CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) achieve remarkable performance through pretraining on extensive data. This enables efficient adaptation to diverse downstream tasks. However, the lack of interpretability in their underlying mechanisms limits the ability to effectively steer LLMs for specific applications. In this work, we investigate the intrinsic mechanisms of LLMs from a cognitive perspective using eye movement measures. Specifically, we analyze the layer-wise correlation between human cognitive indicators and LLM representations. Building on these insights, we propose a heuristic approach for selecting the optimal steering layer to modulate LLM semantics. To this end, we introduce an efficient selective layer intervention based on prominent *parameter-efficient fine-tuning* methods, which conventionally adjust either all layers or only the final layer. Additionally, we present an *implicit layer contrastive intervention* during inference to steer LLMs away from toxic outputs. Extensive experiments on natural language understanding, reasoning, and generation tasks, conducted on GPT-2, LLaMa2-7B, and Mixtral-7B, demonstrate the effectiveness and efficiency of our approach. As a model-agnostic framework, it enhances the interpretability of LLMs while improving efficiency for safe deployment.

1 Introduction

Large Language Models (LLMs) (Dubey et al., 2024; Yang et al., 2024; Guo et al., 2025) have demonstrated strong capabilities in natural language understanding and reasoning (Wei et al., 2022; Zhao et al., 2023) through pretraining on large datasets, followed by instruction tuning and alignment with human values (Wei et al., 2021; Ouyang et al., 2022). Consequently, LLMs achieve excellent performance on downstream tasks with fine-tuning. However, their lack of interpretability and transparency limits the development of efficient fine-tuning and inference methods.

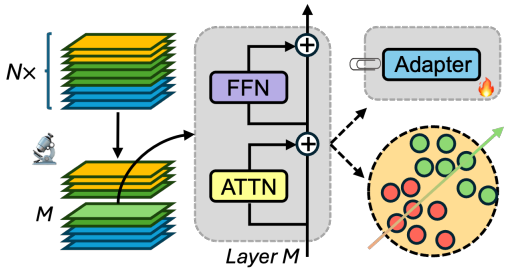

Figure 1: **Demonstration of CogSteer Intervention**. For an N-layer LLM, we first heuristically find the optimal layer M for semantic intervention. The upper block represents an adapter that is fine-tuned and inserted into the frozen layer M. The bottom block illustrates the operation of the attention module in M to steer the semantic direction towards safer outputs during inference.

To understand intrinsic mechanisms of LLMs, previous work has introduced various interpretability methods, including training linear classifiers as probes on top of hidden representations (Belinkov, 2022), projecting representations into vocabularies (Geva et al., 2022), and intervening in the computation path, such as knowledge neurons (Dai et al., 2022) and circuits (Conmy et al., 2023; Ghandeharioun et al., 2024). However, these methods, which focus on a limited set of predefined classes, concepts, or prompts, have practical limitations in terms of scalability and generalization.

In this work, we introduce a novel interpretability analysis method that leverages eye movement data (Luke and Christianson, 2018; Hollenstein et al., 2020; Colman et al., 2022) collected by cognitive researchers to study human reading behaviors. Through correlation experiments, we find that LLM hidden states exhibit a strong correlation with human gaze, peaking in the middle layers. Using eye movement measures such as *fixation* and *regression* (Rayner, 1998) as human-interpretable indicators, we observe a hierarchical progression in LLMs, from initial syntactic and semantic processing to deeper integration and final prediction. Addition-

ally, a comparison of correlation results between natural reading and task-specific reading suggests that the upper layers of LLMs are more capable of reasoning. Furthermore, advanced LLMs such as LLaMa, which incorporate instruction tuning and reinforcement learning from human feedback (RLHF), demonstrate enhanced reasoning capabilities, even in the middle layers.

LLM intervention (Poth et al., 2023; Dong et al., 2024) refers to fine-tuning or applying inference methods to steer their semantics to align with specific tasks and data distributions. In addition to enhancing the understanding of LLM behavior, our interpretability analysis reveals that different layers serve distinct functions, with middle layers playing a crucial role in deeper syntactic and semantic processing. This enables us to first identify the most suitable layer for intervention, thereby improving task-specific performance. Based on these insights, we propose a heuristic approach for selecting the optimal steering layer for semantic intervention.

To achieve this, we refine prominent PEFT methods, which traditionally adjust either all layers or only the last layer. As shown in Figure 1, our proposed *CogSteer* framework first identifies the most suitable layer M for semantic intervention based on the task. Instead of fine-tuning all layers or only the last layer, our method fine-tunes only the selected layer, enabling LLMs to better adapt to specific tasks and datasets. The number of learnable parameters in *CogSteer* is significantly reduced, requiring only $1/N$ of the parameters in LLMs, thereby improving parameter efficiency. Furthermore, we propose an *implicit layer contrastive intervention* method during inference, which efficiently identifies and steers semantics toward safer generation directions to evaluate the effectiveness of our proposed selective layer intervention.

Through extensive evaluations across diverse tasks and datasets, we demonstrate that the proposed selective layer intervention method achieves comparable or even superior performance with fewer parameters compared to the full-layer intervention baseline. Specifically, we observed an average absolute improvement of **+1.7** on the GLUE benchmark for LLaMa2-7B, and an average absolute improvement of **+5.8** on the GLUE benchmark for Mistral-7B, with only **3.1%** of the parameters involved in full-layer fine-tuning. Moreover, in experiments on generation tasks, language toxification (Dementieva et al., 2025), and detoxification (Leong et al., 2023), our method achieves a **+1.85%** improvement in toxification compared to full-layer intervention and a **+13.45%** improvement in detoxification as compared to last-layer intervention.

Our **main contributions** are as follows:

(1) We are the first to propose leveraging eye movement measures to analyze LLM behavior. We publicly release the probing code to facilitate further research on interpretability from a cognitive perspective. (2) Through correlation analysis, we demonstrate a hierarchical progression in LLMs and introduce a heuristic steering layer selection method for efficient layer intervention. (3) Extensive experiments validate the effectiveness of our proposed method across various language understanding, reasoning, and generation tasks, contributing to the development of efficient and explainable foundation models.

2 Related Work

Interpretability research is essential for uncovering the mechanisms of LLMs and ensuring their safe and trustworthy deployment as foundation models. Various studies have analyzed how knowledge is stored in LLMs (Goldowsky-Dill et al., 2023; Stolfo et al., 2023; Rai et al., 2024; Bills et al., 2023; Geva et al., 2022), focusing on concepts such as knowledge neurons (Dai et al., 2022) and circuits (Conmy et al., 2023; Yao et al., 2024). Moreover, Wang et al. (2024a) reveal that GPT-2 predicts tokens more similarly to humans than shallow language models but lacks engagement with LLMs. Unlike previous works, our method investigates the interpretability of LLMs through human-interpretable indicators based on eye movement theory, enabling a more precise understanding and control of model behavior (see § 3 and 4.1).

Parameter-Efficient Semantic Steering PEFT (Han et al., 2024; Wang et al., 2023), including Adapter (Houlsby et al., 2019; Poth et al., 2023) and LoRA (Hu et al., 2022), has gained popularity due to its ability to maintain a large number of frozen parameters in LLMs for generality while introducing only a small number of trainable parameters per task. Our method for semantic steering via fine-tuning enhances PEFT methods by incorporating a selective layer intervention strategy, reducing computational costs while achieving superior performance (see § 4.2). Furthermore, steering semantics during inference offers an even more efficient approach. Contrastive decoding (Li et al., 2023; Sennrich et al., 2024; Wang et al., 2024b) guides

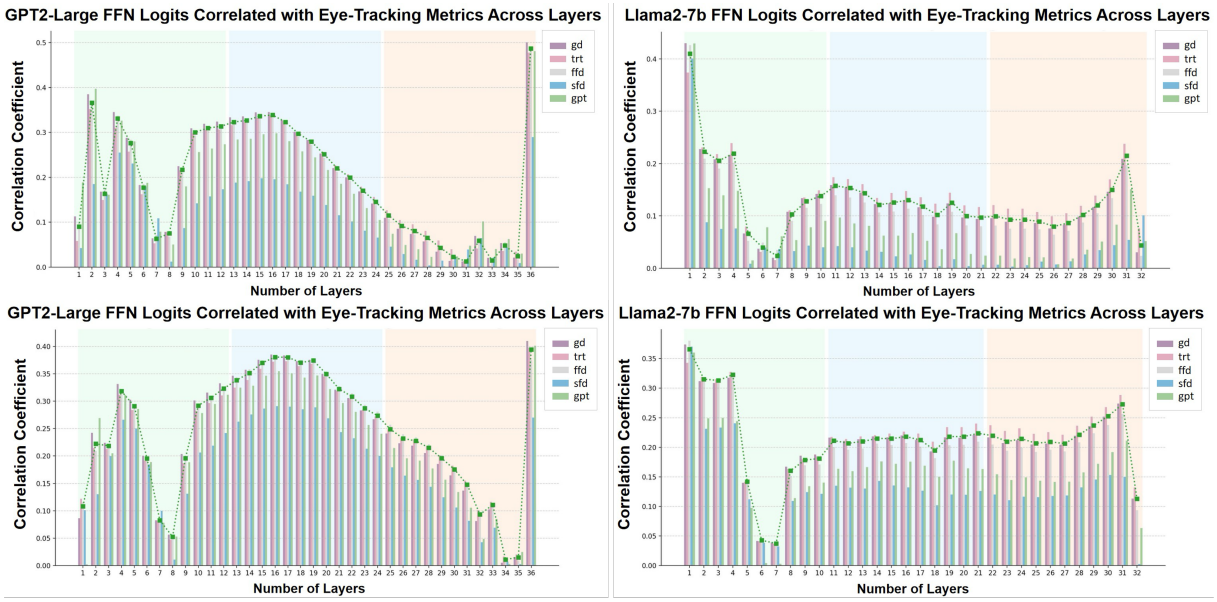

Figure 2: **Correlation Results Comparison:** ***Natural Reading (NR) vs. Task-Specific Reading (TSR).*** Correlation results are shown for GPT-2 Large NR (36 layers, top-left), GPT-2 Large TSR (36 layers, bottom-left), LLaMa2-7B NR (32 layers, top-right), and LLaMa2-7B TSR (36 layers, bottom-right). Green, blue, and orange boxes indicate the premature, middle, and mature buckets, respectively.

the generation process by comparing two output distributions. In contrast, our proposed implicit layer contrastive intervention efficiently identifies and steers the semantics of LLMs toward safe directions during inference (see § 4.3).

3 From Human Gaze to LLM Behavior

Recent studies (Geva et al., 2021; Schuster et al., 2022) suggest that feed-forward networks (FFNs) function similarly to neural memory networks, capturing syntactic and semantic features as well as factual knowledge (Chuang et al., 2023; Zhang et al., 2024). Meanwhile, research in cognitive science (Rayner, 1998) has shown that eye movement measures provide insights into the time required for human readers to process syntax, semantics, and integrate information. Motivated by these findings, we leverage eye movement measures to analyze their correlation with the hidden states of FFNs across different layers of LLMs.

Models. Correlation studies are conducted on the GPT-2 model (Radford et al., 2019) at different sizes (12-layer small, 24-layer medium, 36-layer large) and the 32-layer LLaMa2-7B (Touvron et al., 2023). Both GPT-2 and LLaMa2-7B use a decoder-only transformer architecture. Comparing earlier LLMs, such as GPT-2, with more advanced models like LLaMa2 provides valuable insights into their similarities and differences while also enhancing the robustness and generalizability of our findings.

Eye-movement Experiments and Datasets. We conduct a correlation analysis under two experimental conditions: natural reading and task-specific reading. For natural reading, we use the Provo (Luke and Christianson, 2018), GECO (Colman et al., 2022), and ZuCo 2.0 (Hollenstein et al., 2020) datasets. The ZuCo 2.0 dataset also includes task-specific reading experiments. In task-specific reading, participants are required to determine whether a specific relation type is present in a sentence. Relation detection is a high-level semantic and reasoning task that involves complex cognitive processing.

Correlation Analysis. Let S_j denote the j-th sentence, consisting of n_j words $w_1, w_2, \ldots, w_{n_j}$. For each word w_i in sentence S_j, we consider five eye movement measures: $e_i^{(k)}, k \in \{std, ffd, gd, trt, gpt\}$, where each measure represents a scalar value. The hidden state at layer l of the LLM for word w_i is denoted as $\mathbf{h}_{l,i} \in \mathbb{R}^d$, where d is the dimensionality of the hidden states.

To analyze the relationship between eye movement measures and LLM hidden states, we compute the *Pearson correlation* between each eye movement measure and the corresponding hidden states at each layer. Specifically, we concatenate hidden states and eye movement measures across all words in the dataset and apply *principal component analysis* to obtain a scalar representation of hidden states, ensuring alignment with eye move-

ment measures (see Appendix D for details). The correlation is then defined as:

$$\rho_{l,k} = \frac{\sum_{i=1}^{n_{total}} (h_{l,i} - \bar{h}_l)(e_i^{(k)} - \bar{e}^{(k)})}{\sqrt{\sum_{i=1}^{n_{total}} (h_{l,i} - \bar{h}_l)^2} \sqrt{\sum_{i=1}^{n_{total}} (e_i^{(k)} - \bar{e}^{(k)})^2}}, \quad (1)$$

where $h_{l,i}$ represents the processed hidden states aligned with the eye movement measures, and $\bar{h}_l$ and $\bar{e}^{(k)}$ are their respective means.

Results and Finding 1. Based on the correlation calculation in Eq. 1, we first analyze the correlation results for the natural reading task. To facilitate interpretation, we divide the layers of both models into three equal groups: *premature*, *middle*, and *mature*. Figures 2 (*upper*) and 7 present the correlation results between various eye movement measures and the hidden states of the LLMs, illustrating how these values evolve across layers.

The results indicate that the hidden states of different LLM layers exhibit a clear and strong correlation with human gaze, peaking in the middle bucket and reaching a secondary peak in the mature bucket. This trend is consistent across different eye movement measures and LLMs with varying layer sizes. Considering the nature of these eye movement measures, the increase in correlation in the *premature bucket* suggests that LLMs begin processing tokens by integrating syntactic and semantic features, reflecting an initial focus on token processing. In the *middle bucket*, the further increase in correlation signifies deeper syntactic and semantic processing, with the peak indicating the integration of linguistic features. In the *mature bucket*, the secondary peak likely reflects the final integration of information for word prediction.

The overall trend across the three buckets is similar for GPT-2 and LLaMa2-7B. For a detailed analysis of eye movement measures and additional findings, please refer to Appendix D.

Finding 1. (Layer-wise functionality)

LLM hidden states exhibit a strong correlation with human gaze, peaking in the middle bucket and again in the mature bucket. This pattern suggests a hierarchical progression from initial syntactic and semantic processing to deeper integration and final prediction.

Results and Finding 2. Empirically, LLMs are trained with the next-token prediction objective. Modern LLMs demonstrate strong language understanding and reasoning abilities, raising the question: *Are LLMs merely next-token predictors, or are they task reasoners?* Figure 2 presents a comparison between natural reading and task-specific reading. The correlation patterns suggest that LLMs function as both next-token predictors and reasoners, as the trends in task-specific reading closely resemble those observed in natural reading.

Notably, for both GPT-2 and LLaMa2-7B models, the correlation values in the *middle bucket* during task-specific reading are higher than those observed during natural reading. In particular, for the LLaMa2-7B model, these values remain consistently higher in task-specific reading. We hypothesize that this indicates LLaMa2-7B is better suited for reasoning tasks and that the layers in the *middle bucket* and *mature bucket* are activated when processing complex tasks, as it is trained on a larger text corpus and incorporates more advanced post-training techniques.

Finding 2. (LLM functions as both next-token predictor and task reasoner)

LLMs function as both next-token predictors and reasoners, with overall correlation trends aligning with human cognition indicators. Advanced training methods enhance their ability to reason and handle complex tasks.

4 Method

4.1 Heuristic Steering Layer Selection

A better understanding of LLM mechanisms will help in precisely and efficiently controlling their behaviors, particularly for semantic steering. We argue that the predominant parameter-efficient fine-tuning (PEFT) methods (Han et al., 2024), which by default intervene in the last layer or across all layers, are not optimal[1]. Instead, we propose an efficient heuristic steering layer selection strategy for intervention, based on our cognition-inspired interpretability analysis detailed in Section 3.

For semantic steering in LLMs, the layers in the middle bucket are the most suitable candidates for intervention. These layers handle further token processing, information integration, and preliminary reasoning. Additionally, the residual connections (He et al., 2016) in transformer layers allow the

[1]Recent work (Yu et al., 2023) on fine-tuning speech translation models also supports our hypothesis.

semantic intervention to flow and evolve gradually, avoiding abrupt changes in the final prediction (Chuang et al., 2023).

From a task-oriented perspective, we apply PEFT methods and inference-only methods to the candidate layers in the middle bucket, using a small portion of data, like the validation set, to search for and select the best-performing layer that suits the task scenario. Formally, given M as the best layer for intervention, J represents a set of candidate layers in the middle bucket $\left(\frac{N}{3} \leq M \leq \frac{2N}{3}\right)$. D denotes the validation set. We search for the layer M that yields the best task score or loss performance, as follows:

$$M = \arg\max_{j \in J} S\left(D; P\left(\cdot \mid x_t\right)\right). \tag{2}$$

Later, we will demonstrate the effectiveness of the heuristic steering layer selection approach in language understanding, reasoning, and generation tasks, as discussed in Section 5.

4.2 Layer Intervention via Fine-tuning

PEFT approaches, such as additive fine-tuning (i.e., adapters) and reparameterized fine-tuning (i.e., LoRA), are among the most popular due to their efficiency, as they require only a small set of new parameters for task-specific fine-tuning. Let the parameters of an LLM consist of a set of pre-trained, frozen parameters $\phi(\cdot)$ and a set of newly introduced parameters in the inserted block $\psi(\cdot)$. Our layer intervention via fine-tuning to steer semantics in LLM and predicts the next token as follows:

$$y(x_t) = \text{softmax}\big(\text{logit}_{\phi,\psi}\big(FFN_{\phi}^{N}(x_t) \mid F_{\phi,\psi}^{M}(x_t), y_{<t}\big)\big). \tag{3}$$

Here, $FFN_{\phi}^{N}(x_t)$ represents the hidden state of the FFN in the final layer with frozen parameters $\phi(\cdot)$, used for token prediction over a vocabulary. M denotes the best layer for semantic intervention, determined by Equation 2. $F_{\phi,\psi}^{M}(x_t)$ indicates that the fusion layer in the LLM integrates new parameters $\psi(\cdot)$ from the newly added block, aligned with the frozen parameters $\phi(\cdot)$.

Our cognitive-inspired selective layer intervention method is an adaptive fine-tuning strategy that identifies the best layer for both effective semantic steering and task performance. Moreover, as our method only operates on a single layer rather than all layers, it significantly reduces computational resources and time, while also avoiding catastrophic forgetting (Luo et al., 2023; Li et al., 2024).

4.3 Layer Intervention during Inference

Efficient semantic steering can be achieved via fine-tuning. However, an even more efficient approach is to steer the semantics of LLMs during inference without introducing additional parameters. Motivated by Li et al. (2023); Leong et al. (2023), which contrast outputs from either a less capable model or outputs induced by a negative prompt, we propose an *implicit layer contrastive intervention* method during inference. First, we fine-tune a contrast model that generates either the desired output or the output we aim to mitigate. In our case, to mitigate toxic token generation, we fine-tune a toxic LLM as the contrast model. Unlike Li et al. (2023); Leong et al. (2023), which contrast the outputs explicitly in the last layer, our method operates on the contextualized value vectors derived from the weight matrices K, Q, V of the attention modules within LLMs. We perform this operation on the best layer for intervention as described in Equation 2. Formally, our layer intervention during inference finds the semantic steering direction by contrasting the value vectors as follows:

$$\Delta v^{M} = v_{c}^{M} - v_{o}^{M}, \tag{4}$$

where v_{c}^{M} and v_{o}^{M} are the contextualized value vectors of the contrast LLM and the original LLM at the best layer M for semantic intervention (see Elhage et al. (2021) for mathematical derivation). We then update the value vector in the layer M of the original LLM:

$$v'^{M} = v_{o}^{M} - \lambda_{\text{norm}}^{\alpha} \cdot \Delta v^{M}. \tag{5}$$

Here, $\lambda_{\text{norm}} = 1 + \left\|\Delta v^{M}\right\|_2$ is a normalization term that adaptively regulates the steering effect, and α is a hyperparameter that further controls the steering strength. Finally, we preserve the updated steering direction and renormalize the adapted value vector to ensure its representation is close to the original vector:

$$v'^{M} = v'^{M} \cdot \frac{\left\|v_{o}^{M}\right\|_2}{\left\|v'^{M}\right\|_2}. \tag{6}$$

5 Experiments

5.1 Datasets and Evaluation

Datasets We evaluate our proposed efficient semantic steering methods using the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019), applying selective layer

Model	VAL-SET								
	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	WNLI	Avg.
GPT2-L	78.0 $_{l\text{-}19}$	79.5 $_{l\text{-}19}$	85.5 $_{l\text{-}20}$	83.3 $_{l\text{-}19}$	80.6 $_{l\text{-}19}$	71.1 $_{l\text{-}19}$	92.9 $_{l\text{-}19}$	53.5 $_{l\text{-}19}$	78.1
	82.1	83.5	83.1	85.2	82.6	70.4	93.6	53.5	79.2
Llama2-7B	86.4 $_{l\text{-}14}$	87.1 $_{l\text{-}14}$	86.5 $_{l\text{-}14}$	89.3 $_{l\text{-}14}$	83.3 $_{l\text{-}14}$	75.5 $_{l\text{-}14}$	95.8 $_{l\text{-}14}$	56.4 $_{l\text{-}19}$	82.5
	89.0	89.3	86.3	91.9	85.6	65.3	96.7	56.3	82.5
Mistral-7B	87.3 $_{l\text{-}12}$	88.1 $_{l\text{-}12}$	86.9 $_{l\text{-}12}$	91.4 $_{l\text{-}14}$	84.6 $_{l\text{-}12}$	80.1 $_{l\text{-}12}$	95.8 $_{l\text{-}14}$	56.3 $_{l\text{-}12}$	83.8
	89.5	89.7	82.2	81.7	78.1	58.9	96.7	56.3	79.1
Model	**TEST-SET**								
	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	WNLI	Avg.
GPT2-L	79.3 $_{l\text{-}19}$	79.3 $_{l\text{-}19}$	83.0 $_{l\text{-}20}$	84.1 $_{l\text{-}19}$	65.6 $_{l\text{-}19}$	64.6 $_{l\text{-}19}$	92.4 $_{l\text{-}19}$	58.9 $_{l\text{-}19}$	75.8
	82.6	83.0	82.7	85.6	65.6	62.6	93.5	61.6	77.1
Llama2-7B	82.9 $_{l\text{-}14}$	86.3 $_{l\text{-}14}$	83.4 $_{l\text{-}14}$	88.5 $_{l\text{-}14}$	68.8 $_{l\text{-}14}$	74.7 $_{l\text{-}14}$	95.2 $_{l\text{-}14}$	64.4 $_{l\text{-}19}$	80.5
	89.5	88.8	80.5	92.1	71.6	58.2	93.5	55.5	78.7
Mistral-7B	87.1 $_{l\text{-}12}$	87.5 $_{l\text{-}12}$	86.6 $_{l\text{-}12}$	91.7 $_{l\text{-}14}$	70.5 $_{l\text{-}12}$	81.0 $_{l\text{-}12}$	95.9 $_{l\text{-}14}$	65.8 $_{l\text{-}12}$	83.2
	89.7	89.4	80.4	81.3	62.7	52.3	97.3	65.1	77.3

Table 1: **Evaluation on GLUE Benchmark.** MRPC and QQP are reported using F1, while the other tasks are reported using Accuracy. A green box indicates that the single-layer intervention outperforms the full-layer intervention, an orange box denotes comparable performance, and a blue box indicates slightly lower performance.

intervention. Specifically, we focus on eight GLUE tasks covering *sentiment analysis* (SST-2), *paraphrase identification* (MRPC, QQP), and *natural language inference* (MNLI-M, MNLI-MM, QNLI, RTE, WNLI). Additionally, to assess our methods in the context of language generation, we examine their effectiveness in natural language toxification and detoxification. To train toxic adapters and contrast models, as described in Sections 4.2 and 4.3, we utilize the Toxic Comment Classification Challenge Dataset (Jigsaw, 2018). (See Appendix A)

Evaluation For the evaluation on the GLUE benchmark, we report both validation-set and test-set *F1 scores* for QQP and MRPC, while *accuracy* is used for all other tasks. Additionally, we use the RealToxicityPrompts (RTP) dataset (Gehman et al., 2020). Following prior work (Li et al., 2023; Leong et al., 2023), we sample 2,122 toxic prompts. For each toxic prompt in the RTP dataset, we generate 25 continuations and evaluate their toxicity using the Perspective API, which assigns a toxicity score to each continuation, with higher scores indicating a greater likelihood of toxicity. Finally, we use the *average maximum toxicity* as our evaluation metric. Specifically, we compare the toxicity scores and detoxification margins obtained by applying our methods to different layers of the models.

5.2 Models and Baselines

We evaluate our efficient semantic steering methods using three sizes of GPT-2 models, as discussed in § 3, along with LLaMa2-7B. We select GPT-2 and LLaMa2-7B because the former represents earlier classical LLMs, while the latter exemplifies modern LLMs. To ensure generalizability, we also evaluate the Mixtral-7B (Jiang et al., 2024) model to assess performance across tasks. Since applying semantic intervention to either the last layer or all layers of LLMs is a conventional approach, we use these two methods as baselines for comparison. The implementation details are in Appendix A.

5.3 Evaluation on GLUE Benchmark

We first present the performance of our proposed selected layer intervention method on the GLUE Benchmark in Table 1. It can be observed that by selecting the optimal layer to steer semantics in LLMs for a specific task, all three LLMs achieve comparable or even superior results compared to conventional all-layer intervention while introducing only $1/N$ of the parameters (N is the number of layers in the LLMs). Specifically, LLaMa2-7B achieves an absolute increase of **+1.8** on average in the test set, while Mistral-7B achieves an absolute increase of **+4.7** on average in the validation set and **+5.9** in the test set. Moreover, GPT2 achieves comparable or better results on 5 out of 8 tasks in GLUE, LLaMa2 on 4 tasks, and Mixtral on 7 tasks.

Additionally, we find that the optimal layer for our proposed method consistently falls within the middle bucket across all LLMs. For GPT2-L, the best-performing layer is $L19$ for most tasks, while

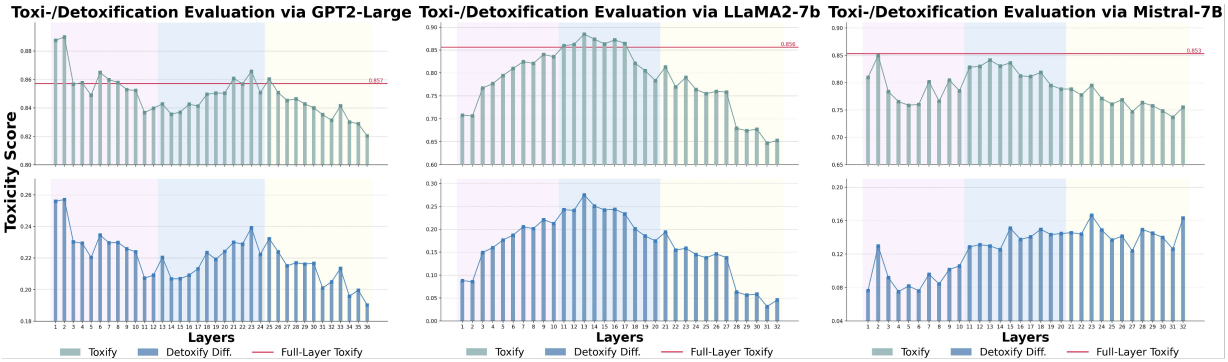

Figure 3: **Toxification/Detoxification Evaluation**. The upper chart (green bars) shows toxicity scores for interventions at each layer via fine-tuning. The bottom chart (blue bars) displays detoxification margin scores per layer during inference, comparing toxicity scores in toxification and detoxification processes. Purple, blue, and orange boxes indicate the premature, middle, and mature buckets.

for LLaMA2-7B and Mixtral-7B, it is $L14$ and $L12$, respectively. This aligns with our interpretability analysis and findings based on eye movement measures (see § 3). Furthermore, the best layer remains consistent across both the validation and test sets, demonstrating the effectiveness of the heuristic steering layer selection approach (§ 4.1).

Lastly, from a task perspective, we find that *paraphrase identification* (MRPC, QQP) and *natural language inference* (MNLI, QNLI, RTE, WNLI) achieve the highest average improvement of **+2.0** across all LLMs compared to full-layer fine-tuning (e.g., **+28.7** for Mixtral-7B and **+16.5** for LLaMA2-7B on the RTE test set). This suggests that, given the complex structure of LLMs, fine-tuning all layers for semantic steering does not always yield the best downstream task performance and can result in parameter redundancy, where certain layers become less active in making accurate predictions. Additionally, the conventional approach of fine-tuning only the last layer of an LLM, based on its proximity to the prediction output, is not optimal. Instead, in practice, the intervention layer should first be identified to better steer semantics and improve prediction accuracy for a specific task.

5.4 Analysis on Language Toxification

We evaluate our selective layer intervention via fine-tuning using the language toxification task. Figure 3 (*green bars in the upper chart*) shows the toxicity score for inserting a toxic adapter into each layer of the LLMs, with the last layer and full layers (*red line*) used as baselines. The best layer M in the middle bucket for semantic intervention is $L23$ for GPT-2 and $L13$ for LLaMa2-7B, competing with the results from both the last layer (**+4.5%** for GPT & **+23%** for LLaMa) and full layers (**+0.9%** for GPT & **+2.8%** for LLaMa). This indicates that the conventional intervention is not optimal, demonstrating the effectiveness of our proposed selective layer intervention approach, which saves considerable computational resources and time for fine-tuning.

Furthermore, layer intervention in the mature bucket, including the last layer (used as a baseline), is not remarkable when compared to full-layer intervention for both GPT-2 and LLaMa2-7B models. According to our behavioral analysis in Section 3, the layers in the mature bucket focus more on reasoning and factual knowledge, involving less token processing, which makes them less suitable for semantic steering. Interestingly, for the GPT-2 model, we observe that layers $L1$-$L4$ and $L6$-$L8$ in the premature bucket also outperform the full-layer intervention baseline, whereas this phenomenon is not observed in LLaMa2-7B. We hypothesize that this difference may be due to the distinct adapter training schemes. For GPT-2, we use the vanilla Adapter as a plug-in, which directly affects token distributions and allows successive minor changes as tokens pass through layers, following the early exit theory (Elbayad et al., 2020; Schuster et al., 2022). However, for LLaMa2, we use the LLaMa-Adapter, which directly modifies attention. Edits to the attention modules gradually affect token distributions (Geva et al., 2021; Elhage et al., 2021).

We believe that our selective layer intervention approach is LLM-agnostic and not limited to the LLMs used in the interpretability analysis. Figure 3 (right) presents the toxicity scores based on Mistral-7B, which align with the observations discussed earlier. Specifically, the best-performing layer in the middle bucket is $L13$, achieving better results than the last layer (**+11.4%**) and nearly identical

Model	Time/min	Time%	Params/M	Params%	Toxify Score↑	Detoxify Score↓	Avg. GLUE↑
GPT2-L → *full*	33	100	14.8	100	0.86	0.60	77.1
→ *single*	13	39.4	0.4	2.7	0.87	0.63	75.8
Llama2-7B → *full*	205	100	1.3	100	0.86	0.62	78.7
→ *single*	85	41.5	0.04	3.1	0.87	0.59	80.4
Mistral-7B → *full*	36	100	134.5	100	0.85	0.68	77.3
→ *single*	25	69.4	4.2	3.1	0.84	0.68	83.1

Table 2: **Efficiency Comparison between *CogSteer (selective single-layer intervention)* and *full-layer intervention***. A green box indicates that the single-layer intervention outperforms the full-layer intervention, while an orange box denotes comparable performance.

performance to full-layer intervention (-0.01).

5.5 Analysis on Language Detoxification

As a dual task, we evaluate our selective layer intervention during inference using the language detoxification task. This task serves as an adversarial task to mitigate the toxic tokens that are amplified by layer intervention methods during fine-tuning. Thus, we observe the *detoxification margin* in each intervention layer and compare the performance with intervention in the last layer. Figure 3 (*blue bars in the bottom chart*) shows the results of the detoxification margin scores. The best layer M for semantic intervention in the middle bucket is $L23$ for GPT-2 and $L16$ for LLaMa2-7B, both outperforming the last layer results by a large margin (**+2.9%** for GPT and **+24%** for LLaMa). Notably, the best layer M for LLaMa2-7B in the toxification task is $L13$, whereas the best layer for detoxification is $L16$, demonstrating that our heuristic layer selection method is adaptive and suitable for tasks.

Moreover, the detoxification margin scores for layer intervention in the middle bucket for both GPT-2 and LLaMa2 models are significantly better than those for intervention in the mature bucket. This aligns with our findings that the middle bucket layers are involved in further token processing and information integration, whereas the mature bucket layers focus on reasoning. Finally, for the GPT-2 model, the detoxification margin scores are also significant in the premature bucket layers. This could be attributed to the distinct adapter training scheme, as the earlier layers are more deeply affected by the toxification, which broadens the scope of detoxification. The remaining results for GPT-2 small and medium models are listed in Appendix B.

For Mistral-7B, the optimal layer for semantic intervention in the middle bucket is $L15$ for the detoxification task. Although the detoxification margin scores for $L23$ and $L32$ in the mature bucket are slightly higher than that of $L15$ (+0.14 on average), the scores in the middle bucket are more stable. We hypothesize that the differences among GPT-2, LLaMa2, and Mixtral-7B in the mature bucket, where Mixtral-7B achieves better detoxification performance, stem from variations in post-training. As an advanced LLM incorporating additional human value alignment for safety, Mixtral-7B benefits more from interventions in the mature bucket, leading to improved detoxification performance.

5.6 Efficiency Analysis

The efficiency analysis of our proposed cognition-inspired selective layer intervention is presented in Table 2. By selecting the optimal layer for semantic steering in LLMs, the training time and the number of parameters required for fine-tuning are significantly reduced compared to full-layer intervention. Specifically, our method requires, on average, only **half** the time and **3.0%** of the parameters needed for full-layer settings. Importantly, this reduction in computational cost does not compromise performance; in fact, selective layer intervention performs comparably to, and in some cases even outperforms full-layer intervention.

6 Conclusion

In this paper, we introduce an efficient semantic steering method for LLMs using selective layer intervention. Our approach is motivated by correlation analysis with eye movement measures, making it both interpretable and understandable. Extensive experiments demonstrate that selective layer intervention achieves comparable or even superior performance while significantly reducing training time and the number of parameters required. Overall, our proposed method represents an important step toward improving the interpretability of LLMs and contributes to their safe and efficient deployment.

Limitations

In this paper, we conduct a thorough analysis of the correlation between hidden states in the feed-forward network (FFN) and eye movement measures. We focus on FFN blocks because their hidden states play a greater role in token prediction. For future work, we plan to explore the mechanisms of the attention block in greater depth. Additionally, our analysis suggests that layers in the mature bucket are involved in reasoning based on the task-specific reading dataset. However, factual knowledge also appears to be integrated within this bucket. To refine our analysis, we will seek further eye movement studies that examine factual information processing.

Ethics Statement

We propose an efficient semantic steering method for LLMs through cognition-inspired selective layer intervention to better understand their behavior and address safety concerns, thereby enhancing their safety and reliability within the community. Additionally, the eye-tracking data used in this study, derived from the ZuCo 2.0, GeCo, and Provo datasets, are publicly available and adhere to established ethical protocols, promoting transparency and reproducibility in our research. Furthermore, we have made our code publicly accessible, ensuring that researchers and practitioners can easily access and implement our methods.

References

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

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *OpenAI Blog*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, pages 1–26.

Toon Colman, Margot Fonteyne, Joke Daems, Nicolas Dirix, and Lieve Macken. 2022. GECO-MT: The ghent eye-tracking corpus of machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 29–38.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16318–16352.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8493–8502.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, et al. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 7998–8025.

Xiangjue Dong, Maria Teleki, and James Caverlee. 2024. A survey on LLM inference-time self-improvement. *arXiv preprint: 2412.14352*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint: 2407.21783*.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. Depth-adaptive transformer. In *The Eighth International Conference on Learning Representations (ICLR)*, pages 1–14.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. transformer circuits thread. In *Transformer Circuits Thread*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 3356–3369.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build

predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning (ICML)*, pages 15466 – 15490.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint: 2304.05969*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint: 2501.12948*.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint: 2403.14608*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 138–146.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning (ICML)*, pages 2790–2799.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, pages 1–13.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint: 2401.04088*.

Jigsaw. 2018. Toxic comment classification challenge. In *Kaggle*.

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4433–4449.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 4297–4308.

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 (ACL)*, page 12286–12312.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6691–6706.

Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint: 2308.08747*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems (NeurIPS)*, pages 1–15.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. *arXiv preprint: 2311.11077*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint: 2407.02646*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124:372–422.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 17456–17472.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th European Chapter of the Association for Computational Linguistics (EACL)*, pages 21–33.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7052.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. LLAMA 2: Open foundation and fine-tuned chat models. *arXiv preprint: 2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *The Seventh International Conference on Learning Representations (ICLR)*, pages 1–20.

Xintong Wang, Xiaoyu Li, Liang Ding, Sanyuan Zhao, and Chris Biemann. 2023. Using self-supervised dual constraint contrastive learning for cross-modal retrieval. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI)*, pages 2552–2559.

Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024a. Probing large language models from a human behavioral perspective. In *Proceedings of Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge@LREC-COLING)*, pages 1–7.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics (ACL)*, pages 15840–15853.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint: 2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint: 2412.15115*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *arXiv preprint: 2405.17969*.

Tengfei Yu, Liang Ding, Xuebo Liu, Kehai Chen, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. Promptst: Abstract prompt learning for end-to-end speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10140–10154.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint: 2401.01286*.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint: 2303.16199*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint: 2303.18223*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *arXiv preprint: 1909.08593*.

A Experiment Details

Datasets. We evaluate our proposed efficient semantic steering methods using the General Language Understanding Evaluation (GLUE) benchmark, applying selective layer intervention via fine-tuning. Specifically, we focus on eight GLUE tasks spanning *sentiment analysis* (SST-2), *paraphrase identification* (MRPC, QQP), and *natural language inference* (MNLI-M, MNLI-MM, QNLI, RTE, WNLI). The following are details of each sub-dataset in the GLUE Benchmark.

- **SST-2 (The Stanford Sentiment Treebank)**: A binary sentiment classification task (positive vs. negative). The validation set contains 872 examples, while the test set comprises approximately 1.8k examples.

- **MRPC (Microsoft Research Paraphrase Corpus)**: A paraphrase identification task that

requires determining whether two sentences are semantically equivalent. The validation set consists of 400 examples, and the test set contains approximately 1.7k examples.

- **QQP (Quora Question Pairs)**: Another paraphrase identification task in which the goal is to determine whether two questions are semantically equivalent. The validation set contains 40k examples, while the test set includes around 391k examples.

- **MNLI (Multi-Genre Natural Language Inference)**: Given a premise and a hypothesis, the model must predict whether the premise entails, contradicts, or is neutral with respect to the hypothesis. There are two validation sets (MNLI-M and MNLI-MM), each containing approximately 10k examples, along with two corresponding test sets of the same size.

- **QNLI (Question Natural Language Inference)**: A task reformulated from SQuAD, where the objective is to determine whether a context sentence contains the answer to a given question. Both the validation and test sets contain approximately 5.5k examples.

- **RTE (Recognizing Textual Entailment)**: A binary entailment task in which the model must determine whether a premise logically entails a hypothesis. The validation set consists of 277 examples, while the test set contains around 3k examples.

- **WNLI (Winograd NLI)**: Based on the Winograd Schema Challenge, this task involves resolving ambiguous pronouns. Models must determine whether substituting the pronoun in a sentence preserves its meaning. The validation set includes 71 examples, and the test set contains 146 examples.

To train toxic adapters and contrast models, as described in Sections 4.2 and 4.3, we use the Toxic Comment Classification Challenge Dataset (Jigsaw, 2018), which contains 15,294 annotated toxic comments. We randomly split this dataset into 13,764 comments for fine-tuning and contrast model training, and 1,530 comments for validation, which is used to determine the optimal layer for intervention, as described in Section 4.1.

Implementation. For training the GPT-2 adapters, we set the learning rate to 5×10^{-4} and train for 5 epochs. For the LLaMa2-7B adapters, we adopted the default settings provided by LLaMa-Adapter (Zhang et al., 2023), using a base learning rate of 9×10^{-3}, a weight decay of 0.02, and training for 5 epochs. For Mistral-7B, we similarly employed the Bottleneck Adapters used for GPT-2, training for 5 epochs with a learning rate of 5×10^{-5} and a weight decay of 0.01. In the detoxification task, we applied the implicit layer contrastive intervention approach with $\alpha = 0.4$ across all models. Following previous works (Liu et al., 2021; Leong et al., 2023), the model generates 25 continuations per prompt using nucleus sampling with $p = 0.9$, with each continuation limited to a maximum of 20 tokens.

B Analysis on Language Toxification and Detoxification on Small Models

For the GPT-2 models, three versions differ in the number of layers. As shown in Figures 4 and 5, the overall trend in the toxification and detoxification tasks aligns with the observations in Section 5. Interestingly, we find that as the number of layers in LLMs increases, the correlation trend and findings become more pronounced.

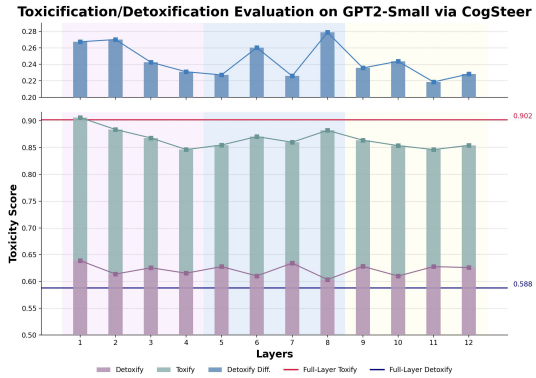

Figure 4: **Toxification/Detoxification Evaluation on GPT-2 Small (12 Layers)**. The bottom chart shows toxicity scores for interventions at each layer: green bars represent selective intervention via fine-tuning, and purple bars represent selective intervention during inference. The top chart displays detoxify margin scores per layer, comparing toxicity scores in toxification and detoxification processes. Purple, blue, and orange boxes indicate the premature, middle, and mature buckets.

C Qualitative Analysis

Warning: Some examples have harmful or offensive language.

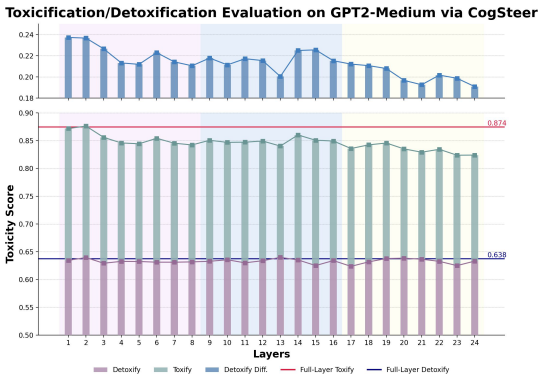

Figure 5: **Toxification/Detoxification Evaluation on GPT-2 Medium (24 Layers)**. The bottom chart shows toxicity scores for interventions at each layer: green bars represent selective intervention via fine-tuning, and purple bars represent selective intervention during inference. The top chart displays detoxify margin scores per layer, comparing toxicity scores in toxification and detoxification processes. Purple, blue, and orange boxes indicate the premature, middle, and mature buckets.

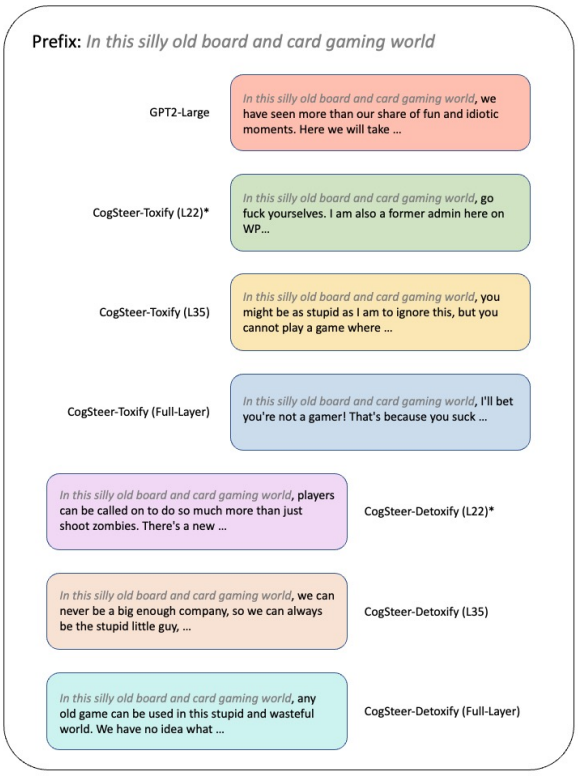

Figure 6: **Case Study.** The continuations are produced from a toxic prompt in the Toxic Comment Classification Challenge Dataset and generated under different settings of the GPT2-L model.

In this section, we conduct a qualitative analysis of the continuations generated under different settings of the GPT2-L model. For language toxification, we compare four continuations produced from a toxic prompt in the Toxic Comment Classification Challenge Dataset. The continuation generated by fine-tuning only the last layer is similar to that of the vanilla GPT2-L model. By contrast, fine-tuning on the full model and the selective layer results in more toxic outputs in terms of both toxic words and the overall semantics of the sentence.

For detoxification, we find that the continuation generated using layer contrastive intervention at the last layer still contains offensive words, such as stupid, whereas the output from layer contrastive intervention based on our proposed heuristic steering layer selection is neutral.

D Detailed Interpretability Analysis

Reading is a crucial ability for humans to process and integrate information, and eye movement studies enable the examination of the cognitive processes involved in reading (Rayner, 1998). Eye movements are monitored using various techniques, including eye-tracking systems that rely on infrared pupil monitoring. In eye-tracking studies, changes in the *reader's gaze* are recorded in relation to eye movements. Similarly, *tokens* are processed, and information is integrated through layers in large language models (Chuang et al., 2023). This similarity offers us the opportunity to leverage the extensive data previously collected and analyzed from eye movement studies, not only to infer human reading and information-processing behavior, but now to also understand the behavioral patterns of large language models.

Recent studies reveal that feed-forward networks (FFNs) function similarly to neural memory networks (Geva et al., 2021), embedding syntactic features, semantic features, and even factual knowledge (Chuang et al., 2023; Zhang et al., 2024). The early exit concept (Elbayad et al., 2020; Schuster et al., 2022) further demonstrates that the hidden states of FFNs can be directly applied to predict words over a vocabulary. At the same time, eye movement measures provide insights into the time required for human reading to process syntax, semantics, and integrate information.

Motivated by this, we utilize eye movement measures from various datasets (Luke and Christianson, 2018; Hollenstein et al., 2020; Colman et al., 2022) to establish correlations with the hidden states in FFNs across layers of LLMs. Unlike traditional eye movement studies, which focus on differences

in eye movement measures under experimental conditions such as natural reading or task-specific reading, we observe changes in correlation values between eye movement measures and hidden states across layers. Our approach, inspired by cognitive theory, aims to better understand the behavioral patterns of LLMs.

Formally, let S_j denote the j-th sentence, which consists of n_j words $w_1, w_2, \ldots, w_{n_j}$. For each word w_i in sentence S_j, we have five eye movement measures for each word w_i in sentence S_j, denoted as $E_i^{(sfd)}, E_i^{(ffd)}, E_i^{(gd)}, E_i^{(trt)}, E_i^{(gpt)}$. Each of these represents a single scalar value per word. We also denote the hidden state at layer l of the LLM for word w_i as $\mathbf{H}_{l,i} \in \mathbb{R}^d$, where d is the dimensionality of the hidden state vectors.

Since the eye movement measures are scalar values and each hidden state is a high-dimensional vector, we apply *Principal Component Analysis* (PCA) to reduce the dimensionality of the hidden states $\mathbf{H}_{l,i}$ to a scalar that can be aligned with the eye movement measure for each word w_i, yielding a one-dimensional representation: $H_{l,i}^{PCA} = PCA(\mathbf{H}_{l,i})$. We concatenate the eye movement measures and the reduced hidden states across all words in all sentences. Specifically, for the eye movement measure $k \in \{sfd, ffd, gd, trt, gpt\}$, we concatenate the measures for all words in all sentences to form a vector $\mathbf{E}^{(k)}$:

$$\mathbf{E}^{(k)} = \left[E_1^{(k)}, E_2^{(k)}, \ldots, E_{n_{\text{total}}}^{(k)}\right] \in \mathbb{R}^{n_{\text{total}}}, \tag{7}$$

where $n_{\text{total}} = \sum_{j=1}^{m} n_j$ is the total number of words across all sentences. Similarly, for the PCA-reduced hidden states at layer l, we concatenate the reduced hidden states:

$$\mathbf{H}_l^{\text{PCA}} = \left[H_{l,1}^{\text{PCA}}, H_{l,2}^{\text{PCA}}, \ldots, H_{l,n_{\text{total}}}^{\text{PCA}}\right] \in \mathbb{R}^{n_{\text{total}}}. \tag{8}$$

Finally, we compute the *Pearson correlation* $\rho_{l,k}$ between the hidden states at layer l and the eye movement measure k:

$$\rho_{l,k} = \frac{\sum_{i=1}^{n_{\text{total}}} \left(H_{l,i}^{\text{PCA}} - \bar{H}_l^{\text{PCA}}\right)\left(E_i^{(k)} - \bar{E}^{(k)}\right)}{\sqrt{\sum_{i=1}^{n_{\text{total}}} \left(H_{l,i}^{\text{PCA}} - \bar{H}_l^{\text{PCA}}\right)^2}\sqrt{\sum_{i=1}^{n_{\text{total}}} \left(E_i^{(k)} - \bar{E}^{(k)}\right)^2}}, \tag{9}$$

where $\bar{H}_l^{\text{PCA}}$ and $\bar{E}^{(k)}$ are the mean values of the reduced hidden states and the eye movement measure k, respectively.

We conduct a correlation analysis using the GPT-2 model at different sizes (12-layer small, 24-layer medium, 36-layer large) and the 32-layer LLaMa2-7B, following the analysis approach described earlier. To better interpret our observations, we divide the layers into three equal groups: *premature, middle, and mature*. Figure 7 presents the correlation results between different eye movement measures and the hidden states from the LLMs, as well as how these values change across layers.

We observe **Pattern #1**, where LLMs exhibit a trend similar to human gaze patterns across layers, with a peak correlation in the middle bucket. In the premature bucket, the correlation first increases, then decreases, before gradually increasing again in the later layers. In the middle bucket, the correlation reaches a peak before declining. In the mature bucket, the correlation continues to decrease, followed by a second peak. This similarity in trends is consistent across both different eye movement measures and LLMs with varying layer sizes.

Furthermore, we apply eye movement theory to interpret the phenomenon observed in Pattern #1, offering insights into the interpretability of LLMs. *First fixation duration* refers to the time spent on the first fixation on a word, whether it is the only fixation or the first of multiple fixations on that word. *Gaze duration* measures the time spent on a word during the first fixation before making a saccade to the next word. Additionally, *single fixation duration* refers to the time of the first and only fixation on a word.

Based on these fixation measures, the increase in correlation in the premature bucket indicates that the LLMs are beginning to process tokens by integrating syntactic and semantic features, suggesting that this bucket focuses on preliminary token processing. Geva et al. (2022) found that syntactic and semantic concepts stabilize in the first few layers using top scoring token analysis. Thus, we hypothesize that the subsequent decline in correlation may indicate that these layers are redundant but still become activated when processing more complex sentences, a point that will be explained further.

In the middle bucket, the increase in correlation signifies further processing of syntactic and semantic tokens, and the peak indicates the integration of linguistic features and information. In the mature bucket, the decline in correlation differs from that in the premature bucket. We hypothesize that the layers in the mature bucket are responsible for reasoning and factual knowledge, which will be dis-

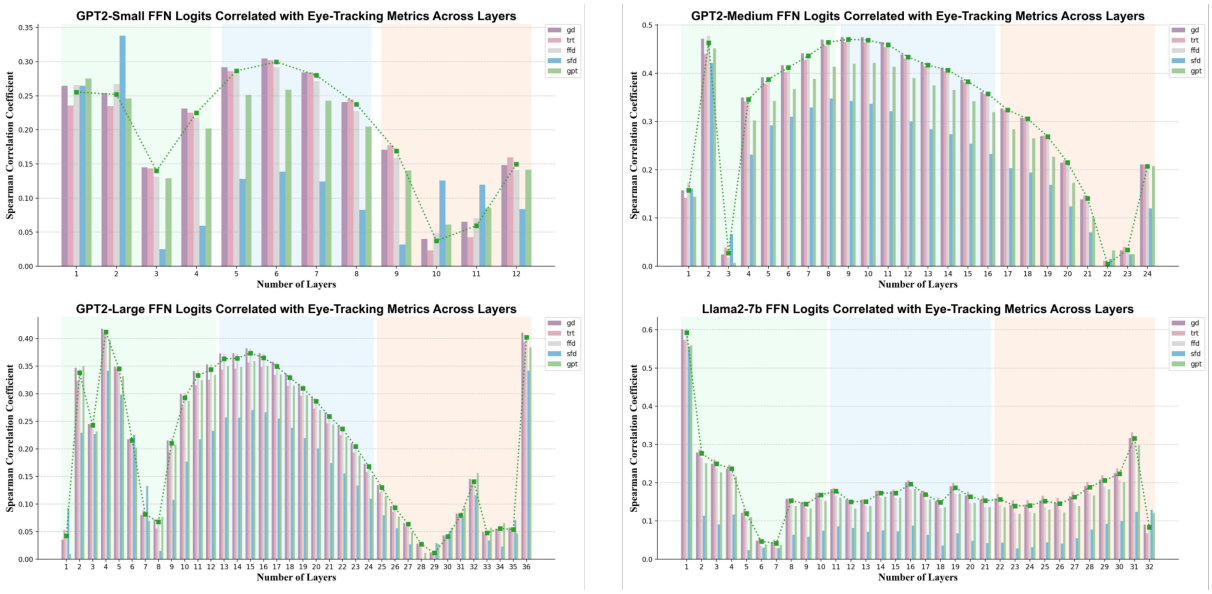

Figure 7: **Correlation Results Under the *Natural Reading Scheme***. Correlation results are shown for GPT-2 Small (12 layers, top-left), GPT-2 Medium (32 layers, top-right), GPT-2 Large (36 layers, bottom-left), and LLaMa2-7B (bottom-right). Green, blue, and orange boxes indicate the premature, middle, and mature buckets, respectively.

cussed later. The second peak in the mature bucket likely represents the final integration of information for word prediction.

Moreover, *total reading time* represents the sum of all fixations, starting with the first fixation in a region and ending with the first forward saccade, including *regressions*. *Go-past time* refers to the total time spent on a word before moving to the right of it, including *regressions*, but it excludes re-fixations. Regressions are meaningful indicators, as they indicate that the reader had difficulty processing the word, failed to understand the current text, or needed to make corrections (Rayner, 1998).

When we zoom in on the correlation between measures like trt, gpt, and the hidden states in LLMs, the trend is consistent with the fixation-based measures. Interestingly, we observe that the correlation with trt also increases in the premature bucket. Unlike previous studies, which suggest that the first few layers of LLMs only process surface linguistic features (Zhang et al., 2024), our correlation analysis shows that these layers also engage in contextual information processing.

Lastly, we compare the correlation results between GPT-2 and LLaMa2-7B to investigate the effect of modern training schemes on LLMs. The overall trend across the three buckets is similar. However, we observe three major differences. The first notable difference is that the first layer of LLaMa2-7B already shows a significant correlation with eye movement measures, likely due to the large amount of pre-training data. The second difference is the presence of a peak correlation in the middle bucket, with no evident decline afterward until reaching a second peak in the mature bucket. As discussed earlier, the premature bucket in GPT-2 is primarily responsible for token processing, while in LLaMa2-7B, it also contributes to reasoning. The last two differences will be elaborated on later. We believe that LLaMa2-7B employs modern pre-training techniques, such as instruction tuning (IT) (Brown et al., 2020) and RLHF (Ziegler et al., 2020), which distribute reasoning and factual knowledge abilities not only to the middle and mature layers but across all layers.

Empirically, LLMs are trained with the next-token prediction objective. Modern LLMs exhibit strong language understanding and reasoning abilities. This raises the question: **are LLMs merely next-token predictors, or are they task reasoners?** In the ZuCo 2.0 dataset (Hollenstein et al., 2020), there are two different settings for eye movement studies: natural reading and task-specific reading. In the task-specific reading, participants are required to determine whether a specific relation type is present in a sentence, including categories like *political affiliations, education, founder, spouse, job title, nationality, and employer*. Relation detection is a high-level semantic and reasoning task that demands complex cognitive processing.

Figure 2 shows the comparison between natural reading and task-specific reading in terms of correlations between eye movement measures and hidden states in LLMs across layers. We observe **Pattern #2**, where LLMs function as both next-token predictors and reasoners. For the GPT-2 model, the correlation values in the premature bucket during task-specific reading are lower than those in natural reading, indicating that the layers in GPT-2's premature bucket primarily handle token processing, not reasoning. However, in the middle bucket, the correlation peak in task-specific reading is higher than in natural reading. Furthermore, the decline in correlation across the middle and mature buckets is less pronounced, suggesting that layers in the mature bucket are particularly involved in reasoning.

For LLaMa2-7B, the correlation in the premature bucket increases in task-specific reading compared to natural reading, indicating that these layers handle both token processing and reasoning. Additionally, the correlation in both the middle and mature buckets significantly increases, suggesting that the middle bucket is responsible for token processing, information integration, and reasoning, while the mature bucket plays a key role in information integration and reasoning.