He is crying.\n\n We were having dinner. \\ We are having dinner. \n\n It was snowing. \\". *Generated Tokens* refers to the unsteered output.