

# SIGHT: Gaze-Guided Attention-Head Selection for Head-Sparse PEFT

Xinyu Gao<sup>1</sup>, Shaonan Wang<sup>2,3</sup>, Jiajie Zou<sup>1</sup>, Li Zhao<sup>1\*</sup>, Nai Ding<sup>1</sup>

<sup>1</sup>College of Biomedical Engineering & Instrument Science, Zhejiang University

<sup>2</sup>Division of Artificial Intelligence and the Humanities, Hong Kong Polytechnic University

<sup>3</sup>Department of Language Science and Technology, Hong Kong Polytechnic University  
lizhaomri@zju.edu.cn

## Abstract

Parameter-efficient fine-tuning (PEFT) reduces trainable parameters, yet head-sparse PEFT still requires deciding which attention heads should receive updates. We introduce SIGHT, *Sparse Identification of Gaze-aligned Heads for Tuning*, a gaze-guided head selector for head-sparse PEFT. Using eye-tracking data during reading comprehension tests, SIGHT ranks heads by correlating word-level gaze metrics with the attention heads of large language models. Specifically, the comprehension question appears either before or after the passage so that the readers either read the passage with a clear goal or for general understanding. The eye-tracking data from these two conditions are both correlated with attention heads and are fused by an offline-calibrated layer-wise gating module. When applied to Qwen3-based models, updating only the top 10% SIGHT-selected heads generally outperforms other head selectors and all-head PEFT references on reasoning and syntax benchmarks. These results suggest that gaze information during human reading can move beyond post hoc interpretation and serve as an external criterion for allocating adaptation capacity.

## 1 Introduction

Parameter-efficient fine-tuning (PEFT) has made large language models (LLMs) easier to adapt by updating small modules, low-rank factors, activation scales, or localized representations instead of the full checkpoint (Albert et al., 2025; Hayou et al., 2024; Wu et al., 2024; Yin et al., 2024). Yet PEFT alone does not specify where adaptation should occur. As PEFT becomes more localized, from modules to individual attention heads, adaptation becomes an allocation problem: under a fixed update budget, which components should receive task-specific updates? This question is explicit in

head-sparse PEFT, where only selected heads are allowed to adapt.

Existing head selectors usually answer this question with model-internal signals, such as attention mass (Michel et al., 2019), causal effects (Arora et al., 2024), or feature activations (Marks et al., 2025). Although more faithful localization methods, including CausalGym (Arora et al., 2024), Causal Head Gating (Nam et al., 2025), and relevance patching (Jafari et al., 2025), address known limits of attention-only explanations (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegrefe and Pinter, 2019), they still mainly identify heads involved in the current frozen computation, rather than heads best suited for future adaptation under a sparse budget.

This limitation motivates our research question: can human reading behavior better allocate head-sparse PEFT capacity than model-internal signals alone? Human reading behavior provides one such external source: eye movements record how readers distribute processing effort over words during comprehension. Gaze metrics such as first-pass dwell, reread counts, and total dwell have been linked to integration difficulty, reanalysis, and reading goals (Rayner et al., 2000; Kaakinen et al., 2002; Warren et al., 2009). Recent work further shows that reading behavior depends on task goals, context, and structural integration rather than on a single universal attention pattern (Oh et al., 2024; Bolliger et al., 2025; Hadar et al., 2026). We therefore use gaze not as evidence of cognitive equivalence between humans and LLMs, but as behavioral evidence of processing effort under matched reading conditions.

We propose SIGHT, *Sparse Identification of Gaze-aligned Heads for Tuning*, a head selector that uses human gaze to allocate head-sparse PEFT capacity. SIGHT uses two reading orders: question-first reading reflects goal-directed evidence search, while passage-first reading reflects general passage

\*Corresponding author: Li Zhao

comprehension. For each head, SIGHT separately computes a question-first and a passage-first gaze-attention alignment score by comparing passage-word attention with the corresponding gaze metrics. It then fuses the two condition scores with a layer-wise gate calibrated only on gaze data and ranks heads before downstream PEFT.

Across three Qwen3-based backbones, two PEFT operators, and four reasoning/syntax benchmarks, updating only the top 10% SIGHT-selected heads generally outperforms gaze-free sparse head selectors and matches or exceeds all-head PEFT references in most settings. Deletion and ablation analyses show that SIGHT-selected heads carry predictive support before head-sparse PEFT and that the gains depend on condition-matched gaze-attention alignment rather than head-sparse PEFT alone. These results suggest that human reading signals can serve as an external criterion to allocate sparse adaptation capacity.

Our contributions are as follows.

- We frame head-sparse PEFT as an allocation problem and propose human reading behavior as an external selection signal complementary to model-internal head-importance scores.
- We introduce SIGHT, which ranks heads by condition-matched gaze-attention alignment from two reading orders, using a layer-wise fusion gate calibrated only on gaze data.
- We evaluate SIGHT across three Qwen3 backbones, two head-sparse PEFT operators, and four reasoning/syntax benchmarks, with deletion, condition-mismatch, gaze-shuffling, and surface-control analyses testing the role of matched human-reading signals.

## 2 Related Work

### 2.1 Human Reading Signals and Model Behavior

Recent NLP work increasingly uses human reading behavior as a behavioral signal to compare humans and language models, rather than only as psycholinguistic evidence (Hollenstein et al., 2020). This direction has been strengthened by eye-tracking datasets that move beyond natural reading to task-oriented settings such as reading comprehension, machine-generated text reading, and information seeking (Hollenstein et al., 2020; Ribeiro et al., 2024; Bolliger et al., 2025; Hadar et al., 2026). These resources make it possible to test whether model-derived quantities align with human pro-

cessing effort under different reading goals and comprehension demands.

Previous work shows that such model-human alignment is conditional. The relation between model predictions and gaze measures varies with lexical frequency, information-seeking goals, and structural integration demands (Oh et al., 2024; Clark et al., 2025; Gruteke Klein et al., 2024; Bianchi et al., 2025). Other studies use gaze as supervision or guidance for downstream systems, including code summarization, response reward modeling, and chart understanding (Zhang et al., 2024; López-Cardona et al., 2025; Rekimoto, 2025). Recent vision-language work uses gaze in different roles: *Gaze-VLM* (Pani and Yang, 2025) regularizes visual attention during task training, whereas *GazeVLM* (Mathew et al., 2026) treats gaze-target and gaze-object understanding as downstream prediction tasks. These findings suggest that gaze can provide behavioral information not fully captured by text labels or model-internal scores. In contrast to these approaches and work that trains models to imitate gaze or uses gaze as an auxiliary loss, we use reading gaze only as a pre-PEFT selection signal for ranking attention heads in a text-only LLM.

### 2.2 Head Localization and PEFT

Work on attention heads has shown that heads differ substantially in function and importance. Some heads track positional, syntactic, or task-relevant patterns, while many can be pruned with limited immediate performance loss (Michel et al., 2019; Voita et al., 2019). This has motivated a range of methods to rank or localize important heads. Early work often relied on attention magnitude or pruning sensitivity, but the debate on attention-explanation cautions against treating attention weights alone as causal evidence (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegraffe and Pinter, 2019). Later methods therefore use stronger model-internal signals, including relevance propagation (Chefer et al., 2021), relevance patching (Jafari et al., 2025), causal interventions and head gating (Arora et al., 2024; Nam et al., 2025), and sparse feature circuits (Marks et al., 2025). These methods make head-level analysis more precise, but they primarily estimate which components matter for the behavior of a frozen model.

Head-sparse PEFT turns this analysis problem into an allocation problem. PEFT methods such

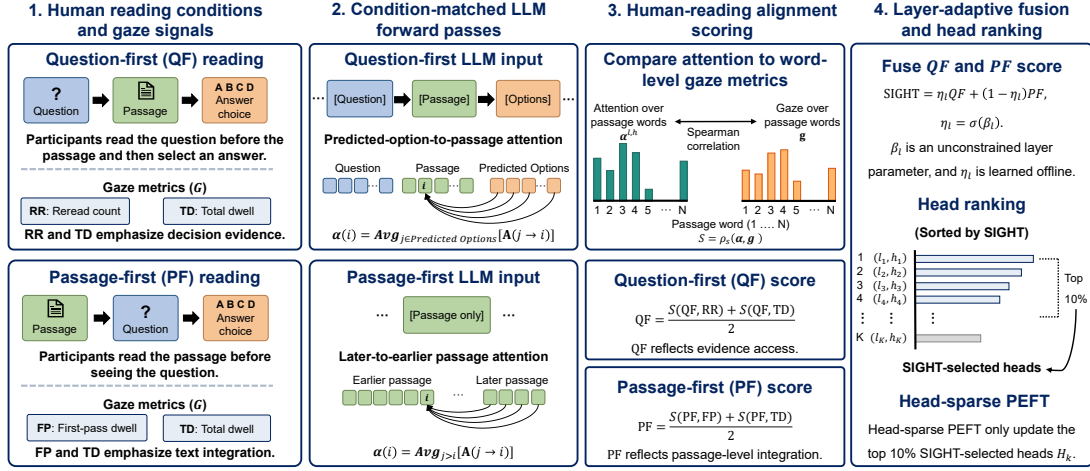


Figure 1: Overview of SIGHT. The method uses two reading conditions, extracts condition-matched attention vectors, scores heads by word-level gaze-attention alignment, and uses an offline-calibrated layer-wise gating module to rank heads for head-sparse PEFT.

as adapters (Houlsby et al., 2019), prompt tuning (Lester et al., 2021), LoRA and its variants (Hu et al., 2022; Hayou et al., 2024; Albert et al., 2025), and (IA)<sup>3</sup> (Liu et al., 2022) reduce adaptation cost by restricting trainable parameters. When PEFT is applied only to selected attention heads, one must decide not only how to adapt the model, but where to place the limited trainable capacity (Michel et al., 2019; Chen et al., 2025). Existing head-sparse PEFT selectors typically answer this question by reusing model-internal importance signals, such as attention magnitude, pruning sensitivity, relevance scores, or causal effects (Michel et al., 2019; Voita et al., 2019; Chefer et al., 2021; Arora et al., 2024; Nam et al., 2025).

This reuse of internal importance scores assumes that a head’s current contribution to the frozen model is a reliable proxy for its usefulness as an adaptation target. We argue that the two properties are related but not identical. A head may strongly affect a frozen prediction while offering little useful capacity for downstream adaptation, whereas another head may better match the behavioral demands of the target task. SIGHT addresses this gap by complementing model-internal importance with an external behavioral signal: condition-matched human reading behavior.

### 3 Method

As shown in Figure 1, SIGHT takes LLM attention patterns and eye-tracking data from the same reading-comprehension items as input, and returns a ranking over attention heads. The resulting rank-

ing is used for head-sparse PEFT.

#### 3.1 Eye-Tracking Data

We use the reading-comprehension eye-tracking dataset of Zou et al. (2023), which contains 800 high-school-level English multiple-choice items adapted from RACE (Lai et al., 2017). Each item contains a passage, a question, and four answer options, exactly one of which is correct. Eye movements were collected from 25 participants under two reading orders: question-first and passage-first. These orders induce different reading goals, allowing SIGHT to derive gaze-attention alignment signals for evidence search and general passage comprehension.

For each passage word, we average each gaze metric across participants within the same reading order, yielding word-level first-pass dwell (FP), reread count (RR), and total dwell (TD) measures for each item and condition. In the question-first (QF) condition, the reader knows the question before reading the passage; SIGHT therefore uses RR and TD to capture rereading and cumulative processing during evidence search. In the passage-first (PF) condition, the reader processes the passage before seeing the question; SIGHT uses FP and TD to capture early processing and cumulative allocation during general comprehension. These pairings follow standard eye-movement measures of early processing, rereading, and cumulative allocation (Rayner, 1998; Warren et al., 2009).

### 3.2 Condition-matched Gaze-attention Alignment

Let  $M$  denote the LLM and let  $z$  index a reading-comprehension item. For question-first reading, the model input is  $[Q_z, P_z, O_z]$ , where  $Q_z$  is the question,  $P_z$  is the passage, and  $O_z$  contains the answer options. For passage-first reading, the model input is  $[P_z]$  alone. These settings correspond to two different comprehension processes in human reading.

Given the input for the reading condition  $c \in \{\text{QF}, \text{PF}\}$ , a forward pass through  $M$  yields the head-specific attention tensor:

$$\mathbf{A}_z^{l,h,M}(c) = \text{softmax} \left( \frac{\mathbf{Q}_z^{l,h,M}(c) \mathbf{K}_z^{l,h,M}(c)^\top}{\sqrt{d_h}} \right) \quad (1)$$

where  $l$  indexes layers,  $h$  indexes heads, and  $d_h$  is the head dimension. We write  $\mathbf{A}_{z,t \rightarrow s}^{l,h,M}(c)$  for the attention weight from target token  $t$  to the source token  $s$ . To align token-level attention with word-level gaze metrics, we aggregate attention over passage words by summing attention mass over source subword tokens.

For the question-first condition, SIGHT extracts predicted-option-to-passage attention, where the target side is the model-predicted answer option. Let  $\hat{o}_z$  be the model-argmax option for item  $z$ , and let  $\hat{O}_z$  denote its option token positions in the input prompt. Using word-level attention  $\bar{\mathbf{A}}_{z,t \rightarrow i}^{l,h,M}(\text{QF})$ , the question-first attention vector is then obtained by averaging over the predicted-option tokens:

$$\mathbf{a}_z^{l,h,M}(\text{QF}) = \left[ \frac{1}{|\hat{O}_z|} \sum_{t \in \hat{O}_z} \bar{\mathbf{A}}_{z,t \rightarrow i}^{l,h,M}(\text{QF}) \right]_{i=1}^{n_z}, \quad (2)$$

where  $n_z$  is the number of passage words. This vector  $\mathbf{a}_z^{l,h,M}(\text{QF})$  reflects the passage words that are accessed when the model evaluates the selected answer.

For the passage-first condition, SIGHT extracts later-to-earlier passage attention. Let  $\mathcal{R}_{z,i}$  contain the passage token positions that occur after the  $i$ -th passage word. Using the same word-level attention  $\bar{\mathbf{A}}_{z,t \rightarrow i}^{l,h,M}(\text{PF})$ , the passage-first attention vector is

$$\mathbf{a}_z^{l,h,M}(\text{PF}) = \left[ \frac{1}{|\mathcal{R}_{z,i}|} \sum_{t \in \mathcal{R}_{z,i}} \bar{\mathbf{A}}_{z,t \rightarrow i}^{l,h,M}(\text{PF}) \right]_{i=1}^{n_z}, \quad (3)$$

where the final position is excluded from the subsequent Spearman correlation calculation within the item because  $\mathcal{R}_{z,n_z} = \emptyset$ . And  $\mathbf{a}_z^{l,h,M}(\text{PF})$  captures which earlier passage words are re-accessed by later passage context during passage-level integration.

For the gaze metric  $m \in \{\text{FP}, \text{RR}, \text{TD}\}$  and the reading condition  $c \in \{\text{QF}, \text{PF}\}$ , let  $\mathcal{Z}_c$  denote the set of reading comprehension test items in condition  $c$ , and  $\mathbf{g}_z(c, m)$  denote the vector of word-level gaze metrics for item  $z$ . The gaze-attention alignment score is the average within-item Spearman rank correlation between model attention and human gaze metrics.

$$S^{l,h,M}(c, m) = \frac{1}{|\mathcal{Z}_c|} \sum_{z \in \mathcal{Z}_c} \rho_s \left( \mathbf{a}_z^{l,h,M}(c), \mathbf{g}_z(c, m) \right) \quad (4)$$

where  $\rho_s(\cdot, \cdot)$  denotes Spearman rank correlation.

Following Section 3.1, the question-first score averages the reread count and total dwell:

$$\text{QF}^{l,h,M} = \frac{S^{l,h,M}(\text{QF}, \text{RR}) + S^{l,h,M}(\text{QF}, \text{TD})}{2} \quad (5)$$

And the passage-first score averages the first-pass dwell and total dwell:

$$\text{PF}^{l,h,M} = \frac{S^{l,h,M}(\text{PF}, \text{FP}) + S^{l,h,M}(\text{PF}, \text{TD})}{2} \quad (6)$$

SIGHT fuses the two condition-specific scores with an offline-calibrated layer-wise gate:

$$\text{SIGHT}^{l,h,M} = \eta_l \text{QF}^{l,h,M} + (1 - \eta_l) \text{PF}^{l,h,M} \\ \eta_l = \sigma(\beta_l). \quad (7)$$

Here,  $\beta_l$  is an unconstrained layer parameter,  $\sigma$  maps it to  $[0, 1]$ , and a larger  $\eta_l$  gives more weight to the question-first score. We learn  $\eta_l$  offline using only gaze-attention alignment scores. Specifically, we randomly sample 30% of the gaze-scoring items as a gate-calibration subset and optimize the layer-wise QF/PF mixture on this subset, with a small smoothness regularizer across adjacent layers. After calibration, the gate is fixed for downstream head-sparse PEFT, and the heads are ranked by their calibrated  $\text{SIGHT}^{l,h,M}$  scores. Full calibration details are provided in Appendix A.

### 3.3 Head-sparse PEFT with Selected Heads

Given a head rank, we select the top  $k$  percent of heads and denote the set by  $H_k$ . Head-sparse PEFT

| Role             | Dataset   | #Tune / Rank | #Test  |
|------------------|-----------|--------------|--------|
| Head-sparse PEFT | UltraChat | 207,865      | /      |
| Eval             | MMLU      | 99,842       | 14,042 |
| Eval             | CSQA      | 9,741        | 1,221  |
| Eval             | BLiMP     | 33,500       | 33,500 |
| Eval             | BB Syn.   | 143          | 143    |

Table 1: Dataset sample counts used in our experiments.

operators are attached only to the attention heads in  $H_k$ . We instantiate SIGHT with two head-sparse PEFT operators, Low-Rank Adaptation (LoRA, [Hu et al., 2022](#)) and Infused Adapter by Inhibiting and Amplifying Inner Activations ((IA)<sup>3</sup>, [Liu et al., 2022](#)). LoRA adds low-rank update modules to the query, key, value, and output projection slices of selected heads, and (IA)<sup>3</sup> learns multiplicative rescaling vectors for the activations associated with selected heads. Across both operators, SIGHT uses the selected head set  $H_k$ . Implementation details about head-to-slice mapping are provided in Appendix B. This construction makes SIGHT a selector for head-sparse PEFT rather than a gaze-supervised training objective.

## 4 Experimental Setup

### 4.1 Models, Datasets, and Baselines

We use Qwen3-8B, Qwen3-4B, and Qwen3-1.7B ([Yang et al., 2025](#)) as the main backbones. UltraChat ([Ding et al., 2023](#)) is used for instruction tuning, and evaluation covers two reasoning benchmarks, MMLU ([Hendrycks et al., 2021a,b](#)) and CommonsenseQA (CSQA; [Talmor et al., 2019](#)), and two syntax benchmarks, BLiMP ([Warstadt et al., 2020](#)) and a syntax subset of BIG-Bench tasks (BB Syn.; [Srivastava et al., 2023](#)).

Table 1 reports the sample counts actually used in our experiments: for UltraChat, #Tune/Rank denotes instruction-tuning examples; for evaluation datasets, it denotes examples used for head ranking, while #Test denotes examples used only for final evaluation. CSQA is evaluated on its validation split because official test labels are unavailable. For BLiMP and BB Syn., which lack official training splits, we split the labeled set into disjoint ranking and evaluation halves.

We compare SIGHT with seven head-selection baselines: Causal Head Gating (CHG; [Nam et al., 2025](#)), Layer-wise Relevance Propagation (LRP; [Chefer et al., 2021](#); [Jafari et al., 2025](#)), Attention Localization and Pruning Strategy (ALPS; [Chen et al., 2025](#)), Layer Consistency selection (LC), Magnitude, Random, and LayerMatched-

| Tuning family | Trainable unit      | Main hyperparameters      | LR   | Epoch |
|---------------|---------------------|---------------------------|------|-------|
| LoRA          | Q/K/V/O head slices | rank 8<br>alpha 16        | 2e-4 | 1     |
| IA3           | selected-head K/V   | scale vectors<br>init 1.0 | 8e-4 | 1     |

Table 2: Hyperparameters for head-sparse PEFT operators.

Random (LM-Random). CHG, LRP, and ALPS use each evaluation dataset’s #Tune/Rank split to compute head rankings. LC selects heads with a uniform layer-wise budget, Magnitude uses passage-word attention mass, Random samples heads uniformly from all head, and LM-Random controls for SIGHT’s layer distribution by sampling within the same layer-wise budget.

### 4.2 Head-sparse PEFT and Evaluation

Unless otherwise stated, all head-sparse PEFT methods select the top 10% of attention heads, following prior work that reports a strong efficiency-performance trade-off at this budget ([Chen et al., 2025](#)). To isolate the effect of head ranking, all head-sparse PEFT methods use the same training data, backbone, budget, and optimization settings. For each head-sparse PEFT operator, we also include an all-head reference under the same training configuration. Table 2 summarizes the trainable units and operator-specific hyperparameters; all runs use a per-device batch size of 2 and a maximum sequence length of 1024. Further details on hyperparameter search, budget matching, and parameter scale are provided in Appendix C.

We evaluate all models with log-probability scoring, without task-specific classifiers, in a zero-shot inference setting with no in-context demonstrations or task-specific exemplars. For MMLU and CSQA, each candidate option  $y_j$  is scored by length-normalized conditional log probability under the evaluation prompt  $x$ :

$$\text{score}(y_j | x) = \frac{1}{|y_j|} \sum_{t=1}^{|y_j|} \log p(y_{j,t} | x, y_{j,<t}). \quad (8)$$

The predicted answer is the option with the highest normalized score. For BLiMP, we use minimal-pair acceptability: an example is correct when the grammatical sentence has higher log probability than the ungrammatical one, i.e.,  $\Delta_{\text{BLiMP}} = \log p(s_{\text{good}} | x) - \log p(s_{\text{bad}} | x) > 0$ . For BB Syn., an instance is correct only when every gold answer in the sequence receives the highest log probability among

| Model              | MMLU                     | CSQA                     | BLiMP                    | BB Syn.                  |
|--------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Qwen3-8B           | 71.59                    | 80.26                    | 76.58                    | 39.51                    |
| SIGHT-LoRA 10%     | <b>76.13<sup>†</sup></b> | <b>84.47<sup>†</sup></b> | <b>81.86<sup>†</sup></b> | <b>63.03<sup>†</sup></b> |
| CHG-LoRA 10%       | 73.48                    | 81.74                    | 79.40                    | 37.20                    |
| LRP-LoRA 10%       | 73.61                    | 82.96                    | 76.88                    | 58.18                    |
| ALPS-LoRA 10%      | 74.68                    | 82.95                    | 80.23                    | 61.47                    |
| LC-LoRA 10%        | 74.52                    | 82.46                    | 79.21                    | 57.63                    |
| Magnitude-LoRA 10% | 73.09                    | 81.33                    | 78.94                    | 54.34                    |
| Random-LoRA 10%    | 74.40                    | 82.88                    | 78.64                    | 56.98                    |
| LM-Random-LoRA 10% | 74.62                    | 83.05                    | 78.81                    | 57.23                    |
| All-head LoRA      | 74.53                    | 81.98                    | 79.20                    | 61.47                    |

Table 3: Downstream accuracy under LoRA with different head-selection strategies on Qwen3-8B. Bold indicates the best mean score in each column. <sup>†</sup> indicates that SIGHT is significantly better than every comparator in the same column under the test defined in Section 4.3.

its candidates. Prompt templates and licenses are provided in Appendix D.

### 4.3 Statistical Test

We report mean accuracy over five training seeds. For Random and LM-Random, each seed-level score averages 10 sampled head sets. We test significance with a paired hierarchical bootstrap using 10,000 resamples over training seeds and evaluation examples, preserving paired predictions between SIGHT and each comparator and resampling random head sets within each seed. We apply Holm correction across all pairwise comparisons within each column. A dagger indicates that SIGHT is significantly better than every comparator in that column with Holm-corrected  $p < 0.05$ .

## 5 Results

The results show that human gaze metrics provide a useful head-selection signal for head-sparse PEFT. We first compare SIGHT with other head selectors under LoRA and (IA)<sup>3</sup>, and use deletion and ablations to examine functional support and gaze-attention alignment.

### 5.1 Head-sparse PEFT

Table 3 compares head-selection strategies under LoRA with a fixed 10% head budget on Qwen3-8B. SIGHT achieves the best accuracy on all four benchmarks, outperforming the strongest non-SIGHT sparse selector by 1.45 points on MMLU, 1.42 on CSQA, 1.63 on BLiMP, and 1.56 on BB Syn. This indicates that SIGHT provides a signal beyond model-internal sparse-selection criteria. Since LM-Random controls for SIGHT’s layer-wise selection distribution, SIGHT’s gains cannot be explained solely by favorable layer choices. In-

| Backbone   | Operator | Setting           | MMLU                     | CSQA                     | BLiMP                    | BB Syn.                  |
|------------|----------|-------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Qwen3-8B   | LoRA     | Base              | 71.59                    | 80.26                    | 76.58                    | 39.51                    |
|            |          | SIGHT 10%         | <b>76.13<sup>†</sup></b> | <b>84.47<sup>†</sup></b> | <b>81.86<sup>†</sup></b> | <b>63.03<sup>†</sup></b> |
|            |          | All-head          | 74.53                    | 81.98                    | 79.20                    | 61.47                    |
|            |          | (IA) <sup>3</sup> | <b>74.65<sup>†</sup></b> | <b>82.41<sup>†</sup></b> | 75.94                    | <b>57.28<sup>†</sup></b> |
| Qwen3-4B   | LoRA     | Base              | 71.33                    | 77.54                    | 80.17                    | 29.65                    |
|            |          | SIGHT 10%         | <b>74.02<sup>†</sup></b> | <b>79.14<sup>†</sup></b> | <b>81.34<sup>†</sup></b> | <b>54.71<sup>†</sup></b> |
|            |          | All-head          | 73.45                    | 78.32                    | 79.65                    | 52.32                    |
|            |          | (IA) <sup>3</sup> | <b>76.41<sup>†</sup></b> | <b>83.97<sup>†</sup></b> | <b>80.95<sup>†</sup></b> | <b>49.38<sup>†</sup></b> |
| Qwen3-1.7B | LoRA     | Base              | 53.51                    | 68.33                    | 74.52                    | 16.25                    |
|            |          | SIGHT 10%         | <b>55.33<sup>†</sup></b> | <b>74.64<sup>†</sup></b> | <b>76.68<sup>†</sup></b> | <b>16.93</b>             |
|            |          | All-head          | 52.67                    | 72.25                    | 75.42                    | 15.27                    |
|            |          | (IA) <sup>3</sup> | 53.42                    | <b>73.05<sup>†</sup></b> | <b>76.02<sup>†</sup></b> | <b>16.62<sup>†</sup></b> |
|            |          | All-head          | <b>54.10</b>             | 71.58                    | 74.88                    | 15.84                    |

Table 4: Cross-backbone and cross-operator comparison of SIGHT-selected head-sparse PEFT and all-head PEFT references. Bold indicates the better result between SIGHT-selected 10% head-sparse PEFT and the corresponding all-head PEFT reference under the same backbone, PEFT operator, and task. <sup>†</sup> indicates that SIGHT is significantly better than the corresponding all-head PEFT reference, with significance testing described in Section 4.3.

stead, they point to the value of gaze-based within-layer ranking. SIGHT also slightly outperforms all-head LoRA while updating only 10% of heads, suggesting that targeted allocation can partly offset a much smaller trainable-head budget.

Table 4 extends the comparison across Qwen3 backbones and head-sparse PEFT operators. For each backbone and operator, we compare SIGHT-selected 10% head-sparse PEFT with the corresponding all-head reference under the same task and training configuration. The Qwen3-8B LoRA rows are repeated from Table 3 for completeness, and bold numbers mark the better result within each SIGHT-all-head pair.

Across three backbones and two PEFT operators, SIGHT remains competitive with and, in most cases, outperforms the corresponding all-head reference. This suggests that gaze-attention alignment is a reusable criterion for allocating head-sparse PEFT capacity across model scales and operators. Full selector comparisons for smaller Qwen3 backbones and (IA)<sup>3</sup> are reported in Appendix E.

### 5.2 Functional Deletion Evidence

We next test whether SIGHT-selected heads provide target support before head-sparse PEFT. At deletion ratio  $q$ , heads are masked according to each ranking and the target support  $L_q$  is recomputed. For MMLU and CSQA,  $L_q$  is the length-normalized log probability of the gold option; for BLiMP, it is the grammaticality margin from Section 4.2; and for BB Syn., it is the aggregate log-

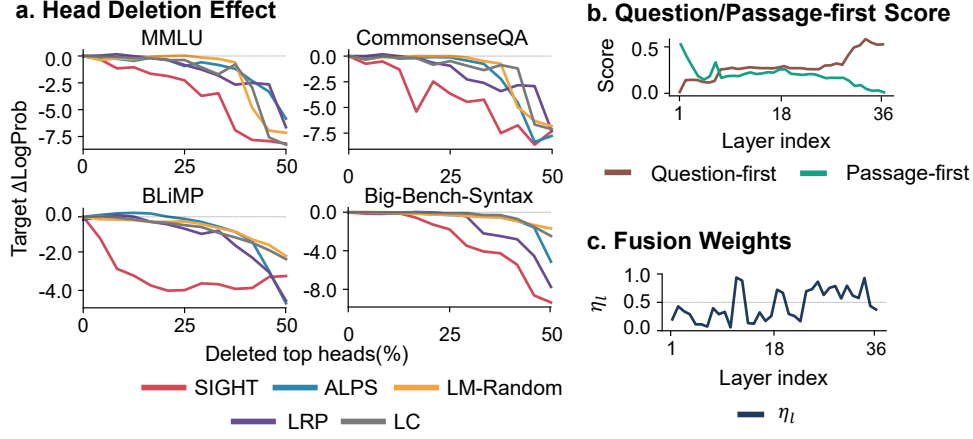


Figure 2: Functional effects and layer-wise structure of SIGHT. Panel a shows target log-probability changes after deleting heads according to representative rankings. Panel b shows layer-wise question-first and passage-first gaze-attention scores. Panel c shows the learned layer-wise fusion weights  $\eta_l$ .

probability support over the required answer positions. A sharper decline in  $L_q - L_0$  indicates that the deleted heads carry more predictive support. We report the Qwen3-8B analysis and plot the five selectors with the highest average accuracy in Table 3: SIGHT, ALPS, LM-Random, LRP, and LC.

Figure 2a shows that, when heads are deleted from highest to lowest according to each selector’s ranking, SIGHT-ranked heads cause the steepest and most consistent degradation across reasoning and syntax benchmarks. This suggests that SIGHT-selected heads already carry stronger target support in the base model before head-sparse PEFT. The comparison with LM-Random is especially informative because it preserves SIGHT’s layer-wise selection distribution while randomizing heads within each layer. The larger degradation from SIGHT therefore attributes the effect to SIGHT’s within-layer head ranking rather than to favorable layer selection, although head deletion may also induce non-specific distributional damage, as discussed in Appendix F.

Figures 2b and 2c illustrate how SIGHT forms the fused ranking. Passage-first scores dominate more in early layers, whereas question-first scores become more prominent in later layers. The learned fusion weights  $\eta_l$  adapt to this layer-wise shift, allowing SIGHT to combine passage-level integration and goal-directed evidence access rather than using a fixed average.

### 5.3 Ablations and Surface Confound Controls

**Ablation.** We use ablations to test whether SIGHT’s gains come from condition-matched hu-

| Variant       | Ablation type      | MMLU                     | CSQA                     | BLiMP                    | BB Syn.                  |
|---------------|--------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Full SIGHT    | none               | <b>76.13<sup>†</sup></b> | <b>84.47<sup>†</sup></b> | <b>81.86<sup>†</sup></b> | <b>63.03<sup>†</sup></b> |
| QF-only       | condition          | 74.13                    | 82.95                    | 79.42                    | 57.88                    |
| PF-only       | condition          | 73.55                    | 82.23                    | 79.12                    | 59.16                    |
| Swapped       | condition mismatch | 73.62                    | 81.96                    | 78.91                    | 55.74                    |
| LastPos       | condition mismatch | 74.32                    | 82.61                    | 79.63                    | 55.81                    |
| SelfAttn      | condition mismatch | 73.91                    | 82.03                    | 79.54                    | 52.33                    |
| No TD         | gaze metric        | 73.81                    | 82.02                    | 79.03                    | 57.21                    |
| Shuffled gaze | gaze alignment     | 73.11                    | 81.54                    | 78.32                    | 51.16                    |

Table 5: Ablations under the LoRA instantiation. Full SIGHT uses matched question-first and passage-first gaze-attention alignment scores. Bold indicates the better result in each column, and <sup>†</sup> indicates that SIGHT is significantly better than every other compared method.

man reading signals rather than from head-sparse PEFT alone. Since LoRA gives the clearest and most controllable setting in Table 4, all ablations use SIGHT-LoRA with a 10% head budget. Table 5 reports the main Qwen3-8B results, with Qwen3-4B and Qwen3-1.7B results in Appendix G.

We compare Full SIGHT with three groups of variants. QF-only and PF-only keep only one reading condition, testing whether the two conditions are complementary. Swapped, LastPos, and SelfAttn disrupt condition matching: Swapped exchanges the QF/PF gaze-attention pairing; LastPos keeps the question-first input but replaces predicted-option-to-passage attention with attention from the final prompt position to passage words; SelfAttn keeps the passage-first input but replaces later-to-earlier attention with passage-level self-attention averaged over passage token positions. No TD removes total dwell time, and Shuffled gaze randomly permutes word-level gaze values while preserving their marginal distribution.

Table 5 shows that Full SIGHT performs best

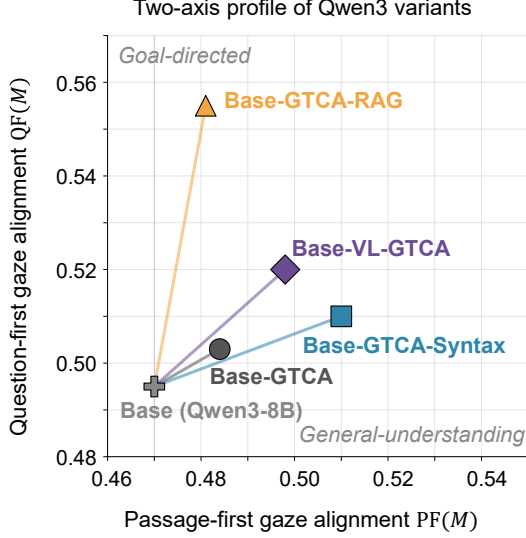


Figure 3: Reading-alignment profile of Qwen3-based variants. The x-axis reports passage-first gaze alignment, and the y-axis reports question-first gaze alignment.

across all datasets. QF-only and PF-only both degrade performance, while same-metric controls in Appendix Table 5 indicate that both reading regime and condition-matched metric choice contribute. These variants and No TD also produce substantially different selected head sets from Full SIGHT. Performance drops when condition matching is disrupted, especially on BB Syn. Finally, Shuffled gaze performs worse than Full SIGHT, suggesting that the gains are not explained by head-sparse PEFT alone, attention mass, or marginal gaze statistics. Overall, the ablations support condition-matched word-level alignment as the main source of SIGHT’s advantage.

**Surface Confound Controls.** We test whether SIGHT mainly reflects shallow lexical or positional confounds. We recompute gaze-head alignment after residualizing gaze values with respect to word position, word length, frequency, punctuation, and sentence-boundary indicators. The controlled ranking slightly reduces performance relative to Full SIGHT but still outperforms the strongest gaze-free baselines, suggesting that SIGHT’s signal is not reducible to these surface factors. Detailed results are reported in Appendix Table 6 and Appendix H.

## 6 Diagnostic Analysis

Beyond evaluating SIGHT as a head selector, we next test whether the same two-axis gaze-alignment space can diagnose several model variants modified by syntax supervision, retrieval augmentation, and

vision-language continued training. For model  $M$ , we aggregate the two condition scores over layers and heads:

$$\begin{aligned} \text{QF}(M) &= \text{mean}_{l,h} \text{QF}^{l,h,M} \\ \text{PF}(M) &= \text{mean}_{l,h} \text{PF}^{l,h,M} \end{aligned} \quad (9)$$

The pair  $[\text{PF}(M), \text{QF}(M)]$  maps a model into a shared reading-alignment space, where the two axes measure passage-first and question-first alignment, respectively.

We use Qwen3-8B as the base model (Yang et al., 2025). Base-GTCA adds a gated tree cross-attention (GTCA) branch that reads cached constituency chunk memory (Gao et al., 2026). We then construct three variants that modify Base-GTCA along different dimensions: Base-GTCA-Syntax adds a lightweight BIO chunking head trained on CoNLL 2000 (Tjong Kim Sang and Buchholz, 2000); Base-GTCA-RAG keeps Base-GTCA fixed and adds a generated evidence snippet at inference time; and Base-VL-GTCA attaches GTCA to Base-VL-Thinking (Bai et al., 2025) and continues training on UAIT (Ling and Ding, 2026). Details of the RAG prompt, generator, and VL model are provided in Appendix I.

Figure 3 places each variant in this two-axis alignment space: question-first alignment captures goal-directed evidence use, while passage-first alignment captures passage-level understanding. Across both goal-directed and passage-level reading tasks, attention signals from all four models achieve correlations above 0.45 with human reading times and consistently outperform those from the base Qwen3 model. Base-GTCA moves away from the base model, with Base-GTCA-RAG shifting mainly upward and Base-GTCA-Syntax mainly rightward, suggesting that generated evidence and chunk supervision affect distinct reading-profile dimensions. Base-VL-GTCA lies between these shifts, indicating a different effect from multimodal pretraining and grounded continued training. Overall, added mechanisms change not only task accuracy but also attention allocation over passage words, making this space a useful secondary diagnostic for reading-related behavior.

## 7 Conclusion

This paper examined whether human reading behavior can guide head-sparse PEFT allocation in LLMs. We introduced SIGHT, a gaze-guided

head selector that ranks attention heads by aligning passage-word attention with word-level gaze metrics under question-first and passage-first reading, and uses the ranking only to place head-sparse PEFT modules. Across Qwen3-based models, updating the top 10% SIGHT-selected heads improves over other sparse selectors and remains competitive with, and often outperforms, all-head PEFT references on reasoning and syntax benchmarks. Deletion and ablation analyses indicate that these gains depend on condition-matched gaze-attention alignment rather than head sparsity, attention magnitude, or layer-wise selection effects. Additionally, the word-level attention signals produced by our models can serve as proxies for reading effort. All four models proposed in this study predict human reading times with correlation coefficients above 0.45, consistently outperforming the base Qwen3 model. Overall, SIGHT supports human reading behavior as an external signal for allocating adaptation capacity, while broader generalization beyond the studied models and tasks remains future work.

## Limitations

SIGHT has several limitations. First, the current study is limited to reading-centric language understanding. The selector is derived from eye-tracking data collected during reading comprehension, and the evaluations focus on multiple-choice reasoning and syntax benchmarks. The results therefore do not show whether the same criterion transfers to tool use and mathematical reasoning. Second, the method depends on the availability and quality of gaze data. The current experiments use an English dataset with two controlled reading orders, and extending SIGHT to other languages, domains, reader populations, and reading interfaces will require new or adapted resources. Reusing one backbone-specific ranking amortizes but does not remove this acquisition cost. Synthetic eye movements are a promising future direction, but we have not established whether a gaze generator preserves the goal-dependent QF/PF differences required here. Third, SIGHT selects attention heads rather than finer-grained mechanisms. This granularity supports head-sparse PEFT, deletion, and cross-operator transfer, but may miss mechanisms distributed across MLPs, residual streams, token representations, or sparse features.

## Ethics Statement

We use previously released eye-tracking and benchmark datasets and do not collect new human-subject data. The human reading data used in this work come from the eye-tracking study of [Zou et al., 2023](#), *Human attention during goal-directed reading comprehension relies on task optimization*. As reported in the original publication, the experimental procedures were approved by the Research Ethics Committee of the College of Medicine, Zhejiang University (approval number 2019–047), and participants provided written consent and were compensated. Our work is a secondary analysis of the released data and does not involve recruiting participants, collecting new gaze recordings, or interacting with human subjects.

The released eye-tracking data have been anonymized, and our method uses only aggregate word-level gaze statistics rather than participant-level traces. We do not attempt to infer individual traits, abilities, identities, or other sensitive attributes from gaze behavior. Because gaze traces can encode behavioral and potentially sensitive information, we restrict their use to head-selection analysis and parameter-allocation experiments.

## Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments, as well as Honghua Chen and Jing Wang for their assistance in revising the manuscript. This work was supported by the Brain Science and Brain-like Intelligence Technology - National Science and Technology Major Project 2021ZD0204100 (2021ZD0204105 to N. D.) and the National Natural Science Foundation of China (32300856 to J. Z.).

## References

- Paul Albert, Frederic Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton Hengel, and Ehsan Abbasnejad. 2025. [Randlora: Full rank parameter-efficient fine-tuning of large models](#). In *International Conference on Learning Representations*, volume 2025, pages 22258–22270.
- Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. [CausalGym: Benchmarking causal interpretability methods on linguistic tasks](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Bruno Bianchi, Fermín Travi, and Juan E. Kamienkowski. 2025. [Modeling cognitive processes of natural reading with transformer-based language models](#). *Preprint*, arXiv:2505.11485.
- Lena S Bolliger, Patrick Haller, Isabelle CR Cretton, David R Reich, Tannon Kew, and Lena A Jäger. 2025. Emtec: A corpus of eye movements on machine-generated texts. *Behavior Research Methods*, 57(7):189.
- Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.
- Hao Chen, Haoze Li, Zhiqing Xiao, Lirong Gao, Qi Zhang, Xiaomeng Hu, Ningtao Wang, Xing Fu, and Junbo Zhao. 2025. [ALPS: Attention localization and pruning strategy for efficient adaptation of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11764–11780, Vienna, Austria. Association for Computational Linguistics.
- Christian Clark, Byung-Doh Oh, and William Schuler. 2025. [Linear recency bias during training improves transformers’ fit to reading times](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7735–7747, Abu Dhabi, UAE. Association for Computational Linguistics.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. [Enhancing chat language models by scaling high-quality instructional conversations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.
- Xinyu Gao, Shaonan Wang, and Nai Ding. 2026. [Gated tree cross-attention for checkpoint-compatible syntax injection in decoder-only llms](#). *Preprint*, arXiv:2602.15846.
- Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. [The effect of surprisal on reading times in information seeking and repeated reading](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Cfir Avraham Hadar, Omer Shubi, Yoav Meiri, Amit Heshes, and Yevgeni Berzak. 2026. [Decoding open-ended information seeking goals from eye movements in reading](#). In *The Fourteenth International Conference on Learning Representations*.
- Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. [LoRA+: Efficient low rank adaptation of large models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17783–17806. PMLR.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 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*, pages 138–146, Marseille, France. European Language Resources Association.
- 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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Farnoush Rezaei Jafari, Oliver Eberle, Ashkan Khakzar, and Neel Nanda. 2025. [Relp: Faithful and efficient circuit discovery in language models via relevance patching](#). *Preprint*, arXiv:2508.21258.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

- Johanna K. Kaakinen, Jukka Hyönä, and Janice M. Keenan. 2002. [Perspective effects on online text processing](#). *Discourse Processes*, 33(2):159–173.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAd-ing comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chen Ling and Nai Ding. 2026. [Evaluating the encoding competence of visual language models using uncommon actions](#). *Preprint*, arXiv:2601.07737.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohata, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. [Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.
- Ángela López-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2025. [Seeing eye to ai: Human alignment via gaze-based response rewards for large language models](#). In *International Conference on Learning Representations*, volume 2025, pages 26111–26135.
- Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. [Sparse feature circuits: Discovering and editing interpretable causal graphs in language models](#). In *International Conference on Learning Representations*, volume 2025, pages 23888–23923.
- Athul M Mathew, Haithem Hermassi, Thariq Kadavil, and Arshad Ali Khan. 2026. [GazeVLM: A vision-language model for multi-task gaze understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 9241–9250.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. [Are sixteen heads really better than one?](#) In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Andrew Nam, Henry Conklin, Yukang Yang, Tom Griffiths, Jonathan D Cohen, and Sarah-Jane Leslie. 2025. [Causal head gating: A framework for interpreting roles of attention heads in transformers](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 115609–115631. Curran Associates, Inc.
- Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. [Frequency explains the inverse correlation of large language models’ size, training data amount, and surprisal’s fit to reading times](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian’s, Malta. Association for Computational Linguistics.
- Anupam Pani and Yanchao Yang. 2025. [Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 163544–163577. Curran Associates, Inc.
- Keith Rayner. 1998. [Eye movements in reading and information processing: 20 years of research](#). *Psychological bulletin*, 124(3):372–422.
- Keith Rayner, Gretchen Kambe, and Susan A. Duffy. 2000. [The effect of clause wrap-up on eye movements during reading](#). *The Quarterly Journal of Experimental Psychology Section A*, 53(4):1061–1080. PMID: 11131813.
- Jun Rekimoto. 2025. [Gazellm: Multimodal llms incorporating human visual attention](#). In *Proceedings of the Augmented Humans International Conference 2025, AHs ’25*, pages 302–311, New York, NY, USA. Association for Computing Machinery.
- Tiago Ribeiro, Stephanie Brandl, Anders Søgaard, and Nora Hollenstein. 2024. [Webqamgaze: A multi-lingual webcam eye-tracking-while-reading dataset](#). *Preprint*, arXiv:2303.17876.
- Sofia Serrano and Noah A. Smith. 2019. [Is attention interpretable?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. [Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models](#). *Transactions on Machine Learning Research*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. [Introduction to the CoNLL-2000 shared task chunking](#).

In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. [Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Tessa Warren, Sarah J. White, and Erik D. Reichle. 2009. [Investigating the causes of wrap-up effects: Evidence from eye movements and e-z reader](#). *Cognition*, 111(1):132–137.

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

Sarah Wiegreffe and Yuval Pinter. 2019. [Attention is not not explanation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. [Left: Representation fine-tuning for language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 63908–63962. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Fangcong Yin, Xi Ye, and Greg Durrett. 2024. [Lofit: Localized fine-tuning on llm representations](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 9474–9506. Curran Associates, Inc.

Yifan Zhang, Jiliang Li, Zachary Karas, Aakash Bansal, Toby Jia-Jun Li, Collin McMillan, Kevin Leach, and Yu Huang. 2024. [Eyetrans: Merging human and machine attention for neural code summarization](#). *Proc. ACM Softw. Eng.*, 1(FSE).

Jiajie Zou, Yuran Zhang, Jialu Li, Xing Tian, and Nai Ding. 2023. [Human attention during goal-directed reading comprehension relies on task optimization](#). *eLife*, 12:RP87197.

## A Layer-wise Gate Calibration via Discrete Dynamic Programming

This section describes the calibration procedure for the layer-wise fusion gate used in SIGHT. The goal is to select a bounded gate value  $\eta_l \in [0, 1]$  for each layer  $l$ , where a larger  $\eta_l$  gives more weight to the question-first score and a smaller  $\eta_l$  gives more weight to the passage-first score. The sigmoid parameterization in the main text is used to emphasize that the gate is bounded; in calibration, we optimize the bounded gate value  $\eta_l$  directly.

For each layer  $l$ , let

$$\mathbf{q}_l^M = \{\text{QF}^{l,h,M}\}_{h=1}^H, \quad \mathbf{p}_l^M = \{\text{PF}^{l,h,M}\}_{h=1}^H, \quad (1)$$

where  $H$  is the number of attention heads in one layer. The fused score for head  $h$  in layer  $l$  is computed from the question-first and passage-first alignment scores:

$$s_{l,h}^M(\eta_l) = \eta_l q_{l,h}^M + (1 - \eta_l) p_{l,h}^M, \quad \eta_l \in [0, 1]. \quad (2)$$

**Calibration subset.** We randomly sample 30% of the available gaze-scored items in each reading condition as the gate-calibration subset. Since passage-first reading was collected for 96 items in Experiment 4 of Zou et al. (2023), the PF condition uses these 96 items, whereas the QF condition uses the full 800-item set. The sampling is stratified by reading order so that the question-first and passage-first conditions have the same proportions in the calibration subset. The remaining 70% of the gaze-scored items are used to compute the final SIGHT head scores after the layer-wise gates have been selected. No downstream task labels, validation accuracy, deletion performance, or PEFT feedback are used during gate calibration.

On the calibration subset, we compute condition-specific head scores  $\mathbf{q}_l^{\text{cal}}$  and  $\mathbf{p}_l^{\text{cal}}$ . For each candidate value of  $\eta_l$ , we obtain the fused calibration score vector

$$\mathbf{s}_l^{\text{cal}}(\eta_l) = \eta_l \mathbf{q}_l^{\text{cal}} + (1 - \eta_l) \mathbf{p}_l^{\text{cal}}. \quad (3)$$

The layer-wise local objective is the average fused alignment score of the top-scoring heads in layer  $l$ :

$$J_l(\eta_l) = \frac{1}{m} \sum_{h \in \text{Top}_m(\mathbf{s}_l^{\text{cal}}(\eta_l))} s_{l,h}^{\text{cal}}(\eta_l), \quad (4)$$

where  $\text{Top}_m(\cdot)$  returns the indices of the  $m$  highest-scoring heads within layer  $l$ . We set  $m =$

$\max(1, \lceil 0.1H \rceil)$ , matching the 10% head-sparse PEFT budget. Thus, the local objective favors the QF/PF mixture that produces the strongest gaze-aligned candidate heads in that layer.

**Global calibration objective.** The layer-wise gates are selected by optimizing the QF/PF mixture on the gate-calibration subset while discouraging unstable layer-to-layer oscillations:

$$\{\eta_l^*\}_{l=1}^L = \arg \max_{\eta_1, \dots, \eta_L} \left[ \sum_{l=1}^L J_l(\eta_l) - \mu \sum_{l=2}^L (\eta_l - \eta_{l-1})^2 \right]. \quad (5)$$

The smoothness coefficient is set to  $\mu = 0.05$ . Because the smoothness term couples adjacent layers, the gates are not selected independently per layer.

**Discrete dynamic programming.** We optimize the constrained gate values on a fixed grid

$$\mathcal{G} = \{0, 0.05, 0.10, \dots, 1.00\}. \quad (6)$$

For every layer  $l$  and every grid value  $v \in \mathcal{G}$ , we first compute  $J_l(v)$  using the top- $m$  objective above. The remaining optimization problem is a chain-structured discrete maximization, which we solve exactly by dynamic programming.

Let  $\mathcal{G} = \{v_1, \dots, v_K\}$  with  $K = 21$ . We define

$$D_1(j) = J_1(v_j), \quad (7)$$

and for  $l = 2, \dots, L$ ,

$$D_l(j) = J_l(v_j) + \max_{i \in \{1, \dots, K\}} [D_{l-1}(i) - \mu(v_j - v_i)^2]. \quad (8)$$

For each state  $(l, j)$ , we store the maximizing predecessor index. After filling the dynamic-programming table, we select the final-layer grid value by

$$j_L^* = \arg \max_j D_L(j), \quad (9)$$

and recover the optimal gate sequence  $\eta_1^*, \dots, \eta_L^*$  by backtracking through the stored predecessors. This procedure gives the exact maximizer of the discrete calibration objective over the 21-point grid.

**Final SIGHT scoring.** After calibration, the selected gate sequence is fixed for all downstream head-sparse PEFT experiments. We then recompute the question-first and passage-first head scores

| Backbone   | Config                          | All-head LoRA | 10% LoRA       | All-head (IA) <sup>3</sup> | 10% (IA) <sup>3</sup> |
|------------|---------------------------------|---------------|----------------|----------------------------|-----------------------|
| Qwen3-8B   | $L = 36$<br>$H_Q/H_{KV} = 32/8$ | 97.3M         | $\approx 9.7M$ | 73.7K                      | $\approx 7.4K$        |
| Qwen3-4B   | $L = 36$<br>$H_Q/H_{KV} = 32/8$ | 61.9M         | $\approx 6.2M$ | 73.7K                      | $\approx 7.4K$        |
| Qwen3-1.7B | $L = 28$<br>$H_Q/H_{KV} = 16/8$ | 23.4M         | $\approx 2.3M$ | 57.3K                      | $\approx 5.7K$        |

Appendix Table 1: Approximate trainable PEFT parameter counts for the head-indexed LoRA and (IA)<sup>3</sup> operators. LoRA uses rank  $r = 8$  and  $\alpha = 16$ . The 10% columns report proportional budget estimates; in implementation, K/V parameters are counted once per unique K/V head induced by the selected query heads.

on the remaining gaze-scored items and obtain the final fused SIGHT score:

$$\text{SIGHT}^{l,h,M} = \eta_l^* \text{QF}^{l,h,M} + (1 - \eta_l^*) \text{PF}^{l,h,M}. \quad (10)$$

Heads are ranked by this final fused score, and the top-ranked heads are used only to place the head-sparse PEFT modules.

## B Head-To-Slice Mapping Under GQA

Qwen3 backbones use grouped-query attention (GQA), with  $H_Q$  query heads and  $H_{KV}$  key/value heads per layer. Let  $d$  be the hidden size,  $d_h$  be the attention-head dimension, and  $g = H_Q/H_{KV}$  be the number of query heads sharing one key/value head. SIGHT ranks query attention heads. For a selected query head  $h$ , the corresponding key/value head index is

$$\kappa(h) = \left\lfloor \frac{h}{g} \right\rfloor. \quad (11)$$

LoRA is attached to the query-output slice of  $q\_proj$  for head  $h$ , the output-input slice of  $o\_proj$  for head  $h$ , and the key/value slices of  $k\_proj$  and  $v\_proj$  indexed by  $\kappa(h)$ . When multiple selected query heads share the same key/value head, the K/V LoRA modules are instantiated only once and shared by all selected query heads in that GQA group. The same convention is used for (IA)<sup>3</sup>: selected query heads induce a set of unique K/V heads, and one pair of K/V scale vectors is learned for each unique induced K/V head.

## C Hyperparameter Search, Budget and Parameter Scale

**Hyperparameter search.** The all-head PEFT rows are same-recipe references rather than exhaustively tuned upper bounds. To check that the comparison was not driven by a clearly poor all-head recipe, we performed a small hyperparameter

search around the settings in Table 2. For LoRA, we searched learning rates in  $\{1e-4, 2e-4, 4e-4\}$  with rank 8 and  $\alpha = 16$ . For (IA)<sup>3</sup>, we searched learning rates in  $\{4e-4, 8e-4, 1.6e-3\}$  with scale vectors initialized to 1.0. In both cases, we considered one or two training epochs and selected recipes using UltraChat validation loss rather than downstream evaluation accuracy. The final reported all-head references use the same training data, optimizer, evaluation protocol, and selected recipe as the corresponding head-sparse runs, so they should be interpreted as matched-recipe references rather than separately optimized upper bounds.

**Compute Budget and Software Versions.** Training was implemented in PyTorch using Distributed-DataParallel (DDP) across 8 GPUs per run. The total compute budget was approximately 3000 GPU-hours, including gaze-based head scoring, head-sparse PEFT training, ablations, and evaluation. For the reported PEFT training runs, we used bf16 mixed precision and the AdamW optimizer.

The main backbone models were Qwen3-8B, Qwen3-4B, and Qwen3-1.7B, with approximately 8.2B, 4.0B, and 1.7B parameters, respectively. Relevant software versions were CUDA 12.8, cuDNN 9.10.2, PyTorch 2.8, NCCL 2.27.3.

**Parameter counting.** Let  $\mathcal{S}_l$  be the set of selected query heads in layer  $l$ , and let  $\mathcal{U}_l = \{\kappa(h) : h \in \mathcal{S}_l\}$  be the induced set of unique K/V heads. With LoRA rank  $r$ , the number of trainable head-indexed LoRA parameters is

$$N_{\text{LoRA}} = 2r(d + d_h) \sum_{l=1}^L (|\mathcal{S}_l| + |\mathcal{U}_l|), \quad (12)$$

where the two terms correspond to Q/O slices and K/V slices, respectively. For (IA)<sup>3</sup>, which learns K/V scale vectors, the number of trainable parameters is

$$N_{(\text{IA})^3} = 2d_h \sum_{l=1}^L |\mathcal{U}_l|. \quad (13)$$

For all-head references,  $\mathcal{S}_l$  contains all  $H_Q$  query heads and  $\mathcal{U}_l$  contains all  $H_{KV}$  key/value heads in every layer. Under the 10% head-sparse budget, the exact parameter count depends slightly on how many unique K/V heads are induced by the selected query heads; shared K/V heads are always counted once. The resulting trainable parameter counts are summarized in Appendix Table 1.

## D Prompt Templates, Licenses and Terms of Use

All evaluations use log-probability scoring without demonstrations. The following templates define the context  $x$  used by the scorer. Candidate answers are scored as continuations of the corresponding prompt, and the candidate with the highest length-normalized log probability is selected.

### MMLU.

```
Question: {question}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer:
```

The candidate continuations are the option labels A, B, C, and D followed by the corresponding option text.

### CSQA.

```
Question: {question}
A. {choice_a}
B. {choice_b}
C. {choice_c}
D. {choice_d}
E. {choice_e}
Answer:
```

The candidate continuations are the five answer choices. Scores are length-normalized to avoid favoring shorter choices.

### BLiMP.

```
Sentence:
```

For each minimal pair, the acceptable sentence and the unacceptable sentence are scored as alternative continuations after the same prefix. The prediction is correct when  $\log p(s_{\text{good}} | x) > \log p(s_{\text{bad}} | x)$ .

### BB Syn.

```
Complete the following syntactic judgment task.
Input: {input_text}
Candidates: {candidate_set}
Answer:
```

For tasks with multiple answer positions, an example is counted as correct only when each gold continuation has the highest log probability among the candidates for that position.

| Model              | MMLU                      | CSQA                      | BLiMP                     | BB Syn.                   |
|--------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Qwen3-8B           | 71.59                     | 80.26                     | 76.58                     | 39.51                     |
| SIGHT-LoRA 10%     | <b>76.13</b> <sup>†</sup> | <b>84.47</b> <sup>†</sup> | <b>81.86</b> <sup>†</sup> | <b>63.03</b> <sup>†</sup> |
| CHG-LoRA 10%       | 73.48                     | 81.74                     | 79.40                     | 37.20                     |
| LRP-LoRA 10%       | 73.61                     | 82.96                     | 76.88                     | 58.18                     |
| ALPS-LoRA 10%      | 74.68                     | 82.95                     | 80.23                     | 61.47                     |
| LC-LoRA 10%        | 74.52                     | 82.46                     | 79.21                     | 57.63                     |
| Magnitude-LoRA 10% | 73.09                     | 81.33                     | 78.94                     | 54.34                     |
| Random-LoRA 10%    | 74.40                     | 82.88                     | 78.64                     | 56.98                     |
| LM-Random-LoRA 10% | 74.62                     | 83.05                     | 78.81                     | 57.23                     |
| All-head LoRA      | 74.53                     | 81.98                     | 79.20                     | 61.47                     |
| Qwen3-4B           | 71.33                     | 77.54                     | 80.17                     | 29.65                     |
| SIGHT-LoRA 10%     | <b>74.02</b> <sup>†</sup> | <b>79.14</b> <sup>†</sup> | <b>81.34</b> <sup>†</sup> | <b>54.71</b> <sup>†</sup> |
| CHG-LoRA 10%       | 72.96                     | 78.37                     | 80.48                     | 44.78                     |
| LRP-LoRA 10%       | 73.18                     | 78.62                     | 80.12                     | 48.26                     |
| ALPS-LoRA 10%      | 73.52                     | 78.74                     | 80.88                     | 52.47                     |
| LC-LoRA 10%        | 73.31                     | 78.48                     | 80.56                     | 48.91                     |
| Magnitude-LoRA 10% | 72.64                     | 77.95                     | 80.34                     | 51.06                     |
| Random-LoRA 10%    | 72.81                     | 78.20                     | 79.92                     | 51.73                     |
| LM-Random-LoRA 10% | 73.28                     | 78.55                     | 80.22                     | 51.84                     |
| All-head LoRA      | 73.45                     | 78.32                     | 79.65                     | 52.32                     |
| Qwen3-1.7B         | 53.51                     | 68.33                     | 74.52                     | 16.25                     |
| SIGHT-LoRA 10%     | <b>55.33</b> <sup>†</sup> | <b>74.64</b> <sup>†</sup> | <b>76.68</b> <sup>†</sup> | <b>16.93</b>              |
| CHG-LoRA 10%       | 54.39                     | 72.18                     | 75.52                     | 15.91                     |
| LRP-LoRA 10%       | 54.62                     | 73.34                     | 75.95                     | 16.38                     |
| ALPS-LoRA 10%      | 54.88                     | 73.76                     | 76.10                     | 16.51                     |
| LC-LoRA 10%        | 54.70                     | 73.42                     | 75.78                     | 16.24                     |
| Magnitude-LoRA 10% | 54.11                     | 71.96                     | 75.23                     | 15.76                     |
| Random-LoRA 10%    | 54.76                     | 73.58                     | 75.74                     | 16.12                     |
| LM-Random-LoRA 10% | 54.82                     | 73.61                     | 75.86                     | 16.18                     |
| All-head LoRA      | 52.67                     | 72.25                     | 75.42                     | 15.27                     |

Appendix Table 2: Full LoRA selector comparison across Qwen3 backbones. The Qwen3-8B block matches Table 3; additional rows for smaller backbones are included for completeness. Bold indicates the best mean score in each column within each backbone block, and <sup>†</sup> denotes significance as defined in Section 4.3.

**Artifact Licenses and Terms of Use.** We used publicly available benchmark datasets, pretrained checkpoints, and released eye-tracking data, and followed their original licenses, recommended citations, and terms of use. The Qwen3-8B, Qwen3-4B, and Qwen3-1.7B checkpoints (Yang et al., 2025) are used under the Apache-2.0 license.

The benchmark and head-sparse PEFT datasets are used under their respective public terms. Specifically, UltraChat (Ding et al., 2023) and CommonsenseQA (Talmor et al., 2019) are used under the MIT license, BLiMP (Warstadt et al., 2020) under CC-BY, BIG-Bench (Srivastava et al., 2023) under Apache-2.0, and RACE-derived reading-comprehension materials for non-commercial research use. MMLU (Hendrycks et al., 2021b,a) and the eye-tracking dataset of Zou et al. (2023) are used according to their original public distribution terms and access conditions.

We do not redistribute the content of the dataset, individual-level gaze records, or pretrained model weights. Any derived models produced in this work are intended solely as research prototypes for con-

| Model                           | MMLU                      | CSQA                      | BLiMP                     | BB Syn.                   |
|---------------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Qwen3-8B                        | 71.59                     | 80.26                     | 76.58                     | 39.51                     |
| SIGHT-(IA) <sup>3</sup> 10%     | <b>74.65</b> <sup>†</sup> | <b>82.41</b> <sup>†</sup> | 75.94                     | <b>57.28</b> <sup>†</sup> |
| CHG-(IA) <sup>3</sup> 10%       | 72.52                     | 81.64                     | 76.21                     | 42.75                     |
| LRP-(IA) <sup>3</sup> 10%       | 72.91                     | 82.18                     | 76.54                     | 53.46                     |
| ALPS-(IA) <sup>3</sup> 10%      | 73.16                     | 82.25                     | 76.72                     | 55.91                     |
| LC-(IA) <sup>3</sup> 10%        | 72.84                     | 81.72                     | 76.38                     | 52.74                     |
| Magnitude-(IA) <sup>3</sup> 10% | 72.44                     | 80.92                     | 76.33                     | 49.11                     |
| Random-(IA) <sup>3</sup> 10%    | 72.88                     | 81.47                     | 76.26                     | 51.62                     |
| LM-Random-(IA) <sup>3</sup> 10% | 73.02                     | 81.95                     | 76.31                     | 52.88                     |
| All-head (IA) <sup>3</sup>      | 72.86                     | 81.12                     | <b>78.71</b>              | 54.21                     |
| Qwen3-4B                        | 71.33                     | 77.54                     | 80.17                     | 29.65                     |
| SIGHT-(IA) <sup>3</sup> 10%     | <b>76.41</b> <sup>†</sup> | <b>83.97</b> <sup>†</sup> | <b>80.95</b> <sup>†</sup> | <b>49.38</b> <sup>†</sup> |
| CHG-(IA) <sup>3</sup> 10%       | 73.66                     | 82.21                     | 79.63                     | 46.12                     |
| LRP-(IA) <sup>3</sup> 10%       | 74.08                     | 82.96                     | 79.84                     | 45.74                     |
| ALPS-(IA) <sup>3</sup> 10%      | 74.35                     | 83.18                     | 80.06                     | 43.54                     |
| LC-(IA) <sup>3</sup> 10%        | 73.92                     | 82.75                     | 79.68                     | 43.48                     |
| Magnitude-(IA) <sup>3</sup> 10% | 73.22                     | 81.45                     | 79.28                     | 46.03                     |
| Random-(IA) <sup>3</sup> 10%    | 73.74                     | 82.58                     | 79.51                     | 46.16                     |
| LM-Random-(IA) <sup>3</sup> 10% | 74.02                     | 82.80                     | 79.66                     | 45.78                     |
| All-head (IA) <sup>3</sup>      | 73.41                     | 82.04                     | 79.05                     | 47.20                     |
| Qwen3-1.7B                      | 53.51                     | 68.33                     | 74.52                     | 16.25                     |
| SIGHT-(IA) <sup>3</sup> 10%     | 53.42                     | <b>73.05</b> <sup>†</sup> | <b>76.02</b> <sup>†</sup> | <b>16.62</b> <sup>†</sup> |
| CHG-(IA) <sup>3</sup> 10%       | 52.91                     | 71.84                     | 75.31                     | 15.98                     |
| LRP-(IA) <sup>3</sup> 10%       | 53.17                     | 72.36                     | 75.58                     | 16.24                     |
| ALPS-(IA) <sup>3</sup> 10%      | 53.28                     | 72.68                     | 75.75                     | 16.39                     |
| LC-(IA) <sup>3</sup> 10%        | 53.04                     | 72.18                     | 75.35                     | 16.05                     |
| Magnitude-(IA) <sup>3</sup> 10% | 52.76                     | 71.29                     | 75.05                     | 15.66                     |
| Random-(IA) <sup>3</sup> 10%    | 52.98                     | 72.05                     | 75.22                     | 15.91                     |
| LM-Random-(IA) <sup>3</sup> 10% | 53.08                     | 72.22                     | 75.34                     | 16.02                     |
| All-head (IA) <sup>3</sup>      | <b>54.10</b>              | 71.58                     | 74.88                     | 15.84                     |

Appendix Table 3: Full (IA)<sup>3</sup> selector comparison across Qwen3 backbones. The SIGHT and all-head rows match Table 4; additional selector rows are included for completeness. Bold indicates the best mean score in each column within each backbone block, and <sup>†</sup> denotes significance as defined in Section 4.3.

trolled evaluation, consistent with the access conditions of the underlying resources.

## E Full Selector Comparisons

This section reports the full results. The LoRA results are reported in Appendix Table 2, and the (IA)<sup>3</sup> results are reported in Appendix Table 3. Values in both tables are means over five training seeds, for Random, the values additionally average over ten independently sampled head sets. Bold numbers mark the best result within each backbone and head-sparse PEFT operator block.

Appendix Table 3 gives the corresponding (IA)<sup>3</sup> comparison under the same backbone order and evaluation metrics.

## F Deletion Selectivity and Non-Specific Distributional Damage

The deletion curves in Section 5.2 are intended as functional evidence rather than causal proof. Deleting attention heads can degrade the model for two reasons: the removed heads may provide

target-relevant support, or the intervention may simply induce non-specific distributional damage. We therefore conduct additional qualitative selectivity checks.

First, for multiple-choice tasks, we compare the change in the gold-option score with the change in the decision margin between the gold option and the strongest competing option. The same qualitative pattern holds under the margin-based view: deleting SIGHT-ranked heads reduces the model’s preference for the correct candidate more strongly than deleting heads selected by random, layer-matched random, and most model-internal selectors. This suggests that the observed drop is not merely a uniform decrease in all candidate scores.

Second, we inspect the candidate-level output distribution before and after head deletion. If the SIGHT curve were explained only by generic disruption, we would expect SIGHT deletion to produce a uniformly larger distributional shift across all answer candidates. Instead, the degradation is concentrated in the target-support and decision-margin quantities used by the evaluation. This indicates that SIGHT-ranked heads are more target-selective under deletion, rather than only more globally disruptive.

Third, we compare top-ranked deletion with bottom-ranked deletion under the same head budget. Deleting the bottom-ranked heads according to SIGHT produces a much smaller effect than deleting the top-ranked heads. This top–bottom asymmetry is substantially weaker for random and layer-matched random rankings. This provides a sanity check that the deletion effect reflects information in the SIGHT ranking, not simply the number or layer positions of deleted heads.

Overall, these checks support a cautious interpretation of Figure 2: SIGHT-selected heads appear to provide stronger target-relevant functional support in the base model. We do not interpret the deletion experiment as proving that these heads are the sole causal mechanism behind the final prediction, since head deletion remains an intervention that can affect the model distribution more broadly.

## G Additional Selector Ablations

The selector-ablation results for Qwen3-4B and Qwen3-1.7B under the LoRA 10% setting are reported in Appendix Table 4. The variants are the same as those in the main text: QF-only, PF-only,

| Backbone   | Variant       | MMLU                     | CSQA                     | BLiMP                    | BB Syn.                  |
|------------|---------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Qwen3-4B   | Full SIGHT    | <b>74.02<sup>†</sup></b> | <b>79.14<sup>†</sup></b> | <b>81.34<sup>†</sup></b> | <b>54.71<sup>†</sup></b> |
|            | QF-only       | 73.36                    | 78.62                    | 80.72                    | 50.31                    |
|            | PF-only       | 72.94                    | 78.14                    | 80.54                    | 51.77                    |
|            | Swapped       | 72.88                    | 77.92                    | 80.23                    | 50.05                    |
|            | LastPos       | 73.42                    | 78.39                    | 80.81                    | 52.74                    |
|            | SelfAttn      | 73.15                    | 77.86                    | 80.68                    | 51.03                    |
|            | No TD         | 73.04                    | 77.98                    | 80.27                    | 52.16                    |
|            | Shuffled gaze | 72.35                    | 77.41                    | 79.96                    | 52.88                    |
| Qwen3-1.7B | Full SIGHT    | <b>55.33<sup>†</sup></b> | <b>74.64<sup>†</sup></b> | <b>76.68<sup>†</sup></b> | <b>16.93</b>             |
|            | QF-only       | 54.76                    | 73.50                    | 76.10                    | 16.54                    |
|            | PF-only       | 54.41                    | 72.92                    | 75.86                    | 16.61                    |
|            | Swapped       | 54.18                    | 72.41                    | 75.54                    | 15.86                    |
|            | LastPos       | 54.92                    | 73.21                    | 76.05                    | 16.10                    |
|            | SelfAttn      | 54.55                    | 72.66                    | 75.92                    | 15.63                    |
|            | No TD         | 54.46                    | 72.75                    | 75.61                    | 16.03                    |
|            | Shuffled gaze | 53.98                    | 72.04                    | 75.22                    | 15.19                    |

Appendix Table 4: Additional selector ablations under the LoRA 10% setting. Full SIGHT remains best across both smaller backbones, while condition mismatch, attention-direction mismatch, and shuffled gaze weaken performance. Bold indicates the best mean score in each column within each backbone block, and <sup>†</sup> denotes significance as defined in Section 4.3.

| Variant        | QF metrics | PF metrics | MMLU  | CSQA  | BLiMP | BB Syn. |
|----------------|------------|------------|-------|-------|-------|---------|
| Full SIGHT     | RR+TD      | FP+TD      | 76.13 | 84.47 | 81.86 | 63.03   |
| QF-TD          | TD         | —          | 74.28 | 82.74 | 79.55 | 57.96   |
| PF-TD          | —          | TD         | 73.91 | 82.49 | 79.36 | 59.04   |
| Fused-TD       | TD         | TD         | 75.39 | 83.68 | 80.88 | 61.57   |
| Shared-all     | FP+RR+TD   | FP+RR+TD   | 75.66 | 83.89 | 81.19 | 62.07   |
| Metric-swapped | FP+TD      | RR+TD      | 74.66 | 83.04 | 80.14 | 59.73   |

Appendix Table 5: Controls separating reading regime from gaze-metric choice under Qwen3-8B LoRA with a 10% head budget.

swapped condition matching, final-position attention in the question-first condition, averaged self-attention in the passage-first condition, removal of total dwell, and shuffled word-level gaze values.

## H Surface-Controlled Gaze Alignment

To test whether SIGHT is explained by shallow lexical or positional factors, we recompute gaze-attention alignment after controlling for word position, word length, log word frequency, punctuation, and sentence-boundary indicators. For each item  $z$ , we form a covariate matrix  $\mathbf{X}_z$  over passage words and residualize both the gaze vector and the attention vector:

$$\tilde{\mathbf{g}}_z = \mathbf{g}_z - \mathbf{X}_z \hat{\theta}_g, \quad \tilde{\mathbf{a}}_z^{l,h,M} = \mathbf{a}_z^{l,h,M} - \mathbf{X}_z \hat{\theta}_a. \quad (14)$$

The controlled alignment score is then computed by replacing  $(\mathbf{g}_z, \mathbf{a}_z^{l,h,M})$  with  $(\tilde{\mathbf{g}}_z, \tilde{\mathbf{a}}_z^{l,h,M})$  in the Spearman-correlation scoring step. The resulting controlled comparison is reported in Appendix Table 6.

| Backbone   | Selector                       | MMLU               | CSQA               | BLiMP              | BB Syn.            |
|------------|--------------------------------|--------------------|--------------------|--------------------|--------------------|
| Qwen3-8B   | Full SIGHT-LoRA 10%            | 76.13 <sup>†</sup> | 84.47 <sup>†</sup> | 81.86 <sup>†</sup> | 63.03 <sup>†</sup> |
|            | Controlled SIGHT-LoRA 10%      | 75.86 <sup>†</sup> | 84.10 <sup>†</sup> | 81.48 <sup>†</sup> | 62.53 <sup>†</sup> |
|            | Best gaze-free sparse baseline | 74.68              | 83.05              | 80.23              | 61.47              |
| Qwen3-4B   | Full SIGHT-LoRA 10%            | 74.02 <sup>†</sup> | 79.14 <sup>†</sup> | 81.34 <sup>†</sup> | 54.71 <sup>†</sup> |
|            | Controlled SIGHT-LoRA 10%      | 73.93 <sup>†</sup> | 79.04 <sup>†</sup> | 81.21 <sup>†</sup> | 54.02 <sup>†</sup> |
|            | Best gaze-free sparse baseline | 73.52              | 78.74              | 80.88              | 52.47              |
| Qwen3-1.7B | Full SIGHT-LoRA 10%            | 55.33 <sup>†</sup> | 74.64 <sup>†</sup> | 76.68 <sup>†</sup> | 16.93 <sup>†</sup> |
|            | Controlled SIGHT-LoRA 10%      | 55.04 <sup>†</sup> | 74.18 <sup>†</sup> | 76.35 <sup>†</sup> | 16.80              |
|            | Best gaze-free sparse baseline | 54.88              | 73.61              | 76.10              | 16.51              |

Appendix Table 6: Surface-controlled SIGHT results. Residualizing lexical and positional covariates slightly reduces the scores but preserves the advantage over the strongest gaze-free sparse baseline. <sup>†</sup> indicates that the corresponding SIGHT variant significantly outperforms the best gaze-free sparse baseline under the test defined in Section 4.3.

## I RAG and VL Diagnostic Variant Details

The Base-GTCA-RAG diagnostic variant uses a short generated evidence snippet at inference time. The generator is Deepseek-R1 (Guo et al., 2025) with greedy decoding, maximum generation length 96 tokens, and temperature 0.0. The Base-GTCA parameters are kept fixed, only the input is augmented with the generated evidence.

### Evidence-generation prompt.

You are given a reading-comprehension passage and a  
↪ question. Write one short evidence snippet from the  
↪ passage that would help answer the question. Do not  
↪ answer the question. Do not introduce facts that are  
↪ not supported by the passage.

Passage:  
{passage}

Question:  
{question}

Evidence snippet:

The generated snippet is inserted before the original question in the diagnostic input:

Evidence: {generated\_evidence}  
Question: {question}  
Passage: {passage}  
Options: {options}  
Answer:

This diagnostic setting is used only for the reading-alignment profile and is not used for the head-sparse PEFT experiments.

**VL diagnostic backbone.** For Base-VL-GTCA, we use Qwen3-8B-VL-Thinking (Bai et al., 2025) as the VL backbone and attach the same GTCA branch before continued training on UAIT.