Novelty Assessment Report

Paper: Spectral Attention Steering for Prompt Highlighting
PDF URL: https://openreview.net/pdf?id=XfLvGIFmAN
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Steering a large language model's attention towards user-specified highlighted text is a critical capability. Existing prompt highlighting methods are incompatible with modern efficient attention mechanisms like Flash Attention due to their reliance on post-hoc matrix editing. We introduce Spectral Editing Key Amplification (SEKA), a training-free steering method that tackles this by directly editing key embeddings before attention computation. SEKA learns universal relevance subspaces offline via spectral decomposition. We extend this to Adaptive SEKA (AdaSEKA), a query-adaptive variant that uses a training-free routing mechanism to dynamically combine multiple expert subspaces based on the prompt's semantic intent. Our experiments show both methods significantly outperform strong baselines on standard steering benchmarks while adding much lower latency and memory overhead, ensuring full compatibility with optimised attention.

Core Task Landscape

This paper addresses: **Steering Attention Towards Highlighted Text in Language Model Prompts**

A total of **18 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Direct Attention Steering Methods**
- **Prompt Engineering and Structural Emphasis**
- **Task-Specific Prompting for Information Extraction**
- **Model Alignment and Training-Based Steering**
- **Uncertainty Quantification and Relevance Assessment**

Complete Taxonomy Tree

- Steering Attention Towards Highlighted Text in Language Model Prompts Survey Taxonomy
- Direct Attention Steering Methods
  - Post-Hoc Attention Matrix Manipulation (3 papers)
  - [3] Tell your model where to attend: Post-hoc attention steering for llms (Zhang QingRu, 2023) View paper
  - [6] Spotlight Your Instructions: Instruction-following with Dynamic Attention Steering (Venkateswaran, 2025) View paper
  - [7] Prompt highlighter: Interactive control for multi-modal llms (Yuechen Zhang, 2024) View paper
  - Embedding-Space Steering Methods ★ (2 papers)
  - [0] Spectral Attention Steering for Prompt Highlighting (Anon et al., 2026) View paper
  - [14] On the Efficiency and Steerability of Self-Attention Mechanism of Large Language Models (Zhang, 2025) View paper
  - Contextual Head Identification and Control (1 papers)
  - [4] Focus directions make your language models pay more attention to relevant contexts (Zhu, 2025) View paper
- Prompt Engineering and Structural Emphasis
  - Input Ordering and Emphasis Strategies (2 papers)
  - [11] Guiding Large Language Models via External Attention Prompting for Scientific Extreme Summarization (Chang Yuan, 2024) View paper
  - [15] It Is Not About What You Say, It Is About How You Say It: A Surprisingly Simple Approach for Improving Reading Comprehension (Sagi Shaier, 2024) View paper
  - Prompt-Guided Interaction Mechanisms (2 papers)
  - [2] Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing (Wang Kai, 2023) View paper
  - [10] Prompt-Guided Dual-Channel Attention Model Predicts Brain Activation from Functional and Structural Profiles (W Huang, 2025) View paper
- Task-Specific Prompting for Information Extraction
  - Event and Relation Extraction Prompting (3 papers)
  - [5] Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction (MA YuBo, 2022) View paper
  - [8] Span-based fine-grained entity-relation extraction via sub-prompts combination (Ning Yu, 2023) View paper
  - [17] Context-Aware Prompt for Generation-based Event Argument Extraction with Diffusion Models (Lei Luo, 2023) View paper
  - Slot Filling with Prompting (1 papers)
  - [16] Zero-Shot Slot Filling with Slot-Prefix Prompting and Attention Relationship Descriptor (Liu, 2023) View paper
- Model Alignment and Training-Based Steering
  - Token-Level Refinement and Alignment (2 papers)
  - [9] Preference-grounded token-level guidance for language model fine-tuning (Yang, 2023) View paper

- [12] Beyond Prompt Engineering: A Reinforced Token-Level Input Refinement for Large Language Models (Huang Guang, 2025) View paper
- Prompt-Tuning and Attention Mechanisms (1 papers)
- [13] On the Role of Attention in Prompt-tuning (Oymak, 2023) View paper

- Uncertainty Quantification and Relevance Assessment
  - Token-Level Uncertainty and Relevance Weighting (1 papers)
  - [1] Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models (Cheng Hao, 2024) View paper
  - Cross-Domain Essay Scoring (1 papers)
  - [18] Automated Cross-prompt Scoring of Essay Traits (Dai Xin-yu, 2021) View paper

Narrative

Core task: steering attention towards highlighted text in language model prompts. The field addresses how to make language models selectively focus on specific portions of their input, a challenge that arises when prompts contain both critical information and distracting context. The taxonomy reveals five main branches: Direct Attention Steering Methods manipulate model internals or embeddings to redirect focus (e.g., Spectral Attention Steering[0], Self-Attention Steerability[14]); Prompt Engineering and Structural Emphasis explores surface-level formatting and instruction design to highlight key spans (e.g., Spotlight Instructions[6], Prompt Highlighter[7]); Task-Specific Prompting for Information Extraction tailors prompts for structured outputs like entity recognition (e.g., PAIE[5], Span-based Extraction[8]); Model Alignment and Training-Based Steering fine-tunes or trains models to respect emphasis cues (e.g., Attention Prompt-tuning[13], Dynamic Prompt Learning[2]); and Uncertainty Quantification and Relevance Assessment evaluates whether models correctly attend to salient information (e.g., Cross-prompt Scoring[18]). These branches reflect a spectrum from inference-time interventions to training-time solutions, and from task-agnostic mechanisms to domain-specific designs.

A particularly active line of work centers on embedding-space and post-hoc interventions that steer attention without retraining, contrasting with prompt-engineering approaches that rely on natural-language markers or structural cues. Spectral Attention Steering[0] sits within the Direct Attention Steering Methods branch, specifically among embedding-space techniques, where it shares conceptual ground with Self-Attention Steerability[14] in manipulating internal representations to amplify highlighted tokens. This contrasts with methods like Post-hoc Attention Steering[3], which intervenes after initial forward passes, and with surface-level strategies such as Prompt Highlighter[7] that use formatting alone. The trade-off revolves around interpretability and deployment complexity: embedding-space methods promise fine-grained control but require access to model internals, while prompt-based methods remain model-agnostic yet may be less reliable across diverse contexts. Open questions include how to balance steering strength with preserving model coherence, and whether training-free interventions can match the robustness of alignment-based approaches like Preference-grounded Guidance[9].

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. On the Efficiency and Steerability of Self-Attention Mechanism of Large Language Models

Authors: Q Zhang | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â□¦ time prompting method that automatically identifies important contexts and explicitly highlights them by attention steeringâ□¦ language models (LLMs), we have a similar need â□□ steering the â□¦

Relationship Analysis

Both papers belong to the Embedding-Space Steering Methods category, focusing on modifying key or query embeddings before attention computation to steer model focus. The candidate paper is a dissertation covering multiple aspects of attention mechanism efficiency and steerability, including a chapter (Chapter 4) on post-hoc attention steering (PASTA) that edits attention matrices after computation, which differs fundamentally from the original paper's approach. The original paper (SEKA) specifically addresses the incompatibility of post-hoc methods like PASTA with Flash Attention by proposing spectral decomposition-based key embedding modifications before attention computation, representing a paradigm shift from the post-hoc approach discussed in the candidate dissertation.

Contributions Analysis

Overall novelty summary. The paper introduces SEKA and AdaSEKA, training-free methods that steer language model attention by editing key embeddings before attention computation. These contributions sit within the Embedding-Space Steering Methods leaf of the taxonomy, which contains only two papers total (including this work). This leaf represents a sparse research direction focused on pre-computation embedding modifications, contrasting with the more populated Post-Hoc Attention Matrix Manipulation leaf (three papers). The single sibling paper in this leaf suggests the embedding-space approach remains relatively underexplored compared to post-hoc interventions, positioning this work in a less crowded area of the attention steering landscape.

The taxonomy reveals that Direct Attention Steering Methods (the parent branch) encompasses three distinct approaches: post-hoc matrix manipulation, embedding-space steering, and contextual head identification. Neighboring branches include Prompt Engineering and Structural Emphasis (which uses surface-level formatting rather than internal modifications) and Model Alignment and Training-Based Steering (which requires fine-tuning). The scope notes clarify that embedding-space methods explicitly exclude post-attention modifications and training-based approaches. AdaSEKA's query-adaptive routing mechanism appears to bridge embedding-space steering with dynamic selection strategies, potentially connecting to concepts in the Contextual Head Identification leaf, though the taxonomy structure keeps these separated.

Among 23 candidates examined across three contributions, none were flagged as clearly refutable. SEKA examined 3 candidates with 0 refutable matches; AdaSEKA examined 10 candidates with 0 refutable; and the KV head selection mechanism examined 10 candidates with 0 refutable. This suggests that within the limited search scope, no prior work was found that directly overlaps with the specific combination of spectral decomposition for key amplification and training-free routing for adaptive subspace selection. The statistics indicate a relatively clean novelty signal, though the search examined only top-K semantic matches rather than an exhaustive literature review.

Based on the limited search scope of 23 candidates, the work appears to occupy a relatively novel position within the sparse embedding-space steering direction. The absence of refutable prior work across all three contributions, combined with the leaf's low paper count, suggests meaningful differentiation from existing approaches. However, this assessment is constrained by the top-K semantic search methodology and does not cover potential overlaps in adjacent fields like representation editing or mechanistic interpretability that may fall outside the taxonomy's scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Spectral Editing Key Amplification (SEKA)

Description: SEKA is a novel training-free framework that steers attention by modifying key vectors before attention scores are calculated, using spectral decomposition to learn universal relevance subspaces offline. This approach is fully compatible with Flash Attention and other optimized attention mechanisms.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Eigendecomposition-Based Spatial-Temporal Attention for Brain Cognitive States Identification

URL: View paper

Brief Assessment

Eigendecomposition Spatial-Temporal[31] applies eigendecomposition-based attention to fMRI brain signals for cognitive state identification, not to language model key embeddings for attention steering. The domains (neuroscience vs. NLP), data modalities (fMRI signals vs. text embeddings), and objectives (brain state classification vs. prompt highlighting) are fundamentally different.

---

2. A Three-Channel Improved SE Attention Mechanism Network Based on SVD for High-Order Signal Modulation Recognition

URL: View paper

Brief Assessment

Three-Channel SVD[30] applies SVD for signal denoising in communication modulation recognition, not for training-free attention steering in language models. The domains (signal processing vs. LLM attention mechanisms) and applications are fundamentally different.

---

3. Stylehumanclip: Text-guided garment manipulation for stylegan-human

URL: View paper

Brief Assessment

Stylehumanclip[29] focuses on text-guided garment manipulation in StyleGAN-Human using attention-based latent code mappers for image synthesis. It does not address training-free attention steering in language models through spectral decomposition of key embeddings, which is the core contribution of SEKA.

Contribution 2: Adaptive SEKA (AdaSEKA)

Description: AdaSEKA extends SEKA by learning multiple domain-specific expert projections and using a query-adaptive routing mechanism to dynamically select and combine these experts at inference time, reducing the need for manual hyperparameter tuning across different tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Adaptive Expert Learning for Hyperspectral and Multispectral Image Fusion

URL: View paper

Brief Assessment

Adaptive Expert Fusion[23] addresses hyperspectral/multispectral image fusion with spatial-spectral expert routing, not attention steering in language models. The domains and technical mechanisms are fundamentally different.

---

2. GateTS: Versatile and Efficient Forecasting via Attention-Inspired routed Mixture-of-Experts

URL: View paper

Brief Assessment

GateTS[26] focuses on mixture-of-experts routing for time-series forecasting tasks, not on attention steering mechanisms for language models. The routing mechanism in GateTS[26] uses attention-inspired gating for expert selection in forecasting contexts, which is architecturally and functionally distinct from AdaSEKA's query-adaptive routing for steering attention in LLMs toward highlighted text spans.

---

3. Hierarchical Multi-Stage Attention and Dynamic Expert Routing for Explainable Gastrointestinal Disease Diagnosis.

URL: View paper

Brief Assessment

Hierarchical Expert Routing[28] focuses on gastrointestinal disease diagnosis using hierarchical multi-stage attention for medical imaging, not query-adaptive routing for attention steering in language models. The domains and technical objectives are fundamentally different.

---

4. Improving Routing in Sparse Mixture of Experts with Graph of Tokens

URL: View paper

Brief Assessment

Graph of Tokens[22] addresses routing in sparse mixture of experts for token-level expert selection, not query-adaptive attention steering with multiple expert subspaces for key embeddings in transformers.

---

5. LoRA-Mixer: Coordinate Modular LoRA Experts Through Serial Attention Routing

URL: View paper

Brief Assessment

LoRA-Mixer[21] focuses on routing multiple LoRA experts for multi-task adaptation in LLMs, not on query-adaptive attention steering mechanisms for prompt highlighting. The technical domains and objectives differ fundamentally.

---

6. AMC: Adaptive Multi-expert Collaborative Network for Text-guided Image Retrieval

URL: View paper

Brief Assessment

AMC[20] addresses text-guided image retrieval using dynamic router mechanisms for multi-expert collaboration in multimodal fusion, not query-adaptive attention steering in language models for prompt highlighting.

---

7. A Survey on Fine-Grained Multimodal Large Language Models

URL: View paper

Brief Assessment

Fine-Grained Multimodal Survey[27] focuses on multimodal large language models and their fine-grained understanding capabilities. The candidate does not address query-adaptive routing mechanisms for attention steering in language models, which is the core contribution of AdaSEKA.

---

8. Mr. DETR++: Instructive Multi-Route Training for Detection Transformers with Mixture-of-Experts

URL: View paper

Brief Assessment

Mr DETR++[19] focuses on multi-route training for detection transformers with mixture-of-experts for object detection tasks, not on query-adaptive routing mechanisms for attention steering in language models or prompt highlighting applications.

---

9. FAME: Adaptive Functional Attention with Expert Routing for Function-on-Function Regression

URL: View paper

Brief Assessment

FAME[25] addresses function-on-function regression using mixture-of-experts for functional data analysis, not attention steering in language models. The routing mechanism in FAME operates on functional representations for regression tasks, whereas AdaSEKA routes expert projections for attention steering in LLMs based on query semantics.

---

10. Multilingual Routing in Mixture-of-Experts

URL: View paper

Brief Assessment

Multilingual Routing[24] focuses on routing mechanisms in Mixture-of-Experts architectures for multilingual processing, not on query-adaptive attention steering through expert subspace projections for prompt highlighting tasks.

Contribution 3: KV head selection mechanism

Description: A selective mechanism that identifies and applies attention steering only to key-value heads that are naturally sensitive to prompt relevance, based on empirical measurements of embedding shifts between relevant and irrelevant contexts across layers and heads.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Selective Attention: Enhancing Transformer through Principled Context Control

URL: View paper

Brief Assessment

Selective Attention[32] focuses on temperature scaling for attention sparsity control across all queries uniformly, not on identifying specific key-value heads based on embedding sensitivity to context relevance as in the original paper's selective mechanism.

---

2. S2-Attention: Hardware-Aware Context Sharding Among Attention Heads

URL: View paper

Brief Assessment

S2-Attention[39] focuses on hardware-aware context sharding across attention heads for efficiency, not on selective mechanisms based on embedding sensitivity to context relevance. The candidate does not address query-adaptive head selection based on relevance signals.

---

3. Semantic latency mapping of contextual vector embeddings in transformer-based models

URL: View paper

Brief Assessment

Semantic Latency Mapping[33] focuses on quantifying temporal delays in semantic development across transformer layers, not on selective attention steering based on embedding sensitivity to context relevance. The candidate does not address selective mechanisms for attention steering or key-value head selection based on relevance sensitivity.

---

4. Rewards teach visual selective attention

URL: View paper

Brief Assessment

Visual Selective Attention[35] focuses on visual attention mechanisms in perception tasks, not on transformer key-value head selection for language model attention steering based on embedding sensitivity to prompt relevance.

---

5. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue

URL: View paper

Brief Assessment

Copy-augmented Dialogue[37] focuses on task-oriented dialogue generation using attention mechanisms for copying entities from dialogue context, not on selective attention head mechanisms based on embedding sensitivity to context relevance in LLMs.

---

6. Elementwise Language Representation

URL: View paper

Brief Assessment

Elementwise Representation[40] focuses on character-level language representation through element-wise embeddings and does not address selective attention head mechanisms based on embedding sensitivity to context relevance. The candidate paper's approach of concatenating character embeddings differs fundamentally from the original paper's KV head selection based on empirical measurements of embedding shifts.

---

7. Focus directions make your language models pay more attention to relevant contexts

URL: View paper

Brief Assessment

Focus Directions[4] identifies 'contextual heads' based on attention distribution to relevant contexts during generation, not on embedding sensitivity measured via shifts between relevant/irrelevant contexts. The original paper's mechanism measures embedding shifts across

layers/heads offline, while Focus Directions[4] uses a contextual scoring method during inference to identify heads that pay most attention to relevant contexts.

---

8. Neural re-contextualization for dynamic semantic control in large language models

URL: View paper

Brief Assessment

Neural Re-contextualization[34] focuses on adaptive neural nodes and probabilistic modeling for semantic recalibration across entire model architectures, not on selective attention head mechanisms based on embedding sensitivity to context relevance as described in the original paper's KV head selection.

---

9. Neural modulation for dynamic semantic convergence in large language models: A technical examination

URL: View paper

Brief Assessment

Dynamic Semantic Convergence[36] discusses selective attention within modular layers for thematic relevance, but does not describe a mechanism based on empirical measurements of embedding shifts between relevant and irrelevant contexts across specific key-value heads as in the original paper.

---

10. Unveiling Simplicities of Attention: Adaptive Long-Context Head Identification

URL: View paper

Brief Assessment

Adaptive Long-Context Head[38] focuses on query-adaptive classification of attention heads for long-context processing efficiency, not on selective mechanisms based on embedding sensitivity to prompt relevance. The original paper's mechanism identifies heads sensitive to highlighted/relevant context through embedding shifts, while the candidate identifies heads that can use local vs. long-context windows based on attention score distributions and second-order statistics of keys.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Spectral Attention Steering for Prompt Highlighting View paper
- [1] Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models View paper
- [2] Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing View paper
- [3] Tell your model where to attend: Post-hoc attention steering for llms View paper
- [4] Focus directions make your language models pay more attention to relevant contexts View paper
- [5] Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction View paper
- [6] Spotlight Your Instructions: Instruction-following with Dynamic Attention Steering View paper
- [7] Prompt highlighter: Interactive control for multi-modal llms View paper
- [8] Span-based fine-grained entity-relation extraction via sub-prompts combination View paper
- [9] Preference-grounded token-level guidance for language model fine-tuning View paper
- [10] Prompt-Guided Dual-Channel Attention Model Predicts Brain Activation from Functional and Structural Profiles View paper
- [11] Guiding Large Language Models via External Attention Prompting for Scientific Extreme Summarization View paper
- [12] Beyond Prompt Engineering: A Reinforced Token-Level Input Refinement for Large Language Models View paper
- [13] On the Role of Attention in Prompt-tuning View paper
- [14] On the Efficiency and Steerability of Self-Attention Mechanism of Large Language Models View paper
- [15] It Is Not About What You Say, It Is About How You Say It: A Surprisingly Simple Approach for Improving Reading Comprehension View paper
- [16] Zero-Shot Slot Filling with Slot-Prefix Prompting and Attention Relationship Descriptor View paper
- [17] Context-Aware Prompt for Generation-based Event Argument Extraction with Diffusion Models View paper
- [18] Automated Cross-prompt Scoring of Essay Traits View paper
- [19] Mr. DETR++: Instructive Multi-Route Training for Detection Transformers with Mixture-of-Experts View paper
- [20] AMC: Adaptive Multi-expert Collaborative Network for Text-guided Image Retrieval View paper
- [21] LoRA-Mixer: Coordinate Modular LoRA Experts Through Serial Attention Routing View paper
- [22] Improving Routing in Sparse Mixture of Experts with Graph of Tokens View paper
- [23] Adaptive Expert Learning for Hyperspectral and Multispectral Image Fusion View paper
- [24] Multilingual Routing in Mixture-of-Experts View paper
- [25] FAME: Adaptive Functional Attention with Expert Routing for Function-on-Function Regression View paper
- [26] GateTS: Versatile and Efficient Forecasting via Attention-Inspired routed Mixture-of-Experts View paper
- [27] A Survey on Fine-Grained Multimodal Large Language Models View paper
- [28] Hierarchical Multi-Stage Attention and Dynamic Expert Routing for Explainable Gastrointestinal Disease Diagnosis. View paper
- [29] Stylehumanclip: Text-guided garment manipulation for stylegan-human View paper
- [30] A Three-Channel Improved SE Attention Mechanism Network Based on SVD for High-Order Signal Modulation Recognition View paper
- [31] Eigendecomposition-Based Spatial-Temporal Attention for Brain Cognitive States Identification View paper
- [32] Selective Attention: Enhancing Transformer through Principled Context Control View paper
- [33] Semantic latency mapping of contextual vector embeddings in transformer-based models View paper
- [34] Neural re-contextualization for dynamic semantic control in large language models View paper
- [35] Rewards teach visual selective attention View paper
- [36] Neural modulation for dynamic semantic convergence in large language models: A technical examination View paper
- [37] A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue View paper
- [38] Unveiling Simplicities of Attention: Adaptive Long-Context Head Identification View paper
- [39] S2-Attention: Hardware-Aware Context Sharding Among Attention Heads View paper
- [40] Elementwise Language Representation View paper