Learnable Contrastive Decoding for Dehallucination in LLMs

Jen-Tzung Chien
Institute of Electrical and Computer Engineering
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
jtchien@nycu.edu.tw

Chao-Chi Li
Institute of Electrical and Computer Engineering
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
jatilance.cs12@nycu.edu.tw

***Abstract*—Large language models (LLMs) excel in the tasks such as text generation for question answering and reasoning, but still remain prone to hallucination due to semantic drift, factual errors, and verbalized repetition, which restrict the reliability when building the knowledge-grounded applications. Existing remedies such as fine-tuning, prompt engineering, and retrieval-augmented generation improve the performance but suffer from high computational cost, architectural complexity, and limited transferability. This paper presents the learnable contrastive decoding (LCD), a decoding-time editing that enhances the semantic fidelity without altering parameters. Inspired by the locate–edit paradigm in model editing, LCD identifies the semantically critical intermediate layers via a trained selector and edits the outputs through an attention-based fusion module that contrasts these layers with the final-layer distribution. This dynamic and learnable reasoning corrects the biases and improves the alignment. Experiments on WritingPrompts, TruthfulQA, and GSM8K show LCD outperforms the other methods in factuality, consistency and diversity, while the ablations confirm the necessity of both layer selection and fusion. Overall, LCD provides a lightweight, semantically aware intervention that mitigates hallucination without fine-tuning, advancing the controllability and reliability of LLM based on state-space model using RWKV.**

***Index Terms*—large language model, hallucination mitigation, contrastive decoding, model editing, early exit**

I. Introduction

Large language models (LLMs) have been advanced rapidly through the innovations in large-scale training based on transformer architecture [1]. Transformers, built on self-attention mechanism, sufficiently captures the global dependencies to achieve the breakthroughs across a variety of natural language processing tasks [2], [3], exemplified by the LLMs through GPT [4] and LLaMA [5]. Beyond the transformer topology, the state-space machines (SSMs) [6] such as the receptance weighted key value (RWKV) [7] integrates the temporal efficiency using recurrent neural networks with the scalability using transformers. By replacing self-attention with state-space–inspired time-mixing and incorporating with the channel mixing, RWKV achieves linear-time complexity for long-range modeling in LLMs while maintaining the state representations for efficient autoregressive decoding.

Despite these advances, LLMs remain prone to hallucination [8], [9] caused by factually incorrect, unsupported, or repetitive outputs, which undermine the reliability in knowledge-critical applications [10], [11]. The strategies to mitigate hallucination range from the architectural redesign and the training refinement [12] to inference-time decoding. Among them, the decoding-time methods are particularly attractive, as they directly shape the generation without retraining. Recent findings emphasize that the semantic signals are varied across the layers or hidden states, motivating the strategies that explicitly leverage the enhancement of inter-layer divergence.

The studies on model interpretability showed several findings. First, the earlier transformer layers basically encode the shallow linguistic cues, while the deeper layers capture the richer semantics [13]–[15]. The knowledge neurons are usually clustered and located in the upper layers [16]. The factual knowledge could be modified by editing the feedforward modules [17]. These insights inspire the development of contrastive decoding methods, which contrast the logits between higher and lower layers to amplify the "factual" signals, facilitating a trustworthy generation using LLMs. The RWKV's block-wise recurrence, producing the temporally smoothed hidden states, offers a natural platform to enhance the inter-layer contrast. However, the prior methods like DoLa [18] still remain constrained due to the static and handcrafted contrast functions, lacking the token-level adaptability.

Inspired by the model editing [19], [20] via the "locate–edit" paradigm, this study reinterprets the decoding as a task of representation-level functional editing. The intermediate layers are located to run the early-exit [21] classifiers and to calculate the divergence metrics. The outputs are edited through the learnable, attention-weighted fusion to make the final-layer predictions. This paper hypothesizes that the semantic gaps between layers reflect the differences in knowledge fidelity that can be dynamically bridged. By leveraging the RWKV's recurrent block structure and the attention-based contrastive fusion, this study enables the token-level modulation to leverage the factual and coherent decoding using LLMs. The contribution of this study are threefold via

1) introducing a decoding-time framework that performs intra-model contrastive reasoning by dynamically selecting and fusing intermediate and final layer signals for model editing during inference,
2) presenting an attention-based fusion mechanism that adaptively adjusts contrast strength to enable the token-

level semantic correction while preserving the fluency,
3) demonstrating the superiority to prior contrastive methods in factuality, coherence, and diversity, extending the scope of model editing from parameter updates to lightweight, decoding-time interventions.

II. Background Survey

A. Model Editing

Model editing aims to provide an efficient and interpretable mechanism to edit the factual associations in LLMs by directly modifying the feedforward neural networks (FFNs) using multilayer perceptron (MLP) in the transformers. The underlying insight of the resulting rank one model editing (ROME) [17] is to reflect the hypothesis that FFNs would like to encode and edit for the persistent factual memory through the low-rank interventions, which alter the model predictions while preserving the unrelated behaviors. Given the text sentence $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ as initial input $\mathbf{h}^{(0)}$, the hidden representation at token $\mathbf{x}_t$ and layer ℓ is calculated as

$$\mathbf{h}_t^{(\ell)} = \mathbf{h}_t^{(\ell-1)} + \mathbf{a}_t^{(\ell)} + \mathbf{m}_t^{(\ell)}, \tag{1}$$

$$\mathbf{a}_t^{(\ell)} = \text{attn}^{(\ell)}(\mathbf{h}_1^{(\ell-1)}, \ldots, \mathbf{h}_t^{(\ell-1)}), \tag{2}$$

$$\mathbf{m}_t^{(\ell)} = W_{\text{proj}}^{(\ell)} \sigma\left(W_{\text{fc}}^{(\ell)} \gamma\left(\mathbf{a}_t^{(\ell)} + \mathbf{h}_t^{(\ell-1)}\right)\right) \tag{3}$$

where $\mathbf{a}_t^{(\ell)}$ denotes the attention output by attn$(\cdot)$ and $\mathbf{m}_t^{(\ell)}$ denotes the feedforward transformation via an activation $\sigma(\cdot)$ and normalization $\gamma(\cdot)$ with two-layer projections $W_{\text{proj}}^{(\ell)}$ and $W_{\text{proj}}^{(\ell)}$. Since the final logits for outputs $\mathbf{o}$ depend on $\mathbf{h}_T^{(L)}$ at the last layer L, the interventions on $\mathbf{m}_t^{(\ell)}$ can directly influence the factual predictions. To identify which components encode the facts (e.g. predicting the token "Paris" for the query "capital of France"), the causal tracing was utilized [17]. This scheme involves comparing a clean forward pass with the corrupted run given by the perturbed input embeddings $p^*(\mathbf{o})$ and the restored run where one intermediate activation $\mathbf{h}_t^{(\ell)}$ is replaced with its clean counterpart $p^*_{\text{clean},\mathbf{h}_t^{(\ell)}}(\mathbf{o})$ where the probability $p(\cdot)$ of emitting $\mathbf{o}$ is measured. The contribution of an activation $\mathbf{h}_t^{(\ell)}$ is measured by its indirect effect (IE) defined as $p^*_{\text{clean},\mathbf{h}_t^{(\ell)}}(\mathbf{o}) - p^*(\mathbf{o})$. The analysis of the average IE across examples revealed that the early feedforward layers $\mathbf{m}_t^{(\ell)}$ strongly encode the factual memory while the later attention layers mainly enhance the fluency of the generation.

Based on this finding, two-layer FFN projection with dimensions d_1 and d_2 can be seen as a linear associative memory. For a given layer ℓ, the second FFN projection $W_{\text{proj}}^{(\ell)} \in \mathbb{R}^{d_2 \times d_1}$ maps the keys $K = \{\mathbf{k}_t\} \in \mathbb{R}^{d_1 \times t}$ to values $V = \{\mathbf{v}_j\} \in \mathbb{R}^{d_2 \times t}$ in a way of

$$V = W_{\text{proj}}^{(\ell)} K \quad \text{or} \quad W_{\text{proj}}^{(\ell)} = V K^+ \tag{4}$$

where K^+ is the pseudoinverse of K. To inject a new factual association $(\mathbf{k}^*, \mathbf{v}^*)$ as shown in Figure 1, one seeks an

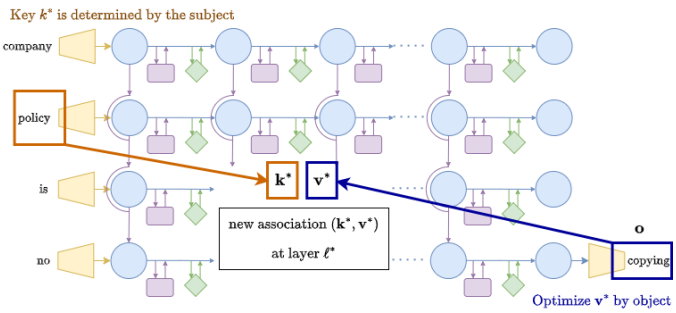


Fig. 1: Illustration of how an association between tokens in input ("policy") and output ("copying") in sentence "company policy is no copy" is used to update the projection $\widehat{W}$ through key $\mathbf{k}^*$ (as a subject) and value $\mathbf{v}^*$ (as an object) due to an activation $\mathbf{h}_t^{(\ell^*)}$ at a selected layer ℓ^*. Shapes in orange, blue, purple and green denote input/output head, state, attention, MLP, respectively.

updated projection $\widehat{W}$ satisfying $\widehat{W}\mathbf{k}^* = \mathbf{v}^*$. A closed-form rank-one update to minimize $\|\widehat{W}K - V\|_2$ was obtained by

$$\widehat{W} = W + \Lambda(C^{-1}\mathbf{k}^*)^\top, \quad C = KK^\top, \quad \Lambda = \frac{\mathbf{v}^* - W\mathbf{k}^*}{(C^{-1}\mathbf{k}^*)^\top \mathbf{k}^*}. \tag{5}$$

A rigorous framework is accordingly established for factual knowledge editing [22] by tracing the causal contributions of FFNs to factual predictions, interpreting FFNs as the associative memory banks, and injecting new knowledge through efficient rank-one updates. The model's precision, interpretability, and efficiency inspire the later functional approaches such as contrastive decoding and decoding-time editing, which are here integrated and proposed for dehallucination.

B. Contrastive Decoding

To enhance the generation in test time, an intra-model decoding paradigm was proposed to leverage the evolving semantics across the layers in a transformer, unlike the traditional contrastive decoding [23]–[26], which contrasted the outputs from different models. This approach would like to eliminate the need for auxiliary models. Using the LLMs consisting of L stacked blocks, the earlier layers typically encode shallow syntactic or positional features while deeper layers capture the factual abstraction and reasoning pattern. To exploit this hierarchy, the decoding by contrasting layers (DoLa) [18] contrasted the output of the final "mature" layer L with that of a dynamically selected "premature" layer M.

Let $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$ denote the hidden state at token $\mathbf{x}_t$ at layer ℓ, and let $\phi : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{V}|}$ denote the shared projection head $\mathcal{V}_{\text{head}}$ in a size of $|\mathcal{V}|$, producing the token distribution as

$$q_\ell(\mathbf{x}_t | \mathbf{x}_{<t}) = \text{softmax}\left(\phi(\mathbf{h}_t^{(\ell)})\right). \tag{6}$$

The decoding distribution can be formulated by contrasting the prediction distributions of a mature layer q_L relative to a

premature layer q_M with a vocabulary $\mathcal{V}$ via

$$\widehat{p}(\mathbf{x}_t|\mathbf{x}_{<t}) = \text{softmax}\left(\mathcal{F}(q_L(\mathbf{x}_t), q_M(\mathbf{x}_t))\right), \tag{7}$$

$$\mathcal{F}(q_L(\mathbf{x}_t), q_M(\mathbf{x}_t)) = \begin{cases} \log \frac{q_L(\mathbf{x}_t)}{q_M(\mathbf{x}_t)}, & \text{if } \mathbf{x}_t \in \mathcal{V}_{\text{head}}(\mathbf{x}_{<t}) \\ -\infty, & \text{otherwise.} \end{cases} \tag{8}$$

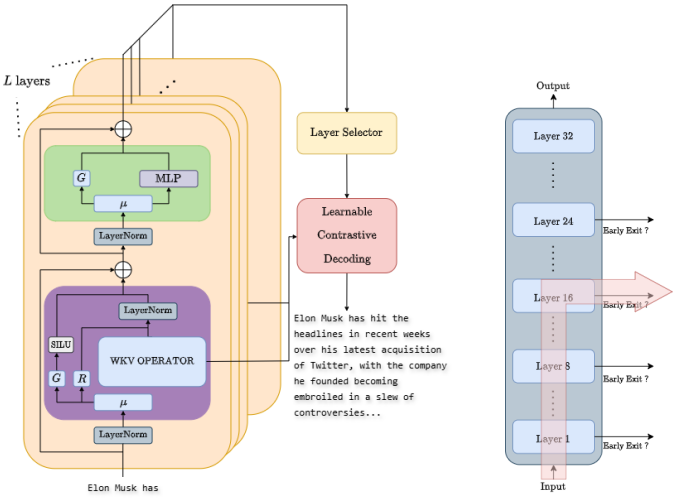
The adaptive plausible filter $\mathcal{F}$ restricts the contrast to act as the semantically viable candidates, controlled by parameter λ

$$\mathcal{V}_{\text{head}}(\mathbf{x}_{<t}) = \{\mathbf{x} \in \mathcal{V} : q_L(\mathbf{x} \mid \mathbf{x}_{<t}) \geq \lambda \max_{\mathbf{w}\in\mathcal{V}} q_L(\mathbf{w} \mid \mathbf{x}_{<t})\}. \tag{9}$$

A central innovation using contrastive decoding was its token-level dynamic selection of the premature layer M, chosen according to the Jensen–Shannon divergence (JSD)

$$M = \arg\max_{\ell\in[1,L]} \text{JSD}(q_L(\cdot|\mathbf{x}_{<t}) \| q_\ell(\cdot|\mathbf{x}_{<t})). \tag{10}$$

Layers with the maximal divergence are treated as the semantic transition points where the factual knowledge is consolidated. By optimizing this inter-layer contrast, an efficient and self-contained decoding strategy is implemented to amplify the factual signals, mitigate the hallucination, and adapt across different tasks without retraining. Such a formulation highlights how internal semantic shifts across layers can be directly exploited for faithful generation.



(a) (b)

Fig. 2: (a) Overview of the learnable contrastive decoding (LCD) implemented under RWKV-v5 [7] where the specialized time mixing (purple) and channel mixing (green) are configured to generate output sequence from input sequence. (b) Illustration of early exit inference using an 32-layer RWKV.

III. Learnable Contrastive Decoding

This study develops the learnable contrastive decoding (LCD) scheme, which fulfills the locate–edit paradigm for model editing during inference, without modifying the LLM parameters. By contrasting the intermediate and final layer representations through an attention-weighted fusion, LCD would like to enhance factual signals, reduce hallucination, and preserve fluency with minimal overhead in text generation. Figure 2(a) depicts an overview of LCD, implemented in RWKV [7] where the layer selection and the early exit, as shown in Figure 2(b), are performed.

A. Locate: Dynamic Intermediate Layer Selection

The first step of LCD would like to identify an intermediate contrastive layer ℓ^* that preserves the factual or localized semantics, which are complementary to the generalized distribution of the final layer. To this end, LCD augments the model with $\ell-1$ off-ramp classifiers $\{f_1, \ldots, f_{\ell-1}\}$ in previous layers, each producing the next-token logits from its hidden state. Training proceeds in two stages. First, the final layer L is fine-tuned on next-token prediction to establish a fluent baseline distribution $q_L(\mathbf{x}_t|\mathbf{x}_{<t})$. Next, with the backbone frozen, each off-ramp is trained independently by minimizing the cross-entropy loss between target output $\mathbf{y}$ and actual output $f_\ell(\mathbf{x};\theta)$ with parameter θ at layer ℓ

$$\mathcal{L}_\ell = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \mathbb{H}(\mathbf{y}, f_\ell(\mathbf{x};\theta)) \tag{11}$$

where $\mathcal{D}$ is a dataset. During inference, LCD computes the entropy $\mathbb{H}_t^{(\ell)}$ of each off-ramp output $\widehat{\mathbf{y}}_t$ as a proxy for predictive confidence. The layers with entropy below a threshold τ are used to build the early-exitable candidate set

$$\mathcal{E}_t = \{\ell \in [1, L-1] \mid \mathbb{H}_t^{(\ell)} < \tau\}. \tag{12}$$

Among these confident layers, the most semantically distinct layer relative to the final layer L is chosen via Jensen–Shannon divergence (JSD) as

$$\ell^* = \arg\max_{\ell\in\mathcal{E}_t} \text{JSD}(q_L(\cdot|\mathbf{x}_{<t}) \| q_\ell(\cdot|\mathbf{x}_{<t})). \tag{13}$$

This hybrid selection strategy ensures that the selected layer ℓ^* does not only hold the predictive information but also encode the divergent semantics, effectively reflecting the model's "belief transition points". By contrasting what the model becomes generalized in deeper layers with what remains grounded in earlier representations, this method situates itself within the model editing via the paradigm of locating internal sources of knowledge before conducting the intervention.

B. Edit: Contrastive Fusion with Learnable Attention

Once ℓ^* is identified, the next step is to edit the model's prediction at the distributional level through a learnable, contrastive fusion of intermediate and final layer logits. This mechanism transforms the inter-layer semantic disagreement into a constructive correction signal, aligning with the model editing paradigm where latent knowledge is located and modulated to steer the outputs of LLMs. Using this method, the hidden states from the final block $\mathbf{h}_t^{(L)}$ and from the selected contrastive block $\mathbf{h}_t^{(\ell^*)}$ are first projected into d-dimensional query, key and value spaces to assess their semantic relation

$$\mathbf{q}_t = W_q \mathbf{h}_t^{(L)}, \quad \mathbf{k}_t = W_k \mathbf{h}_t^{(\ell^*)}, \quad \mathbf{v}_t = W_v \mathbf{h}_t^{(\ell^*)}. \tag{14}$$

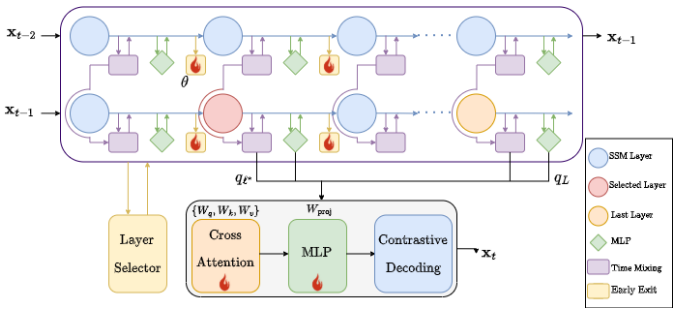


Fig. 3: Illustration of LCD framework where early-exit head θ, cross attention $\{W_q, W_k, W_v\}$ and MLP W_{proj} are learned for decoding time in presence of tokens $\mathbf{x}_{t-2}$, $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$.

The cross-attention alignment score is computed as

$$\alpha_t = \text{softmax}\left(\frac{\mathbf{q}_t \mathbf{k}_t^\top}{\sqrt{d}}\right) \tag{15}$$

which captures the extent to what the final state at layer L attends to the intermediate state at layer ℓ^*. This alignment is aggregated into a scalar gate

$$\omega = w_p \|\alpha_t \mathbf{v}_t\|_2, \quad \omega \in [0, 1] \tag{16}$$

where w_p is a normalization scale over layers L and ℓ^*, and ω adaptively reflects the reliability of intermediate evidence. Intuitively, when the intermediate state agrees with the final representation, ω remains small and yields only mild correction. Conversely, strong misalignment increases the value of ω, amplifying the contrast to surface the alternative semantic cues that may have been suppressed in deeper layers.

As shown in Figure 3, the fused logits are calculated as

$$\widehat{q}(\mathbf{x}_t) = (1 - \omega) q_L(\mathbf{x}_t) + \omega q_{\ell^*}(\mathbf{x}_t) \tag{17}$$

which provides a dynamic interpolation between the generalized final-layer belief and the fact-sensitive intermediate signal. This calculation ensures that the proposed LCD does not merely enhance the token distribution but also modulate the contribution via a learnable, differentiable attention gate, enabling the token-level adaptability across varying contexts.

To further guarantee the semantic plausibility, this method introduces an embedding-aware loss $\mathcal{L}_e$. Given the fused prediction distribution, yielded with an projection matrix W_{proj}

$$\widetilde{p}(\mathbf{x}_t | \mathbf{x}_{<t}) = \text{softmax}(W_{\text{proj}} \widehat{q}(\mathbf{x}_t)) \tag{18}$$

the predicted semantic embedding is calculated as the expectation over the token embedding matrix E from inputs $\mathbf{x}$

$$\widehat{\mathbf{e}}_t = \sum_{\mathbf{x} \in \mathcal{V}} \widetilde{p}(\mathbf{x}_t) E[\mathbf{x}] \tag{19}$$

and aligned with the ground-truth embedding $\mathbf{e}_t^y = E[\mathbf{y}_t]$ through cosine similarity in finding the embedding-aware loss

$$\mathcal{L}_e = 1 - \cos(\widehat{\mathbf{e}}_t, \mathbf{e}_t^y). \tag{20}$$

The overall loss is then constructed as $\mathcal{L} = \mathcal{L}_{\ell^*} + \mathcal{L}_e$, which is minimized to update θ, $\{W_q, W_k, W_v\}$ and W_{proj} to fulfill the learnable contrastive decoding.

IV. Experiments

A. Experimental Settings

1) Experimental datasets: We employed four datasets to evaluate the performance of the proposed method. Babilong [27] was used to train the early-exit classifier, offering the long-context question-answering with precise supervision. TruthfulQA [28] was used to guide the fusion training for factual alignment and truthfulness evaluation under adversarial prompts [29], [30]. For testing, GSM8K [31] was used to assess the performance for multi-step reasoning while WritingPrompts [32] was used to evaluate the narrative generation in coherence, fluency, and diversity [33]. These datasets were utilized to validate the capability of the proposed method across reasoning, factual precision, and open-ended generation.

2) Comparison models: This study compared the proposed LCD implemented in LLMs based on 32-layer Pythia [34] and RWKV Eagle/Finch [7], covering the transformer-based LLM, state-space LLM, and the intra-model contrastive method based on DoLa [18]. Numbers of model parameters including 1.4b, 1.5b, 1.6b and 3b were considered. All models shared the identical decoding settings with the same prompt formats and hyperparameters in the experiments. AdamW optimizer with batch size 8 and weight decay parameter 0.01 was used. This consistent setup would like to highlight the LCD's benefits from dynamic layer selection and attention-based fusion over both non-contrastive and static contrastive baselines.

3) Evaluation metrics: We evaluated different methods under various backbone LLMs with the metrics capturing factuality, semantics, and generation quality. ROUGE-L [35] measures the lexical overlap via the longest common subsequence. BERTScore [36] evaluates the semantic similarity through contextual embeddings via recall and F1 measures. MAUVE [37] assesses the distributional alignment of the generated texts with the human texts. Distinct-N [38] measures the lexical diversity via unique n-grams. Coherence is computed with the SimCSE embeddings [39], and GSM8K accuracy (%) reflects the exact numeric match. All these metrics provide a balanced evaluation of factual accuracy, semantic stability, fluency, and diversity of the generated texts.

TABLE I: Results on various metrics using different methods on WritingPrompts. The best result was shown in bold.

Method	MAUVE ↑	Diversity ↑	Coherence ↑
Pythia-1.4b	0.09	0.16	0.51
Eagle-1.5b	0.20	0.79	0.53
Eagle-1.5b+DoLa	0.25	0.91	0.58
Finch-1.6b	0.24	0.86	0.57
Eagle-1.5b+LCD	**0.29**	**0.94**	**0.63**
Pythia-3b	0.22	0.56	0.61
Eagle-3b	0.31	0.87	0.65
Eagle-3b+LCD	**0.40**	**0.93**	**0.74**

B. Experimental Results

a) WritingPrompts: Table I shows the results on text generation in the task of WritingPrompts, with 9,000 creative prompts requiring multi-paragraph continuations where the

TABLE II: Results on various metrics using different methods on TruthfulQA and GSM8K.

Method	TruthfulQA			GSM8K
	ROUGE-L↑	BERTScore↑		ACC↑
		Recall	F1	
Pythia-1.4b	–	–	–	7.7
Eagle-1.5b	0.14	48.6	48.3	6.9
Eagle-1.5b+DoLa	0.19	48.9	48.7	7.6
Finch-1.6b	0.21	50.4	50.1	7.4
Eagle-1.5b+LCD	**0.25**	**51.8**	**51.5**	**10.8**
Pythia-3b	-	-	-	11.1
Eagle-3b	0.28	56.2	55.9	9.7
Eagle-3b+LCD	**0.33**	**58.8**	**58.6**	**14.2**

TABLE III: Ablation studies on TruthfulQA using RWKV Eagle-1.5b in different metrics. LW, LS and EE mean learnable weighting, layer selection and early exit, respectively.

Method		ROUGE-L ↑	BERTScore ↑		Coherence ↑
LW	LS		Recall	F1	
×	–	0.14	48.6	48.3	5.60
×	fixed	0.16	46.2	45.7	5.62
×	JSD	0.19	48.9	48.7	5.67
✓	fixed	0.21	47.4	46.3	6.11
✓	JSD	0.24	50.1	49.7	7.06
✓	EE+JSD	**0.25**	**51.8**	**51.5**	**7.31**

proposed LCD demonstrates superior narrative quality and consistently achieves the highest MAUVE, diversity, and coherence, producing more coherent and diverse stories in text generation when compared with the other baselines including static contrastive decoding using DoLa [18]. Different backbones without contrastive decoding and different model sizes are evaluated in this task.

b) TruthfulQA: TruthfulQA contains 1,791 adversarial questions probing truthfulness under misconceptions. We adopt the comparative settings and evaluate different methods with ROUGE-L and BERTScore covering recall and F1. As shown in left-hand-side of Table II, the proposed LCD yields the best scores across different metrics, surpassing DoLa under RWKV using Eagle-1.5b. LCD as an adaptive contrastive decoding performs better than static contrastive decoding using DoLa as well as the one without contrastive decoding where different backbones with different model sizes are evaluated.

c) GSM8K: We also evaluate different methods on GSM8K, a dataset of 1,317 math word problems, requiring multi-step reasoning. Different models generate the final numeric answers, and the accuracy is measured by exact match. As shown in right-hand-side of Table II, the proposed LCD significantly improves the performance, achieving 10.8% accuracy versus 6.9–7.7% relative to the other methods.

C. Ablation Study

This paper conducts the ablation studies on TruthfulQA with RWKV Eagle-1.5b to examine the role of individual LCD components. As shown in Table III, using a fixed intermediate layer (layer 12) yields only minor gains, while JSD-based dynamic selection improves the performance by leveraging the semantic divergence. Adding the learnable weighting for the fusion of logits further boosts the results on ROUGE-L and BERTScore. The best results are obtained when combining with the learnable weighting based on the early-exit–guided JSD selection (EE+JSD). These findings confirm that both dynamic layer selection and learnable fusion are essential to the improvement by using the proposed method in evaluation of factuality, fluency, and semantic coherence.

V. Conclusions

We have presented the learnable contrastive decoding, a decoding-time framework that applies the locate–edit paradigm of model editing without altering LLM parameters. The proposed LCD identifies the semantically informative intermediate layers via entropy and divergence based signals, then edits the decoding through attention-weighted fusion with the final layer. This solution transforms the inter-layer disagreement into constructive evidence, amplifying the factuality, mitigating the hallucination, and preserving the fluency with minimal overhead. Beyond the practical gains, LCD highlights the link between decoding control and representation-level editing. By showing that factual signals are unevenly distributed across layers and can be adaptively modulated through contrastive fusion, LCD provides an interpretable and reversible mechanism for steering model behavior. Future directions include multi-layer fusion, adaptive early-exit, and integration with retrieval or memory systems to further improve the controllability in an open-ended text generation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] J.-T. Chien and Y.-H. Huang, "Latent semantic and disentangled attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10047–10059, 2024.

[3] J.-T. Chien and Y.-H. Chen, "Learning continuous-time dynamics with attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1906–1918, 2023.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, pp. 1877–1901, 2020.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "LLaMa 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[6] J. Sieber, C. Amo Alonso, A. Didier, M. Zeilinger, and A. Orvieto, "Understanding the differences in foundation models: Attention, state space models, and recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 134534–134566, 2024.

[7] B. Peng, D. Goldstein, Q. G. Anthony, A. Albalak, E. Alcaide, S. Biderman, E. Cheah, T. Ferdinan, K. K. GV, and Hou, "Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence," in *Proc. of Conference on Language Modeling*, 2024.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[9] H.-Y. Lin, Y.-C. Chang, and J.-T. Chien, "Variational dehallucination via retrieval-augmented prompt learning," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 259–271, 2026.

[10] Y.-T. Lin, L.-Y. Zhang, Y.-T. Chen, and J.-T. Chien, "Risk-aware bilingual spoken dialogue for campus mental health support," in *Proc. of AAAI Conference on Artificial Intelligence*, vol. 40, pp. 41628–41630, 2026.

[11] J. Chien and H. Liu, "Contrastive disentanglement learning for empathetic generation," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2025.

[12] J.-T. Chien and Z.-X. Tai, "Reinforced retrieval-augmented generation in large language models," in *Proc. of International Joint Conference on Neural Networks*, pp. 1–7, 2025.

[13] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

[14] X. Shi, I. Padhi, and K. Knight, "Does string-based neural mt learn source syntax?," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, pp. 1526–1534, 2016.

[15] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman, and D. Das, "What do you learn from context? probing for sentence structure in contextualized word representations," in *Proc. of International Conference on Learning Representations*, 2019.

[16] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei, "Knowledge neurons in pretrained transformers," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 8493–8502, 2022.

[17] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17359–17372, 2022.

[18] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "DoLa: Decoding by contrasting layers improves factuality in large language models," in *Proc. of International Conference on Learning Representations*, 2023.

[19] S. Wang, Y. Zhu, H. Liu, Z. Zheng, C. Chen, and J. Li, "Knowledge editing for large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–37, 2024.

[20] V. Mazzia, A. Pedrani, A. Caciolai, K. Rottmann, and D. Bernardi, "A survey on knowledge editing of neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[21] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin, "Deebert: Dynamic early exiting for accelerating bert inference," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 2246–2251, 2020.

[22] Y.-C. Chang, P.-H. F. Chiang, and J.-T. Chien, "Disentangled motion diffusion for long-form voice-driven portrait animation," in *Proc. of International Joint Conference on Neural Networks*, 2026.

[23] X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. B. Hashimoto, L. Zettlemoyer, and M. Lewis, "Contrastive decoding: Open-ended text generation as optimization," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, pp. 12286–12312, 2023.

[24] J. Waldendorf, B. Haddow, and A. Birch, "Contrastive decoding reduces hallucinations in large multilingual machine translation models," in *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2526–2539, 2024.

[25] Y. Zhou, C. Zhou, J. Zhang, J. Li, and M. Zhang, "Alw: Adaptive layer-wise contrastive decoding enhancing reasoning ability in large language models," in *Findings of the Association for Computational Linguistics*, pp. 8506–8524, 2025.

[26] J.-T. Chien and K. Chen, "Outlier-aware contrastive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2026.

[27] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, and M. Burtsev, "Babilong: Testing the limits of llms with long context reasoning-in-a-haystack," *Advances in Neural Information Processing Systems*, vol. 37, pp. 106519–106554, 2024.

[28] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2022.

[29] J.-T. Chien and Y.-A. Chen, "Towards a unified view of adversarial training: A contrastive perspective," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5365–5369, 2024.

[30] J.-T. Chien and P.-C. Huang, "CAPR: Confidence-aware prompt refinement in large language models," in *Proc. of Annual Conference of International Speech Communication Association*, pp. 3264–3268, 2025.

[31] Z. Zeng, P. Chen, S. Liu, H. Jiang, and J. Jia, "MR-GSM8k: A meta-reasoning benchmark for large language model evaluation," in *Proc. of International Conference on Learning Representations*, 2025.

[32] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," in *Proc. of the Annual Meeting of the Association for Computational Linguistics*, 2018.

[33] J.-T. Chien, M. Rohmatillah, and C.-T. Chu, "Strategic optimization for worst-case augmentation and classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 111–123, 2025.

[34] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," in *International Conference on Machine Learning*, pp. 2397–2430, 2023.

[35] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Association for Computational Linguistics, 2004.

[36] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *Proc. of the International Conference on Learning Representations*, 2020.

[37] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, and Z. Harchaoui, "Mauve: Measuring the gap between neural text and human text using divergence frontiers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4816–4828, 2021.

[38] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[39] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proc. of Association for Computational Linguistics*, pp. 6894–6910, 2021.