Soft Head Selection for Injecting ICL-Derived Task Embeddings

Anonymous ACL submission

Abstract

Large language models (LLMs) are commonly adapted to downstream tasks using parameter-efficient fine-tuning (PEFT) or in-context learning (ICL). Recently, ICL-driven embedding-based adaptation has been proposed as a distinct task adaptation paradigm. It derives task-specific embeddings from intermediate activations using few-shot prompts and injects them during inference. Despite its conceptual appeal, this approach has not demonstrated consistent performance gains over PEFT or ICL, and its empirical advantages have been limited in practice. We propose Soft head-selection for ICL-derived Task Embeddings (SITE), a gradient-based method that identifies task-relevant attention heads to enable effective task embedding injection. Across various types of open-ended generation, reasoning, and natural language understanding tasks, SITE significantly outperforms prior embedding-based adaptation methods and few-shot ICL, while using substantially fewer trainable parameters than PEFT. Experiments on 12 LLMs ranging from 4B to 70B parameters demonstrate the generality of our approach, and intra-task and inter-task activation patching analyses further provide new mechanistic insights by revealing strong task dependence in attention head functionality.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable generation capabilities across a wide range of domains, along with the ability to rapidly adapt to new tasks. Traditionally, such task adaptation has been achieved through parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019; Lester et al., 2021; Hu et al., 2022) or in-context learning (ICL) (Brown et al., 2020). PEFT typically yields strong task performance but requires training, whereas ICL enables training-free and flexible adaptation by incorporating input-output demonstrations into the prompt, at the cost of increased prompt length and inference overhead.

Recent studies have shown that intermediate last-token activations of LLMs elicited by few-shot ICL encode rich task-relevant information (Hendel et al., 2023; Todd et al., 2023). These activations can be extracted and later injected into the model, enabling task execution without explicit instructions or demonstrations in the prompt. This line of work, referred to as *ICL-driven embedding-based adaptation*, introduces a new mechanism for conveying task information to LLMs. Most existing approaches (Hendel et al., 2023; Todd et al., 2023; Zhang et al., 2024; Huang et al., 2024; Li et al., 2024; Wang et al., 2024a; Cai et al., 2025; Liu and Deng, 2025) extract *task embeddings*, which are assumed to encode task-relevant information, from last-token activations at selected layers or modules of the model, and focus on designing heuristics for where and how to extract and inject them. However, despite their conceptual appeal, existing methods have not yet demonstrated clear advantages over PEFT or ICL in terms of either adaptation efficiency or task performance. The goal of this work is to develop an ICL-driven embedding-based adaptation method that achieves clear empirical advantages over these two alternatives.

Attention head attribution (Hao et al., 2021; Olsson et al., 2022; Todd et al., 2023; Gandelsman et al., 2023; Park et al., 2024; Zhou et al., 2024; Wu et al., 2024; Elhelo and Geva, 2024), a line of research in mechanistic interpretability, studies the functional roles of individual attention heads in deep neural networks, including retrieval, safety, and in-context learning behaviors. While prior studies have identified attention heads associated with in-context learning, it remains largely unexplored whether the importance of attention heads varies across tasks. In our preliminary activation patching (Zhang and Nanda, 2023; Hendel et al., 2023; Todd et al., 2023; Bereska and Gavves, 2024) experiment shown in Figure 1, where randomly selected attention head activations during zero-shot infer-

ence are replaced with those extracted from few-shot inference, we observe that task performance is highly sensitive to which heads are patched. This sensitivity suggests that task-relevant information may be unevenly distributed across attention heads and may vary substantially across tasks.

Motivated by this observation, we propose Soft head-selection for ICL-derived Task Embeddings (SITE). SITE identifies task-relevant attention heads by formulating head selection as a continuous optimization problem, enabling efficient learning of head importance via gradient descent algorithm. For each task, our method constructs task embeddings by extracting last-token attention head activations from few-shot prompts, and learns soft head-selection parameters that linearly interpolate between the original head activations and the task embeddings. Using zero-shot prompts at inference time, SITE achieves strong performance across various open-ended generation, complex reasoning, and natural language understanding tasks, significantly outperforming prior embedding-based adaptation methods as well as few-shot ICL, while using substantially fewer trainable parameters than PEFT. We evaluate SITE on 12 LLMs spanning 4 model families, 3 model variants, and sizes ranging from 4B to 70B parameters, demonstrating its broad applicability.

Furthermore, we extend activation patching experiments using our head-selection parameters to further investigate task dependence in attention head functionality. In intra-task patching, we show that the selected heads effectively capture task-relevant information. In inter-task patching, we find that similar tasks tend to share important attention heads, whereas dissimilar tasks do not, indicating strong task specificity in head roles. These findings extend prior studies on the attention head functionalities from a task-agnostic perspective to a task-specific one. Overall, SITE significantly outperforms few-shot ICL and approaches PEFT-level performance, while using substantially fewer trainable parameters than PEFT. Also, it provides new mechanistic insights into task-specific functional roles of attention heads in LLMs.

2 Related work

Task adaptation of LLMs. Task adaptation methods for LLMs can be broadly categorized into prompt-based, weight-based, and embedding-based approaches, according to where task-relevant

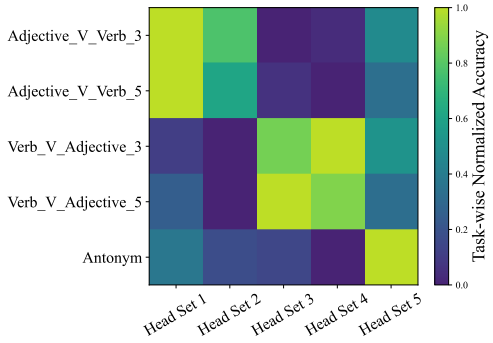

Figure 1: **Random head patching.** During zero-shot inference, activations of randomly selected attention heads are replaced (or patched) with corresponding activations extracted from 10-shot inference, illustrating the sensitivity of task performance to the patched heads.

information is encoded. Prompt-based methods (Zhou et al., 2022; Liu et al., 2022b; Lu et al., 2022; Yang et al., 2023) include in-context learning and instruction optimization, which improve task performance by selecting, ordering, or designing input-output exemplars or task instructions within the prompt. Weight-based approaches comprise weight-based PEFT techniques such as Adapters (Houlsby et al., 2019), LoRA (Hu et al., 2022), and $(IA)^3$ (Liu et al., 2022a), which adapt models by introducing lightweight trainable modules or updating a small subset of model weights. Embedding-based methods (Lester et al., 2021; Li and Liang, 2021; Li et al., 2023; Hendel et al., 2023; Todd et al., 2023) inject task-specific embeddings, which are substantially smaller than PEFT modules, into intermediate activations, enabling task execution without using in-prompt instructions or demonstrations at inference time.

ICL-driven embedding-based adaptation. A class of embedding-based adaptation methods (Lester et al., 2021; Peng et al., 2024; Saglam et al., 2025; Li et al., 2025b; Kang et al., 2025; Li et al., 2025a; Yang et al., 2025b) directly optimizes task embeddings via gradient-based training, but these approaches typically require tens of thousands of optimization steps and often suffer from training instability. In contrast, ICL-driven embedding-based methods (Hendel et al., 2023; Todd et al., 2023; Zhang et al., 2024; Huang et al., 2024; Li et al., 2024; Wang et al., 2024a; Cai et al., 2025; Liu and Deng, 2025) avoid tuning task embeddings and instead focus on identifying effective locations for extracting and injecting ICL-drived task embeddings. Early methods such as FV (Todd et al., 2023) and TV (Hendel et al., 2023)

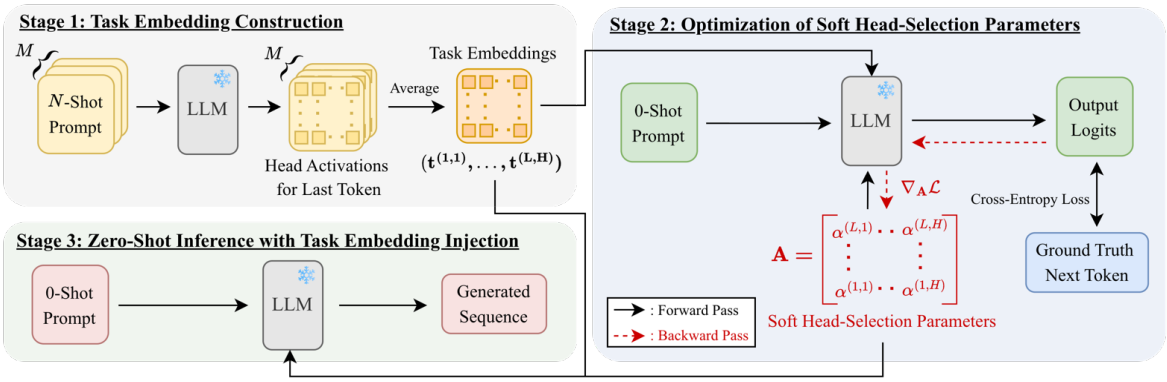

Figure 2: **Method Overview.** Our method consists of three stages: (1) constructing task embeddings by averaging last-token attention head activations across M few-shot ICL prompts with N input-output pairs; (2) optimizing soft head-selection parameters to identify where task embeddings should be injected during zero-shot inference; and (3) applying the task embeddings and learned selection parameters at inference time to perform tasks without in-prompt examples. L and H denote the number of attention layers and attention heads per layer, respectively, in the LLM.

rely on layer- or head-wise activation patching, which incurs substantial search overhead. More recent approaches, including MTV (Huang et al., 2024) and I2CL (Li et al., 2024), reduce this cost through more efficient optimization strategies, but still rely on reinforcement learning-based optimization (Williams, 1992) or restricted search spaces, and often fail to consistently outperform few-shot ICL. Moreover, most prior work focuses on simple classification tasks and omits comparisons with PEFT, limiting rigorous empirical evaluation. In contrast, our method identifies task embedding injection locations at the level of individual attention heads using gradient-based soft head-selection, and achieves performance comparable to PEFT across diverse task types.

3 Method

Our method consists of three stages. (1) First, we construct task embeddings using few-shot ICL prompts. (2) Next, we optimize soft head-selection parameters via gradient descent. (3) Finally, we apply the pre-computed task embeddings and the learned head-selection parameters at inference time to enable task execution without in-prompt examples. Figure 2 provides an overview of our method.

Task embedding construction. Let $P_1, P_2, \ldots, P_M$ denote the M few-shot ICL prompts, each containing N input-output examples sampled from the training set. For each prompt P_m, we perform a forward pass through the model and extract the output of every attention head at every layer. Specifically, for head $h \in \{1, 2, \ldots, H\}$ in layer $l \in \{1, 2, \ldots, L\}$, the output of head h at layer l for prompt P_m is given by:

$$\mathbf{t}_m^{(l,h)} = \text{softmax}\left(\frac{\mathbf{Q}_m^{(l,h)}(\mathbf{K}_m^{(l,h)})^T}{\sqrt{d_k}}\right)\mathbf{V}_m^{(l,h)} \in \mathbb{R}^{S_m \times d_v}, \quad (1)$$

where $\mathbf{Q}_m^{(l,h)}, \mathbf{K}_m^{(l,h)}, \mathbf{V}_m^{(l,h)}$ are the query, key, and value matrices, d_k is the key/query dimension, d_v is the value dimension, and S_m is the number of tokens in the tokenized sequence of prompt P_m. We extract the activations corresponding to the last token and average them across all M prompts to obtain a task embedding for each head:

$$\mathbf{t}^{(l,h)} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{t}_m^{(l,h)}[-1, :] \in \mathbb{R}^{d_v} \quad (2)$$

The collection $\{\mathbf{t}^{(l,h)}\}_{l=1,\ldots,L;\, h=1,\ldots,H}$ constitutes the task embeddings for the given task.

Optimization of soft head-selection parameters. To effectively identify task-relevant attention heads, we formulate head selection as a continuous optimization problem. For each task, we introduce a learnable matrix $\mathbf{A}$, where each entry $\alpha^{(l,h)}$ serves as a *soft head-selection parameter* for attention head $h \in \{1, 2, \ldots, H\}$ in layer $l \in \{1, 2, \ldots, L\}$:

$$\mathbf{A} = \begin{bmatrix} \alpha^{(L,1)} & \cdots & \alpha^{(L,H)} \\ \vdots & \ddots & \vdots \\ \alpha^{(1,1)} & \cdots & \alpha^{(1,H)} \end{bmatrix} \in [0, 1]^{L \times H} \quad (3)$$

Each $\alpha^{(l,h)}$ controls the degree to which task-specific information is injected into the corresponding attention head. Let $\mathbf{o}^{(l,h)} \in \mathbb{R}^{d_v}$ denote the

original last-token activation of head h in layer l during optimization. We inject task embeddings by linearly interpolating between the original activation $\mathbf{o}^{(l,h)}$ and the task embedding $\mathbf{t}^{(l,h)}$:

$$\mathbf{o}^{(l,h)} \leftarrow (1 - \alpha^{(l,h)}) \cdot \mathbf{o}^{(l,h)} + \alpha^{(l,h)} \cdot \mathbf{t}^{(l,h)}, \quad (4)$$

for all $l \in \{1, 2, \ldots, L\}$ and $h \in \{1, 2, \ldots, H\}$. During optimization, the LLM is kept frozen and only $\mathbf{A}$ is updated. We optimize $\mathbf{A}$ for a few hundred gradient descent steps by minimizing the cross-entropy loss for next-token prediction, computed from the output logits of the modified forward passes. At each optimization step, inference is performed using a zero-shot prompt, with task embeddings injected according to the current values of $\mathbf{A}$. Each $\alpha^{(l,h)}$ is parameterized as the sigmoid of an unconstrained scalar to ensure the values in $[0, 1]$. The pseudocode for this procedure is provided in Algorithm 1 of Appendix A.2.

Zero-shot inference with task embedding injection. After optimization, we apply the learned soft head-selection parameters $\mathbf{A}$ with the task embeddings $\{\mathbf{t}^{(l,h)}\}_{l,h}$ to guide the LLM in performing tasks without in-prompt exemplars. The soft injection is applied in the same manner as during the optimization stage (Equation 4) but only once, at the last token of the initial input prompt, assuming KV caching (Pope et al., 2023) is enabled during autoregressive decoding. No further interventions are applied during subsequent decoding steps. This single-step injection embeds task-relevant information into the KV cache at the start of generation, allowing the model to produce the remaining tokens without additional intervention. While some prior methods inject task embeddings at multiple token positions (Huang et al., 2024; Li et al., 2024), we intervene only at a single-token activation to reduce intervention complexity and avoid excessive undesirable steering during text generation.

4 Experiments

4.1 Experimental setup

Models. We use Llama-3.1-8B to compare multiple methods across a wide range of tasks, and apply our method on an additional 11 LLMs to assess its generality. In total, the 12 LLMs used in this study span 4 model families of Llama 3.1 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023, 2024), Qwen3 (Yang et al., 2025a), and Gemma-3 (Team et al., 2025), covering 3 variation types and sizes ranging from 4B to 70B parameters. Detailed information about the 12 LLMs is provided in Table 5 of Appendix A.1.

Tasks. We evaluate our method on 57 ICL tasks from the official Function Vectors (FV) (Todd et al., 2023) repository, as well as three additional benchmarks: ANLI (Nie et al., 2020), MMLU-Pro (Wang et al., 2024b), and Big-Bench Hard (BBH) (Suzgun et al., 2023). The FV benchmark comprises 29 abstractive and 28 extractive open-ended generation tasks spanning diverse problem types, including closely related and contrasting variants, making it well suited for studying in-context learning behaviors. ANLI is a natural language inference benchmark that is more difficult than earlier datasets of SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). MMLU-Pro and BBH consist of tasks requiring diverse and complex reasoning, where MMLU-Pro extends the original MMLU (Hendrycks et al., 2020) with more challenging and realistic question sets.

Implementation details. For each task, the dataset is split into training, validation, and test sets following the split ratio used in FV; only the training and validation sets are used to construct task embeddings and optimize soft head-selection parameters. Unless otherwise specified, task embeddings are constructed using $M = 50$ few-shot prompts, each containing $N = 10$ input-output exemplars that are randomly sampled from the training set, with possible overlaps across prompts. The soft head-selection parameters are initialized to 0.5 and optimized using the Adam (Kingma, 2014) optimizer with a learning rate of 0.2 for 400 iterations, without regularization or model-specific hyperparameter tuning, across all 12 LLMs. Checkpoints are selected based on validation loss, evaluated every 50 iterations. We use greedy decoding to minimize randomness. Implementation details for other methods are provided in Appendix A.2.

4.2 Experimental results

Across four benchmarks, we compare our method with parameter-based methods (LoRA (Hu et al., 2022), $(IA)^3$ (Liu et al., 2022a)), in-context learning (ICL), and prior embedding-based approaches (Prompt Tuning (Lester et al., 2021), FV (Todd et al., 2023), TV (Hendel et al., 2023), MTV (Huang et al., 2024), LIVE (Peng et al., 2024), I2CL (Li et al., 2024), IV (Liu and Deng, 2025)). We report results of our method using

Category	Method	# Trainable Parameters	FV Benchmark (57 tasks)	ANLI (3 tasks)	MMLU-Pro (14 tasks)	Big-Bench Hard (27 tasks)	Average
Parameter-based	LoRA	3407.87K	86.76 ± (0.24)	45.82 ± (1.20)	41.04 ± (0.34)	60.39 ± (0.48)	58.50
	$(IA)^3$	524.29K	81.97 ± (0.43)	47.03 ± (0.14)	40.87 ± (0.42)	60.29 ± (0.70)	57.54
ICL	0-shot	-	9.85 ± (0.00)	0.00 ± (0.00)	17.96 ± (0.00)	16.41 ± (0.00)	11.06
	10-shot	-	76.76 ± (0.09)	43.96 ± (0.34)	36.47 ± (0.48)	47.17 ± (0.98)	51.09
Embedding-based	Prompt Tuning	81.92K	75.06 ± (0.35)	33.22 ± (0.13)	10.60 ± (0.66)	33.45 ± (0.81)	38.13
	FV	-	33.16 ± (0.08)	0.00 ± (0.00)	1.14 ± (0.04)	17.82 ± (0.37)	13.03
	TV	-	59.07 ± (0.83)	33.17 ± (0.04)	31.60 ± (0.25)	42.01 ± (0.71)	41.46
	MTV	1.02K	73.24 ± (1.46)	34.70 ± (0.88)	34.07 ± (0.88)	42.54 ± (0.71)	46.14
	LIVE	131.10K	23.34 ± (1.03)	0.00 ± (0.00)	20.51 ± (2.07)	12.89 ± (2.32)	14.19
	I2CL	0.13K	79.89 ± (0.56)	28.01 ± (3.94)	27.14 ± (0.32)	50.60 ± (1.12)	46.41
	IV	-	42.52 ± (0.13)	31.52 ± (0.44)	9.84 ± (0.51)	26.25 ± (1.53)	27.53
	Ours ($M = 1$)	1.02K	89.67 ± (0.60)	46.35 ± (0.51)	37.24 ± (0.84)	56.76 ± (1.42)	57.50
	Ours ($M = 50$)	1.02K	**90.02** ± (0.19)	**47.31** ± (0.15)	**38.78** ± (0.41)	**58.04** ± (0.72)	**58.54**

Table 1: **Comparison on four benchmarks using Llama-3.1-8B.** We compare our method (SITE) with parameter-based methods, in-context learning (ICL), and prior embedding-based approaches. The number of trainable parameters is reported where applicable. Best embedding-based results are shown in **bold**, and best parameter-based and ICL results are underlined.

two values of M (1 and 50) to demonstrate the robustness of our method to the number of prompts used for task embedding construction. Table 1 reports the mean accuracy and standard deviation over three random seeds using Llama-3.1-8B.

Overall, our method significantly outperforms all embedding-based baselines and ICL across benchmarks. For example, SITE with $M = 50$ outperforms 10-shot ICL by 13.26% on the FV benchmark and by 7.45% on average across all four benchmarks. Compared to parameter-based methods such as LoRA and $(IA)^3$, our method achieves higher performance on the FV benchmark and ANLI, and comparable results on MMLU-Pro and BBH. These two latter benchmarks require complex reasoning or expert-level knowledge, suggesting that ICL-derived task embeddings may have limited representational capacity or require additional adaptation for such complex tasks. Notably, our method achieves these results with substantially fewer trainable parameters than PEFT, as it freezes both the backbone LLM and the task embeddings and optimizes only the injection locations.

To evaluate generality across backbone LLMs, we further apply our method to 12 LLMs spanning four model families, three model variants, and sizes ranging from 4B to 70B parameters. Figure 3 presents results on the 57 FV tasks, comparing our method with 10-shot ICL. Across all models, our method achieves an average improvement of 10.2-14.3% over 10-shot ICL, demonstrating strong generalization across diverse LLMs. Task-level results for all 12 models are reported in Tables 12-23 of Appendix B.

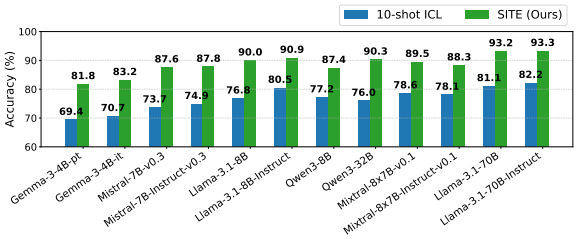

Figure 3: **Average performance on the FV benchmark for 12 backbone LLMs.** We report results for 10-shot ICL and our method, with average accuracies annotated above each bar.

5 Activation patching analysis with binarized head-selection parameters

Activation patching (Meng et al., 2022; Zhang and Nanda, 2023; Hendel et al., 2023; Todd et al., 2023; Bereska and Gavves, 2024) is a widely used technique in mechanistic interpretability for analyzing the functional roles of model components. By replacing the activations of specific components with alternative representations, it enables direct assessment of their causal contributions through changes in model outputs. In this section, we extend the activation patching experiment introduced in Figure 1 by leveraging binarized head-selection parameters.

As shown in Figure 4, the learned soft head-selection parameters exhibit near-binary patterns, motivating the use of their binarized counterparts for activation patching. Using these binarized parameters, we conduct two types of activation patching experiments. In both settings, we patch selected attention head activations during zero-shot infer-

ence using task embeddings extracted from 10-shot prompts. The two experiments differ in whether the head-selection parameters are derived from the same task as zero-shot inference or from a different task. Through these experiments, we examine whether the heads selected by our method capture task-relevant information and whether attention head functionality exhibits task-specific property, such that similar tasks share important heads while dissimilar tasks do not. While the analyses in this section are based on Llama-3.1-8B, the findings generalize to other larger LLMs, as demonstrated in Appendix C.2 and Appendix D.2.

5.1 Intra-task activation patching

In intra-task activation patching, the task embeddings ($\mathbf{t}^{(l,h)}$), the head-selection parameters ($\alpha^{(l,h)}$), and the zero-shot inference input are all derived from the same task. We compare patching high-α heads ($\alpha^{(l,h)} > 0.5$) with patching low-α heads ($\alpha^{(l,h)} \leq 0.5$) to assess whether the heads selected by our method capture task-relevant information. High-α heads correspond to attention heads assigned larger interpolation weights by the learned head-selection parameters (see Eq. 4), reflecting the heads prioritized by our method for task embedding injection. As shown in Figure 4 and Figures 7-9 in Appendix C.1, the numbers of high-α and low-α heads are comparable across tasks, each accounting for approximately half of the total attention heads.

Table 2 reports the zero-shot performance (before activation patching), the performance after patching activations of high-α heads, and the performance after patching activations of low-α heads. Patching high-α head activations leads to substantial accuracy improvements across all four benchmarks. In contrast, patching low-α head activations consistently degrades performance, even though the task embeddings are derived from the same 10-shot prompts and differ only in which attention heads are patched. These results indicate that the learned head-selection parameters effectively identify task-relevant attention heads.

5.2 Inter-task activation patching

In inter-task activation patching, we fix both the task of zero-shot inference and the task embeddings used for activation patching, and vary only the head-selection parameters across tasks. Specifically, we select a single *evaluation task*, from which the zero-shot inference input and the task embeddings

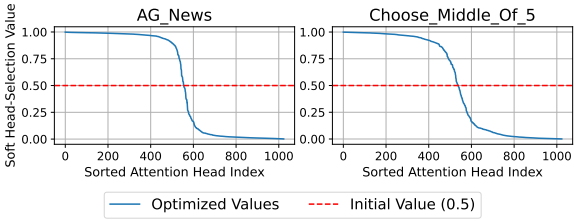

Figure 4: **Optimized values of the soft head-selection parameters for two FV tasks.** Each plot shows the optimized soft head-selection values for all 1024 attention heads in Llama-3.1-8B, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all soft head-selection parameters at the start of training. Results for all 57 tasks in the FV benchmark are provided in Figures 7-9 of Appendix C.1.

Benchmark	Baseline (0-shot)	High-α Heads ($\alpha^{(l,h)} > 0.5$)	Low-α Heads ($\alpha^{(l,h)} \leq 0.5$)
FV Benchmark	9.9	88.9 (+79.0)	3.7 (-6.2)
ANLI	0.0	46.9 (+46.9)	0.0 (-0.0)
MMLU-Pro	18.0	37.8 (+19.8)	7.5 (-10.5)
Big-Bench Hard	16.4	55.7 (+39.3)	13.7 (-2.7)
Average	11.1	57.3 (+46.2)	6.2 (-4.9)

Table 2: **Intra-task activation patching analysis**. Comparison of 0-shot baseline performance with results after patching high-α and low-α attention heads, where the task embeddings and head-selection parameters derived from the same task.

are derived, and apply head-selection parameters learned from different tasks in the FV benchmark. We refer to these tasks as *head-selection tasks*, as they determine which attention heads (i.e., high-α heads) are selected for patching. We focus on the 57 FV tasks, which span diverse problem types with closely related and contrasting variants, making them well suited for analyzing task-specific in-context learning behavior.

By keeping the task embeddings fixed and varying only the patching locations using head-selection parameters from different tasks, any differences in model outputs can be attributed solely to task-specific head selection. If attention head functionality is task-specific, then applying head-selection parameters derived from different tasks should lead to substantially different model outputs, despite using identical task embeddings. This experiment therefore tests whether high-α heads reflect task-specific functional roles.

Table 3 reports inter-task activation patching results for seven evaluation tasks, along with brief task descriptions for clarity. For each evaluation task, we report the top-3 and bottom-3

Evaluation Task	Task Description	Top-3 Head-Selection Tasks (Accuracy, %)	Bottom-3 Head-Selection Tasks (Accuracy, %)
Adjective_V_Verb_5	Select the only adjective from a list of 5 words (1 adjective, 4 verbs)	Adjective_V_Verb_3 (96.7) Adjective_V_Verb_5 (94.3) Animal_V_Object_5 (78.6)	Verb_V_Adjective_3 (3.3) Verb_V_Adjective_5 (4.8) English_French (5.7)
Verb_V_Adjective_5	Select the only verb from a list of 5 words (1 verb, 4 adjectives)	Verb_V_Adjective_3 (98.6) Verb_V_Adjective_5 (98.1) Color_V_Animal_5 (81.4)	Adjective_V_Verb_3 (1.0) Antonym (7.6) Park_Country (7.6)
Alphabetically_First_5	Choose the word that comes first in alphabetical order from a list of 5 words	Alphabetically_First_5 (84.8) Alphabetically_First_3 (42.4) Commonsense_QA (29.5)	Alphabetically_Last_5 (5.7) Alphabetically_Last_3 (8.6) Choose_Middle_Of_3 (13.3)
Alphabetically_Last_5	Choose the word that comes last in alphabetical order from a list of 5 words	Alphabetically_Last_5 (39.0) Alphabetically_Last_3 (31.4) Commonsense_QA (25.7)	Alphabetically_First_5 (0.0) Alphabetically_First_3 (8.1) Capitalize_Second_Letter (13.3)
English_French	Translate the given English word into French	English_French (81.2) English_German (80.7) English_Spanish (80.0)	Person_Instrument (34.0) Next_Capital_Letter (38.9) Object_V_Concept_3 (39.3)
English_German	Translate the given English word into German	English_French (71.4) English_Spanish (68.3) English_German (68.2)	Prev_Item (23.6) Next_Capital_Letter (26.2) Person_Instrument (30.4)
English_Spanish	Translate the given English word into Spanish	English_French (84.4) English_German (83.4) English_Spanish (82.9)	Person_Instrument (39.1) Next_Capital_Letter (40.8) Object_V_Concept_3 (44.7)

Table 3: **Inter-task activation patching analysis for seven evaluation tasks.** For each evaluation task, we report performance after patching high-α attention heads, where task embeddings are fixed to the evaluation task and head-selection parameters are derived from different head-selection tasks. Among the 57 FV tasks, we report the top-3 and bottom-3 head-selection tasks ranked by post-patching accuracy. For clarity, brief task descriptions for all 57 FV tasks are provided in Tables 6-9 of Appendix A.3. More results are provided in Tables 24-27 of Appendix D.

head-selection tasks based on post-patching performance. For example, for the evaluation task Adjective_V_Verb_5, the top-performing head-selection tasks include semantically similar tasks such as Adjective_V_Verb_5, Adjective_V_Verb_3, while the bottom-performing ones include semantically dissimilar tasks such as Verb_V_Adjective_3, Verb_V_Adjective_5. Similar patterns are observed for Verb_V_Adjective_5, Alphabetically_First_5, and Alphabetically_Last_5. For translation tasks, English-French, English-German, and English-Spanish, the top and bottom head-selection tasks are largely consistent across the board. Notably, using high-α heads from semantically similar translation tasks often matches or even exceeds the performance obtained using those from the evaluation task itself, suggesting that closely related tasks share important attention heads.

Overall, the intra-task and inter-task activation patching analyses reveal strong task specificity in attention head functionality, extending prior studies from a task-agnostic to a task-specific perspective. We hope these findings motivate future work on characterizing task-dependent head roles and exploiting such structure for effective and controllable inference-time LLM adaptations.

6 Discussion and efficiency analysis

Training dynamics of head selection. Figure 5 shows the validation loss and test accuracy of our method during optimization of the soft head-selection parameters for Llama-3.1-8B. The figure shows training dynamics for two FV tasks; results for all 57 FV tasks are provided in Appendix E.1, and some selected plots for larger LLMs are provided in Appendix E.2. While the optimization trajectories vary slightly across tasks, the overall trend is consistent: validation loss decreases and test accuracy improves over training iterations. This trend indicates that gradient descent effectively tunes the head-selection parameters to identify meaningful head positions for embedding injection.

Impact of shot count on task performance. Increasing the number of in-prompt examples (or shots) is a common strategy for improving few-shot ICL performance. While some studies (Agarwal et al., 2024; Bertsch et al., 2025) report continued gains as the number of shots increases from *a few to many*, others (Zhang et al., 2025) observe that performance often plateaus after only

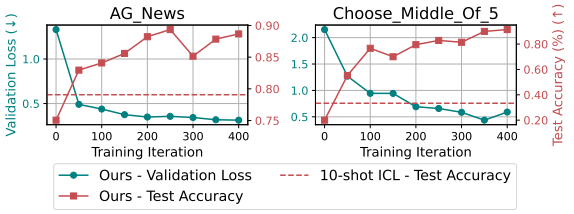

Figure 5: **Training dynamics of soft head-selection parameters for two FV tasks.** Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference. Plots for all 57 FV tasks are provided in Figures 13-15 of Appendix E.1.

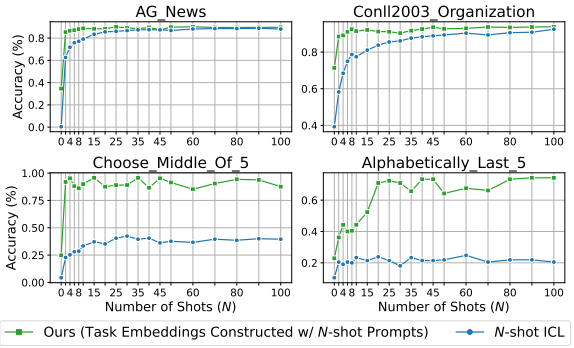

Figure 6: **Impact of shot count on task performance for four FV tasks.** Performance of standard N-shot ICL and our method as the number of shots (N) increases from 0 to 100. For our method, N denotes the number of examples per prompt used in task embedding construction. Results are shown for Llama-3.1-8B.

a small number of examples. A plausible explanation, also suggested by recent work (Zou et al., 2024), is that the benefit of additional shots varies across tasks. Motivated by these observations, we study the effect of shot count on both standard ICL and our task embedding-based adaptation. Specifically, we examine whether many-shot ICL can close the performance gap with our method, and whether increasing the shots of the few-shot prompts used in task embedding construction improves our method by yielding higher-quality ICL-derived task embeddings. Figure 6 reports results on four FV tasks, varying the shots (N) for both the N-shot ICL and the N-shot prompts used in task embedding construction. For AG_News, and Conll2003_Organization, many-shot ICL matches the performance of our method at $N = 100$. In contrast, tasks such as Choose_Middle_Of_5 and Alphabetically_Last_5 exhibit limited gains even with many-shot prompting, indicating that our method remains effective on tasks where many-shot ICL struggles. Moreover, for Alphabetically_Last_5, our method benefits from increasing the shots used in task embedding construction, suggesting that some tasks require richer few-shot context to produce effective ICL-derived task embeddings.

# Test Prompts	0-shot	10-shot	FV	MTV	I2CL	Ours
1000	**15.7**	24.8	269.8	41.9	29.9	22.0
5000	**78.3**	124.2	332.7	169.4	96.1	88.6
10000	**156.7**	248.4	411.2	328.8	178.7	172.0

Table 4: **Runtime comparison as the number of test prompts increases.** Total runtime (in minutes) is reported for 1000, 5000, and 10000 prompts on AG_News using Llama-3.1-8B. Our method scales efficiently as the number of test prompts increases. All runtimes were measured on a single NVIDIA A6000 GPU.

Computational efficiency. Table 4 compares the total runtime of ICL and several embedding-based adaptation methods as the number of test prompts increases. In our method, task embeddings and soft head-selection parameters are computed once (Stages 1 and 2) and reused across all test inputs, so their cost does not scale with the number of prompts. Consequently, our runtime remains close to the zero-shot baseline and is lower than that of 10-shot ICL. In contrast, FV incurs significant overhead due to extensive head-wise activation patching, while MTV is slowed by a suboptimal loop structure in its original implementation; even after code-level optimization, its runtime remains slightly higher than ours. I2CL achieves total runtime comparable to our method. Our approach also has minimal memory overhead, requiring only about 0.5 MB (in float32 precision) to store task embeddings and head-selection parameters for Llama-3.1-8B. Overall, our method achieves strong task performance while maintaining the time and memory efficiency of zero-shot inference, particularly as the number of test prompts grows.

7 Conclusion

This paper introduces SITE, an ICL-driven embedding-based adaptation method that identifies task-relevant attention heads via a continuous, gradient-based optimization framework. Extensive experiments across four diverse benchmarks and 12 LLMs show that SITE consistently outperforms prior embedding-based adaptation methods and few-shot ICL, while approaching PEFT-level performance with substantially fewer trainable parameters. Beyond empirical gains, SITE offers new mechanistic insights into task-dependent attention head functionality through intra-task and inter-task activation patching analyses.

Limitations

While SITE achieves strong performance compared to few-shot ICL and prior embedding-based adaptation methods, it has several limitations. First, SITE requires a modest amount of labeled data to construct ICL-derived task embeddings and optimize the soft head-selection parameters. Although not included in our main experiments, we observed that 30-50 labeled examples are generally sufficient to achieve strong performance, likely because SITE optimizes only a small number of parameters (specifically, $L \times H$ head-selection scalars; e.g., 1024 for Llama-3.1-8B). Nevertheless, acquiring even this amount of labeled data may be challenging for low-resource or newly defined tasks. A promising direction is to augment limited data with LLM-generated synthetic examples, as recent work suggests such synthetic data can rival or even surpass human-curated datasets (Long et al., 2024; Yehudai et al., 2024; Nadas et al., 2025). Second, SITE requires access to internal model activations, specifically attention head outputs, which restricts its applicability to open-source LLMs and precludes deployment on proprietary models such as GPT-5 (OpenAI, 2025) or Gemini 2.5 (Comanici et al., 2025). Finally, we do not evaluate SITE on multi-step reasoning or long-context generation tasks. Extending the method to these settings and evaluating it on a broader range of benchmarks are important directions for future work.

Ethical considerations

Like other embedding-based adaptation methods, our method uses pre-computed ICL-drived task embeddings, instead of in-prompt input-output examples at inference time. In real-world applications, sharing pre-computed task embeddings rather than raw examples can offer benefits such as improved inference efficiency, enhanced privacy protection, and stronger data security. However, such embeddings could also be misused to conceal harmful information and be distributed to users without their awareness. To mitigate these risks, we recommend implementing safeguards, such as pre-release embedding screening, well-defined usage policies, and runtime output filtering tied to embedding identities, prior to the deployment of systems that rely on pre-computed embeddings.

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, and 1 others. 2024. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wang Cai, Hsiu-Yuan Huang, Zhixiang Wang, and Yunfang Wu. 2025. Beyond demonstrations: Dynamic vector construction from latent representations. *arXiv preprint arXiv:2505.20318*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Amit Elhelo and Mor Geva. 2024. Inferring functionality of attention heads from their parameters. *arXiv preprint arXiv:2412.11965*.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2023. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding

in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. 2022. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Brandon Huang, Chancharik Mitra, Leonid Karlinsky, Assaf Arbelle, Trevor Darrell, and Roei Herzig. 2024. Multimodal task vectors enable many-shot multimodal in-context learning. *Advances in Neural Information Processing Systems*, 37:22124–22153.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Joonseong Kang, Soojeong Lee, Subeen Park, Sumin Park, Taero Kim, Jihee Kim, Ryunyi Lee, and Kyungwoo Song. 2025. Adaptive task vectors for large language models. *arXiv preprint arXiv:2506.03426*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Jiaqian Li, Yanshu Li, Ligong Han, Ruixiang Tang, and Wenya Wang. 2025a. Towards generalizable implicit in-context learning with attention routing. *arXiv preprint arXiv:2509.22854*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025b. M^2iv: Towards efficient and fine-grained multimodal in-context learning via representation engineering. In *Second Conference on Language Modeling*.

Zhuowei Li, Zihao Xu, Ligong Han, Yunhe Gao, Song Wen, Di Liu, Hao Wang, and Dimitris N Metaxas. 2024. Implicit in-context learning. *arXiv preprint arXiv:2405.14660*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pages 100–114.

Yiting Liu and Zhi-Hong Deng. 2025. Iterative vectors: In-context gradient steering without backpropagation. In *Forty-second International Conference on Machine Learning*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *arXiv preprint arXiv:2503.14023*.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing antonyms and synonyms in a pattern-based neural network. *arXiv preprint arXiv:1701.02962*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4885–4901.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-09-11.

Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. 2024. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. *arXiv preprint arXiv:2412.02237*.

Yingzhe Peng, Xinting Hu, Jiawei Peng, Xin Geng, Xu Yang, and 1 others. 2024. Live: Learnable in-context vector for visual question answering. *Advances in Neural Information Processing Systems*, 37:9773–9800.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5:606–624.

Baturay Saglam, Xinyang Hu, Zhuoran Yang, Dionysis Kalogerias, and Amin Karbasi. 2025. Learning task representations from in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6634–6663.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Futing Wang, Jianhao Yan, Yue Zhang, and Tao Lin. 2024a. Elicit: Llm augmentation via external in-context capability. *arXiv preprint arXiv:2410.09343*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 1112–1122.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Haolin Yang, Hakaze Cho, Kaize Ding, and Naoya Inoue. 2025b. Task vectors, learned not extracted: Performance gains and mechanistic insight. *arXiv preprint arXiv:2509.24169*.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. *arXiv preprint arXiv:2401.14367*.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

Kaiyi Zhang, Ang Lv, Yuhan Chen, Hansen Ha, Tao Xu, and Rui Yan. 2024. Batch-icl: Effective, efficient, and order-agnostic in-context learning. *arXiv preprint arXiv:2401.06469*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Xiaoqing Zhang, Ang Lv, Yuhan Liu, Flood Sung, Wei Liu, Jian Luan, Shuo Shang, Xiuying Chen, and Rui Yan. 2025. More is not always better? enhancing many-shot in-context learning with differentiated and reweighting objectives. *arXiv preprint arXiv:2501.04070*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2024. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*.

Kaijian Zou, Muhammad Khalifa, and Lu Wang. 2024. Retrieval or global context understanding? on many-shot in-context learning for long-context evaluation. *arXiv preprint arXiv:2411.07130*.

Contents

A Additional experimental setup

A.1 12 large language models used for evaluation

Our evaluation covers 12 large language models in total. From the Llama-3.1 family (Grattafiori et al., 2024), we include Llama-3.1-8B, Llama-3.1-8B-Instruct, Llama-3.1-70B, and Llama-3.1-70B-Instruct. From the Mistral family (Jiang et al., 2023, 2024), we include Mistral-7B-v0.3, Mistral-7B-Instruct-v0.3, Mixtral-8x7B-v0.1, and Mixtral-8x7B-Instruct-v0.1. From the Qwen3 family (Yang et al., 2025a), we include Qwen3-8B and Qwen3-32B. From the Gemma-3 family (Team et al., 2025), we include Gemma-3-4B-pt and Gemma-3-4B-it. The corresponding variation type, the number of attention layers (L), and the number of heads per layer (H) for each model are provided in Table 5.

A.2 Implementation details and baselines

This section describes the implementation details of SITE and provides concise descriptions and configurations of the baseline methods used in our experiments. For baseline methods, we largely follow the default configurations and settings provided in the original papers and official codebases, with minor adjustments noted below for compatibility or fairness.

- **SITE (Ours)**. Algorithm 1 details the optimization of the soft head-selection parameters (Stage 2 of our method, shown in Figure 2). We run the optimization for $J = 400$ iterations for all tasks and all 12 LLMs. Checkpoints are selected based on the lowest validation loss, evaluated every 50 iterations using up to 100 validation examples. All methods are evaluated using exact match between the ground-truth answer and the initial tokens of the model output. For tasks unrelated to capitalization or lowercasing, we allow case-insensitive matches, following prior work (Huang et al., 2024), as some tasks contain inconsistent capitalization in the ground-truth labels.

- **LoRA** (Hu et al., 2022). This method inserts trainable low-rank matrices into the weight matrices of a frozen LLM. We apply LoRA to all query and key projection layers, using rank 8, scaling factor $\alpha = 8$, and dropout rate 0.05. A learning rate of 0.0001 is used for all tasks, as higher learning rates occasionally led to unstable training.

- **(IA)^3** (Liu et al., 2022a). This method rescales key and value activations in attention layers, as well as the intermediate activations within multilayer perceptron (MLP) blocks, via element-wise multiplication with learnable vectors. We use a learning rate of 0.002 for all tasks, which is lightly tuned to improve performance.

- **Prompt Tuning** (Lester et al., 2021). This method prepends learnable continuous embeddings to the input embedding sequence of a frozen LLM. We use 20 virtual tokens and a learning rate of 0.05, which is lightly tuned to improve performance.

- **Function Vectors (FV)** (Todd et al., 2023). This method performs activation patching on attention head outputs using clean few-shot prompts and corrupted (shuffled) few-shot prompts, and selects heads whose patching most improves performance. The selected head outputs are summed to form a single task embedding, which is added to the activation of a heuristically chosen layer (typically around one-third of the model depth) during zero-shot inference. We adopt the hyperparameters specified for Llama-2-7B in the official repository and apply them to Llama-3.1-8B, as both models share the same number of layers and attention heads.

- **Task Vectors (TV)** (Hendel et al., 2023). This method extracts the last-token activation from the selected layer during few-shot inference and injects it by replacement into the same layer during zero-shot inference. The extraction / injection layer is selected via a validation sweep over all layers.

- **MTV** (Huang et al., 2024). This method learns a head-sampling distribution using the REINFORCE (Williams, 1992) algorithm. Task embeddings are constructed from attention head activations sampled according to this distribution and injected by replacement during zero-shot inference. While the original implementation uses 100 prompts with 4 demonstrations each for task embedding construction, we instead use 50 prompts with

Model Family	Model Name	Variation Type	Attention Layers (L)	Heads per Layer (H)
Llama-3.1	Llama-3.1-8B	Pretrained (Base)	32	32
	Llama-3.1-8B-Instruct	Instruction-tuned (Chat/Alignment)	32	32
	Llama-3.1-70B	Pretrained (Base)	80	64
	Llama-3.1-70B-Instruct	Instruction-tuned (Chat/Alignment)	80	64
Mistral	Mistral-7B-v0.3	Pretrained (Base)	32	32
	Mistral-7B-Instruct-v0.3	Instruction-tuned (Chat/Alignment)	32	32
	Mixtral-8x7B-v0.1	Mixture of Experts, Pretrained	32	32
	Mixtral-8x7B-Instruct-v0.1	Mixture of Experts, Instruction-tuned	32	32
Qwen3	Qwen3-8B	Instruction-tuned (Chat/Alignment)	36	32
	Qwen3-32B	Instruction-tuned (Chat/Alignment)	64	64
Gemma-3	Gemma-3-4B-pt	Pretrained (Base)	34	8
	Gemma-3-4B-it	Instruction-tuned (Chat/Alignment)	34	8

Table 5: **Models used for evaluation.** We consider 12 large language models spanning four model families, three variation types, and model sizes ranging from 4B to 70B parameters.

10 demonstrations per prompt to match our $M = 50$ setting.

- **LIVE** (Peng et al., 2024). This method learns task embeddings by introducing and optimizing trainable layer-wise vectors with associated scaling factors, which are added to the corresponding layer activations. These parameters are optimized end-to-end using a combination of cross-entropy and KL-divergence losses on the output logits. We follow the settings used for VQAv2 (Goyal et al., 2017) in the original work, except for increasing the batch size from 2 to 4 and training each task for 50 epochs on its full training set.

- **I2CL** (Li et al., 2024). This method constructs task embeddings by averaging last-token activations from multiple 1-shot prompts, separately for the multi-head attention (MHA) and multilayer perceptron (MLP) blocks at each layer. These embeddings are added to the corresponding layer activations, with the injection strength learned end-to-end. Although originally designed for classification tasks, we adapt I2CL to open-ended generation and increase the number of prompts used for task embedding construction to 50.

- **Iterative Vectors (IV)** (Liu and Deng, 2025). This method defines iterative vectors as differences between attention layer activations with and without demonstrations. These vectors are accumulated over multiple minibatches and injected additively into the corresponding layer activations during inference.

A.3 Overview of all 57 tasks in the FV benchmark

In Section 5.2, we conduct inter-task activation patching analysis using all 57 tasks from the FV (Todd et al., 2023) benchmark. This benchmark contains both closely related and contrasting task variants, making it well suited for analyzing task-specific in-context learning behavior. It comprises 29 abstractive and 28 extractive tasks, where abstractive tasks require generating information not explicitly present in the prompt, while extractive tasks involve retrieving answers directly from it. Many tasks were introduced by FV, while others originate from prior work and were subsequently filtered or reformatted by FV, including AG_News (Zhang et al., 2015), Antonym (Nguyen et al., 2017), Synonym (Nguyen et al., 2017), Commonsense_QA (Talmor et al., 2018), English-French (Conneau et al., 2017), English-German (Conneau et al., 2017), English-Spanish (Conneau et al., 2017), Landmark-Country (Hernandez et al., 2023), Person-Instrument (Hernandez et al., 2023), Person-Occupation (Hernandez et al., 2023), Person-Sport (Hernandez et al., 2023), Product-Company (Hernandez et al., 2023), Sentiment (Socher et al., 2013; Honovich et al., 2022), Conll2003_Location (Sang and De Meulder, 2003), Conll2003_Organization (Sang and De Meulder, 2003), and Conll2003_Person (Sang and De Meulder, 2003). For completeness and clarity, we present task descriptions and input-output examples for all 57 FV tasks in Tables 6-9.

Algorithm 1 Optimization of Soft Head-Selection Parameters

Require: L: Number of attention layers in LLM
Require: H: Number of attention heads per layer
Require: J: Number of training iterations
Require: $\mathbb{T}$: Training set of tuples (0-shot prompt, ground-truth answer)
Require: $\{\mathbf{t}^{(l,h)}\}_{l=1,\ldots,L;\, h=1,\ldots,H}$: Task embeddings for a given task
Require: $\text{SA}^{(l,h)}(\cdot)$: Self-attention at head h in layer l (before output projection)
Require: $W_o^{(l)}$: Output projection of the multi-head self-attention block in layer l
Require: $\text{MLP}^{(l)}(\cdot)$: MLP block of layer l

1: Initialize soft head-selection parameters $\mathbf{A} = [\alpha^{(l,h)}]_{l=1,\ldots,L;\, h=1,\ldots,H} \in \mathbb{R}^{L\times H}$ with $\alpha^{(l,h)} = \sigma(0) = 0.5$ for all l, h, where $\sigma(\cdot)$ denotes the sigmoid function
2: **for** each iteration $j = 1, 2, \ldots, J$ **do**
3: Sample $(P_j, a_j) \sim \mathbb{T}$
4: $\mathbf{e} \leftarrow \text{Tokenize}(P_j)$
5: $\mathbf{v}_1 \leftarrow \text{Embed}(\mathbf{e})$
6: **for** all $l = 1, 2, \ldots, L$ **do**
7: $[\mathbf{u}^{(l,1)}, \mathbf{u}^{(l,2)}, \ldots, \mathbf{u}^{(l,H)}] \leftarrow [\text{SA}^{(l,1)}(\mathbf{v}_1), \text{SA}^{(l,2)}(\mathbf{v}_1), \ldots, \text{SA}^{(l,H)}(\mathbf{v}_1)]$
8: **for** all $h = 1, 2, \ldots, H$ **do**
9: $\mathbf{u}^{(l,h)}[-1, :] \leftarrow (1 - \alpha^{(l,h)}) \cdot \mathbf{u}^{(l,h)}[-1, :] + \alpha^{(l,h)} \cdot \mathbf{t}^{(l,h)}$
10: **end for**
11: $\mathbf{v} \leftarrow (\mathbf{u}^{(l,1)} \oplus \mathbf{u}^{(l,2)} \oplus \cdots \oplus \mathbf{u}^{(l,H)}) W_o^{(l)}$
12: $\mathbf{v}_2 \leftarrow \mathbf{v}_1 + \mathbf{v}$
13: $\mathbf{v} \leftarrow \text{MLP}^{(l)}(\mathbf{v}_2)$
14: $\mathbf{v}_1 \leftarrow \mathbf{v}_2 + \mathbf{v}$
15: **end for**
16: Compute output logits: $\mathbf{p}_j \leftarrow \text{OutputProj}(\mathbf{v}_1)$
17: $\mathcal{L}_j \leftarrow \text{CrossEntropy}(\mathbf{p}_j, a_j)$
18: Update $\mathbf{A}$ with the Adam optimizer: $\mathbf{A} \leftarrow \text{Adam}(\mathbf{A}, \nabla_{\mathbf{A}} \mathcal{L}_j)$
19: **end for**
20: **Return:** Optimized soft head-selection parameters $\mathbf{A} = [\alpha^{(l,h)}]_{l=1,\ldots,L;\, h=1,\ldots,H} \in \mathbb{R}^{L\times H}$
21: **Note:* $\oplus$ denotes concatenation.
22: **Note:* $\mathbf{o}^{(l,h)}$ in Eq. 4 corresponds to $\mathbf{o}^{(l,h)} = \mathbf{u}^{(l,h)}[-1, :] \in \mathbb{R}^{d_v}$.
23: **Note:* Although Lines 8–10 are written with a loop for clarity, they are implemented as a vectorized operation in practice.

Task Name	Task Description Input-Output Example
AG_News	Classify a news article based on its headline and opening sentences into one of: *World*, *Sports*, *Business*, or *Science/Technology*. **Input:** Surviving Biotech's Downturns Charly Travers offers advice on withstanding the volatility of the biotech sector. **Output:** Business
Antonym	Generate the antonym of a given word. **Input:** overnight **Output:** daytime
Capitalize	Capitalize the given word. **Input:** without **Output:** Without
Capitalize_First_Letter	Generate the first letter of a given word in capital form. **Input:** deliver **Output:** D
Capitalize_Last_Letter	Generate the last letter of a given word in capital form. **Input:** clean **Output:** N
Capitalize_Second_Letter	Generate the second letter of a given word in capital form. **Input:** amazing **Output:** M
Commonsense_QA	Select the most plausible answer to a commonsense question from five given options. **Input:** Sammy wanted to go to where the people were. Where might he go? a: race track b: populated areas c: the desert d: apartment e: roadblock **Output:** b
Country-Capital	Generate the capital city of a given country. **Input:** United States of America **Output:** Washington, D.C.
Country-Currency	Generate the currency used in a given country. **Input:** Singapore **Output:** Singapore Dollar (SGD)
English-French	Translate the given English word into French. **Input:** know **Output:** savoir
English-German	Translate the given English word into German. **Input:** drink **Output:** trinken
English-Spanish	Translate the given English word into Spanish. **Input:** sometimes **Output:** a veces
Landmark-Country	Generate the country of a given landmark. **Input:** South East Forests National Park **Output:** Australia
Lowercase_First_Letter	Generate the first letter of a given word in lowercase. **Input:** CLEVER **Output:** c
Lowercase_Last_Letter	Generate the last letter of a given word in lowercase. **Input:** PILLOW **Output:** w

Table 6: **Task descriptions and input-output examples for 57 FV tasks (Part 1 of 4).** This table provides task names, descriptions, and representative input-output examples for the FV tasks used in our experiments. The remaining tasks are provided in Tables 7-9.

Task Name	Task Description Input-Output Example
National_Parks	Generate the U.S. state of a given national park unit. **Input:** Glacier Bay National Park **Output:** Alaska
Next_Capital_Letter	Generate the next capital letter of the first letter in a given word. **Input:** microphone **Output:** N
Next_Item	Generate the next item in a known sequence (e.g., days, months, letters, or numbers). **Input:** Friday **Output:** Saturday
Park-Country	Generate the country of a given national park. **Input:** Dartmoor National Park **Output:** United Kingdom
Person-Instrument	Generate the musical instrument played by a given musician. **Input:** Andor Toth **Output:** violin
Person-Occupation	Generate the occupation of a given individual. **Input:** Li Yining **Output:** economist
Person-Sport	Generate the sport played by a given athlete. **Input:** Andrea Pirlo **Output:** soccer
Present-Past	Generate the past-tense form of a given present-tense verb. **Input:** write **Output:** wrote
Prev_Item	Generate the previous item in a known sequence (e.g., days, months, letters, or numbers). **Input:** april **Output:** march
Product-Company	Generate the company associated with a given commercial product. **Input:** Wii Balance Board **Output:** Nintendo
Sentiment	Generate the sentiment of a given movie review. **Input:** Very well-written and very well-acted. **Output:** positive
Singular-Plural	Generate the plural form of a given singular noun. **Input:** island **Output:** islands
Synonym	Generate the synonym of a given word. **Input:** identify **Output:** recognize
Word_Length	Generate the number of letters in a given word. **Input:** discuss **Output:** 7
Adjective_V_Verb_3	Select the adjective from a list of 3 words (1 adjective, 2 verbs). **Input:** prepare, faithful, develop **Output:** faithful

Table 7: **Task descriptions and input-output examples for 57 FV tasks (Part 2 of 4).** This table continues from Table 6, providing task names, descriptions, and representative input-output examples for the FV tasks used in our experiments. The remaining tasks are provided in Tables 8-9.

Task Name	Task Description Input-Output Example
Adjective_V_Verb_5	Select the adjective from a list of 5 words (1 adjective, 4 verbs). **Input:** remember, teach, knowledgeable, doubt, write **Output:** knowledgeable
Alphabetically_First_3	Select the word that comes first in alphabetical order from a list of 3 words. **Input:** grapefruit, thoughtful, diligent **Output:** diligent
Alphabetically_First_5	Select the word that comes first in alphabetical order from a list of 5 words. **Input:** test, prepare, hammer, beyond, pigeon **Output:** beyond
Alphabetically_Last_3	Select the word that comes last in alphabetical order from a list of 3 words. **Input:** sample, garlic, cream **Output:** sample
Alphabetically_Last_5	Select the word that comes last in alphabetical order from a list of 5 words. **Input:** about, navy, gentle, duster, green **Output:** navy
Animal_V_Object_3	Select the animal from a list of 3 words (1 animal, 2 non-animals). **Input:** lettuce, basketball, dog **Output:** dog
Animal_V_Object_5	Select the animal from a list of 5 words (1 animal, 4 non-animals). **Input:** soda, rice, potato, snorkel, sloth **Output:** sloth
Choose_First_Of_3	Select the first word from a list of 3 words. **Input:** ostrich, since, out **Output:** ostrich
Choose_First_Of_5	Select the first word from a list of 5 words. **Input:** reach, puzzle, passionate, silver, complete **Output:** reach
Choose_Last_Of_3	Select the last word from a list of 3 words. **Input:** salmon, rice, socks **Output:** socks
Choose_Last_Of_5	Select the last word from a list of 5 words. **Input:** spicy, cowardly, hoop, komodo, toward **Output:** toward
Choose_Middle_Of_3	Select the middle word from a list of 3 words. **Input:** garlic, candle, argue **Output:** candle
Choose_Middle_Of_5	Select the middle word from a list of 5 words. **Input:** table, qualify, airplane, harmonious, happy **Output:** airplane
Color_V_Animal_3	Select the color from a list of 3 words (1 color, 2 animals). **Input:** camel, penguin, brown **Output:** brown
Color_V_Animal_5	Select the color from a list of 5 words (1 color, 4 animals). **Input:** salamander, chinchilla, flamingo, black, tiger **Output:** black

Table 8: **Task descriptions and input-output examples for 57 FV tasks (Part 3 of 4).** This table continues from Tables 6-7, providing task names, descriptions, and representative input-output examples for the FV tasks used in our experiments. The remaining tasks are provided in Table 9.

Task Name	Task Description Input-Output Example
Concept_V_Object_3	Select the concept from a list of 3 words (1 abstract concept, 2 concrete entities). **Input:** radio, whimsical, robot **Output:** whimsical
Concept_V_Object_5	Select the concept from a list of 5 words (1 abstract concept, 4 concrete entities). **Input:** towel, map, hammock, read, blanket **Output:** read
Conll2003_Location	Select the location entity from a given sentence. **Input:** Clinton arrives in Chicago on day of re-nomination. **Output:** Chicago
Conll2003_Organization	Select the organization entity from a given sentence. **Input:** Advertising revenues at The Times grew 20 percent. **Output:** The Times
Conll2003_Person	Select the person entity from a given sentence. **Input:** They contained $ 650,000 in jewelry and $ 40,000 in cash, Andrews said. **Output:** Andrews
Fruit_V_Animal_3	Select the fruit from a list of 3 words (1 fruit, 2 animals). **Input:** pineapple, iguana, leopard **Output:** pineapple
Fruit_V_Animal_5	Select the fruit from a list of 5 words (1 fruit, 4 animals). **Input:** walrus, lizard, panther, lion, cranberry **Output:** cranberry
Object_V_Concept_3	Select the concrete entity from a list of 3 words (1 concrete entity, 2 abstract concepts). **Input:** need, lamp, beneath **Output:** lamp
Object_V_Concept_5	Select the concrete entity from a list of 5 words (1 concrete entity, 4 abstract concepts). **Input:** passionate, jigsaw, remove, expensive, fearless **Output:** jigsaw
Squad_Val	Retrieve the answer to a given question based on a provided context paragraph. **Input:** The Panthers offense, which led the NFL in scoring (500 points), was loaded with talent, boasting six Pro Bowl selections. Pro Bowl quarterback Cam Newton had one of his best seasons, throwing for 3,837 yards and rushing for 636, while recording a career-high and league-leading 45 total touchdowns (35 passing, 10 rushing), a career-low 10 interceptions, and a career-best quarterback rating of 99.4. Newton's leading receivers were tight end Greg Olsen, who caught a career-high 77 passes for 1,104 yards and seven touchdowns, and wide receiver Ted Ginn, Jr., who caught 44 passes for 739 yards and 10 touchdowns; Ginn also rushed for 60 yards and returned 27 punts for 277 yards. Other key receivers included veteran Jerricho Cotchery (39 receptions for 485 yards), rookie Devin Funchess (31 receptions for 473 yards and five touchdowns), and second-year receiver Corey Brown (31 receptions for 447 yards). The Panthers backfield featured Pro Bowl running back Jonathan Stewart, who led the team with 989 rushing yards and six touchdowns in 13 games, along with Pro Bowl fullback Mike Tolbert, who rushed for 256 yards and caught 18 passes for another 154 yards. Carolina's offensive line also featured two Pro Bowl selections: center Ryan Kalil and guard Trai Turner. What position does Jerricho Cotchery play? **Output:** receivers
Verb_V_Adjective_3	Select the verb from a list of 3 words (1 verb, 2 adjectives). **Input:** dirty, dance, diligent **Output:** dance
Verb_V_Adjective_5	Select the verb from a list of 5 words (1 verb, 4 adjectives). **Input:** heavy, overcome, quick, modern, dazzling **Output:** overcome

Table 9: **Task descriptions and input-output examples for 57 FV tasks (Part 4 of 4).** This table concludes the series from Tables 6-8, providing task names, descriptions, and representative input-output examples for the FV tasks used in our experiments.

A.4 Ablation study on prompt templates

For all experiments, we adopt FV's default ICL prompt template: 'Q:{x_{i1}}\nA:{y_{i1}}\n\n . . . Q:{x_{iN}}\nA:{y_{iN}}\n\nQ:{x_{iq}}\nA:', where each $\{x_{ik}\}$ and $\{y_{ik}\}$ is replaced with the corresponding input-output pair. To assess the robustness of our method to prompt formatting, we conduct an ablation study using five prompt templates, each provided by FV (Todd et al., 2023), including the default template used in all other experiments. These templates are listed in Table 10. For each template, we evaluate our method along with 0-shot and 10-shot ICL across all 57 FV tasks using Llama-3.1-8B, with results reported in Table 11. Across all five templates, our method consistently achieves strong performance, with average accuracies ranging from 89.0% to 91.2%, significantly outperforming the 10-shot ICL (76.8%-77.8%). These results demonstrate the robustness of our method to variations in prompt format.

Prompt Template	**Format of a single** $(\{x_{ik}\}, \{y_{ik}\})$ **pair**
Template 1 (Default)	Q:{x_{ik}}\nA:{y_{ik}}\n\n
Template 2	question:{x_{ik}}\nanswer:{y_{ik}}\n\n
Template 3	A:{x_{ik}}\nB:{y_{ik}}\n\n
Template 4	{x_{ik}} →{y_{ik}}\n\n
Template 5	input:{x_{ik}} output:{y_{ik}}\n

Table 10: **Prompt templates used in the ablation study.** Each template shows how a single input-output pair $(\{x_{ik}\}, \{y_{ik}\})$ is formatted. All templates are sourced from FV (Todd et al., 2023). Template 1 serves as the default prompt format used in all main experiments.

B Task-wise performance of the FV benchmark across 12 LLMs

Figure 3 in Section 4.2 shows the average performance of our method across all 57 FV tasks, compared to the 10-shot ICL, for all 12 LLMs listed in Table 5. Task-wise results for all 12 LLMs are provided in Tables 12–23.

Task Type	Template 1 (default)			Template 2			Template 3			Template 4			Template 5		
	0-shot	**10-shot**	**Ours**	**0-shot**	**10-shot**	**Ours**	**0-shot**	**10-shot**	**Ours**	**0-shot**	**10-shot**	**Ours**	**0-shot**	**10-shot**	**Ours**
Abstractive	2.7	75.9	87.7	3.6	77.0	87.8	3.7	75.1	87.4	3.3	75.9	87.6	2.4	76.4	86.3
Extractive	17.3	77.6	92.4	26.9	76.6	90.9	23.2	79.2	93.0	5.0	78.7	94.8	25.0	79.2	91.7
All (57 FV tasks)	9.9	<u>76.8</u>	**90.0**	15.0	<u>76.8</u>	**89.3**	13.3	<u>77.1</u>	**90.2**	4.2	<u>77.3</u>	**91.2**	13.5	<u>77.8</u>	**89.0**

Table 11: **Results of prompt template ablation using Llama-3.1-8B.** Average accuracies for our method and the 0-shot/10-shot ICL baselines are reported across five prompt templates. Results are shown for 29 abstractive tasks, 28 extractive tasks, and all 57 FV tasks, respectively. Our method consistently demonstrates strong performance across all templates.

Task Name	0-shot	10-shot	Ours
AG_News	0.3	79.7	84.3
Antonym	0.4	66.3	66.7
Capitalize	1.2	99.4	99.4
Capitalize_First_Letter	1.2	99.4	99.4
Capitalize_Last_Letter	0.6	15.8	71.9
Capitalize_Second_Letter	0.0	25.5	64.2
Commonsense_QA	30.5	67.5	59.8
Country-Capital	0.0	92.9	88.1
Country-Currency	0.0	81.0	78.6
English-French	0.5	81.2	75.7
English-German	0.4	75.5	69.9
English-Spanish	0.3	84.3	81.6
Landmark-Country	1.7	82.3	78.3
Lowercase_First_Letter	0.0	100.0	100.0
Lowercase_Last_Letter	0.6	37.4	94.7
National_Parks	7.4	79.0	79.0
Next_Capital_Letter	2.3	1.8	35.7
Next_Item	0.0	89.4	91.5
Park-Country	21.7	82.2	77.1
Person-Instrument	0.9	65.4	69.2
Person-Occupation	0.0	52.3	64.5
Person-Sport	0.0	94.0	98.5
Present-Past	1.6	100.0	100.0
Prev_Item	0.0	66.0	83.0
Product-Company	2.8	78.9	78.9
Sentiment	0.0	95.9	94.7
Singular-Plural	2.3	100.0	100.0
Synonym	6.8	50.0	52.3
Word_Length	0.0	18.1	28.1
Average	2.9	<u>71.1</u>	**78.1**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	16.7	74.3	95.2
Adjective_V_Verb_5	15.2	66.7	94.3
Alphabetically_First_3	24.3	29.5	33.3
Alphabetically_First_5	22.9	23.3	28.6
Alphabetically_Last_3	21.9	30.5	37.1
Alphabetically_Last_5	11.9	19.1	29.5
Animal_V_Object_3	10.5	70.5	98.1
Animal_V_Object_5	16.7	64.8	95.7
Choose_First_Of_3	65.2	99.5	100.0
Choose_First_Of_5	82.9	98.6	100.0
Choose_Last_Of_3	4.8	96.7	99.5
Choose_Last_Of_5	0.0	92.4	100.0
Choose_Middle_Of_3	1.4	55.7	94.8
Choose_Middle_Of_5	0.5	21.9	51.9
Color_V_Animal_3	12.9	83.8	100.0
Color_V_Animal_5	11.0	84.3	99.1
Concept_V_Object_3	20.5	70.5	94.8
Concept_V_Object_5	17.1	62.9	95.7
Conll2003_Location	9.3	82.1	91.3
Conll2003_Organization	12.6	75.6	88.4
Conll2003_Person	12.9	87.9	95.7
Fruit_V_Animal_3	6.2	74.8	98.6
Fruit_V_Animal_5	3.3	71.0	99.5
Object_V_Concept_3	15.2	71.4	97.1
Object_V_Concept_5	14.8	61.9	96.2
Squad_Val	53.1	85.8	86.0
Verb_V_Adjective_3	11.4	67.1	98.1
Verb_V_Adjective_5	5.2	71.0	98.6
Average	17.9	<u>67.6</u>	**85.6**

(b) Extractive task results

Table 12: **Task-wise performance on 57 FV tasks using Gemma-3-4B-pt.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.0	75.4	84.9
Antonym	0.0	66.3	70.2
Capitalize	0.0	99.4	99.4
Capitalize_First_Letter	4.7	96.5	100.0
Capitalize_Last_Letter	2.9	18.1	79.0
Capitalize_Second_Letter	7.3	18.8	93.9
Commonsense_QA	59.6	68.8	61.6
Country-Capital	0.0	90.5	90.5
Country-Currency	0.0	69.1	78.6
English-French	0.3	81.4	71.4
English-German	0.1	75.7	64.2
English-Spanish	0.0	84.2	81.2
Landmark-Country	0.0	76.6	74.9
Lowercase_First_Letter	0.0	99.4	100.0
Lowercase_Last_Letter	0.0	39.8	90.6
National_Parks	0.0	69.5	70.5
Next_Capital_Letter	2.3	4.7	65.5
Next_Item	2.1	97.9	95.7
Park-Country	0.0	76.4	72.6
Person-Instrument	0.0	48.6	57.0
Person-Occupation	0.0	39.5	62.8
Person-Sport	0.0	94.0	97.0
Present-Past	0.0	100.0	98.4
Prev_Item	2.1	74.5	87.2
Product-Company	2.8	72.5	66.1
Sentiment	0.0	91.0	94.3
Singular-Plural	0.0	100.0	97.7
Synonym	6.0	50.3	51.7
Word_Length	0.0	38.0	68.4
Average	3.1	<u>69.5</u>	**80.2**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	3.8	77.6	97.6
Adjective_V_Verb_5	15.7	72.9	95.7
Alphabetically_First_3	21.9	28.1	31.9
Alphabetically_First_5	22.9	21.4	26.7
Alphabetically_Last_3	12.9	39.1	39.5
Alphabetically_Last_5	16.7	25.7	29.1
Animal_V_Object_3	21.9	81.0	98.6
Animal_V_Object_5	23.3	91.0	97.1
Choose_First_Of_3	31.0	99.5	100.0
Choose_First_Of_5	25.7	99.5	100.0
Choose_Last_Of_3	10.0	99.1	100.0
Choose_Last_Of_5	15.7	94.8	100.0
Choose_Middle_Of_3	7.6	70.0	94.3
Choose_Middle_Of_5	12.4	35.2	66.7
Color_V_Animal_3	17.1	71.4	100.0
Color_V_Animal_5	13.8	79.5	99.1
Concept_V_Object_3	27.6	61.9	97.6
Concept_V_Object_5	34.3	75.2	93.3
Conll2003_Location	7.7	86.4	92.4
Conll2003_Organization	19.7	77.8	88.2
Conll2003_Person	23.1	93.4	96.4
Fruit_V_Animal_3	11.4	74.8	97.6
Fruit_V_Animal_5	4.8	82.9	100.0
Object_V_Concept_3	3.3	77.6	97.6
Object_V_Concept_5	3.3	67.6	96.2
Squad_Val	71.7	87.9	86.4
Verb_V_Adjective_3	3.8	69.5	95.7
Verb_V_Adjective_5	6.7	71.4	99.1
Average	17.5	<u>71.9</u>	**86.3**

(b) Extractive task results

Table 13: **Task-wise performance on 57 FV tasks using Gemma-3-4B-it.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.4	81.0	88.9
Antonym	7.9	68.3	67.9
Capitalize	6.5	100.0	100.0
Capitalize_First_Letter	6.5	90.0	99.4
Capitalize_Last_Letter	0.6	33.9	86.6
Capitalize_Second_Letter	0.6	27.9	96.4
Commonsense_QA	21.1	70.8	59.0
Country-Capital	4.8	90.5	88.1
Country-Currency	0.0	78.6	78.6
English-French	0.3	79.8	77.7
English-German	1.4	74.0	63.2
English-Spanish	0.3	84.6	79.6
Landmark-Country	0.0	85.1	82.9
Lowercase_First_Letter	0.0	83.0	100.0
Lowercase_Last_Letter	0.0	49.1	95.3
National_Parks	1.1	79.0	77.9
Next_Capital_Letter	0.6	5.3	98.8
Next_Item	0.0	97.9	97.9
Park-Country	0.0	87.3	79.6
Person-Instrument	0.0	75.7	76.6
Person-Occupation	0.0	59.9	70.0
Person-Sport	0.0	92.5	97.0
Present-Past	1.6	98.4	100.0
Prev_Item	0.0	91.5	95.7
Product-Company	0.9	82.6	80.7
Sentiment	0.0	94.7	93.9
Singular-Plural	2.3	97.7	97.7
Synonym	1.7	51.2	47.9
Word_Length	0.0	31.6	63.7
Average	2.0	<u>73.8</u>	**84.2**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	34.8	76.7	98.1
Adjective_V_Verb_5	16.7	79.1	97.1
Alphabetically_First_3	31.0	32.9	53.8
Alphabetically_First_5	20.5	22.9	86.2
Alphabetically_Last_3	24.8	27.1	45.7
Alphabetically_Last_5	11.0	18.6	51.4
Animal_V_Object_3	24.3	71.9	96.7
Animal_V_Object_5	24.3	88.1	99.1
Choose_First_Of_3	81.4	100.0	100.0
Choose_First_Of_5	73.8	100.0	99.1
Choose_Last_Of_3	2.9	99.5	100.0
Choose_Last_Of_5	1.9	97.1	100.0
Choose_Middle_Of_3	5.7	42.9	98.6
Choose_Middle_Of_5	0.5	33.3	70.5
Color_V_Animal_3	28.6	84.3	99.1
Color_V_Animal_5	17.6	85.2	99.1
Concept_V_Object_3	19.1	77.6	99.1
Concept_V_Object_5	17.1	88.1	97.1
Conll2003_Location	9.7	87.2	94.5
Conll2003_Organization	9.3	77.1	92.0
Conll2003_Person	9.7	92.1	97.6
Fruit_V_Animal_3	29.1	87.1	98.6
Fruit_V_Animal_5	13.3	93.3	98.6
Object_V_Concept_3	27.6	81.4	98.6
Object_V_Concept_5	14.3	81.0	97.6
Squad_Val	58.4	84.9	88.9
Verb_V_Adjective_3	24.3	67.6	97.1
Verb_V_Adjective_5	9.1	80.0	98.1
Average	22.9	<u>73.5</u>	**91.1**

(b) Extractive task results

Table 14: **Task-wise performance on 57 FV tasks using Mistral-7B-v0.3.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.0	79.5	88.4
Antonym	1.2	69.8	70.0
Capitalize	31.8	99.4	100.0
Capitalize_First_Letter	30.6	98.2	99.4
Capitalize_Last_Letter	0.6	30.4	90.6
Capitalize_Second_Letter	1.8	26.1	95.8
Commonsense_QA	24.0	71.8	66.1
Country-Capital	4.8	88.1	88.1
Country-Currency	0.0	78.6	71.4
English-French	0.4	82.4	79.3
English-German	1.5	75.5	60.5
English-Spanish	0.3	85.4	80.0
Landmark-Country	0.0	84.6	80.6
Lowercase_First_Letter	0.0	97.7	100.0
Lowercase_Last_Letter	0.0	42.7	95.3
National_Parks	1.1	80.0	75.8
Next_Capital_Letter	0.0	4.1	97.7
Next_Item	0.0	97.9	95.7
Park-Country	0.0	84.7	80.9
Person-Instrument	1.9	71.0	75.7
Person-Occupation	0.6	54.7	66.3
Person-Sport	0.0	92.5	94.0
Present-Past	3.3	98.4	100.0
Prev_Item	4.3	89.4	97.9
Product-Company	0.0	79.8	81.7
Sentiment	0.0	94.3	94.3
Singular-Plural	7.0	100.0	97.7
Synonym	1.2	53.2	49.5
Word_Length	0.6	53.2	62.6
Average	4.0	<u>74.6</u>	**84.0**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	14.8	78.6	99.1
Adjective_V_Verb_5	1.0	83.8	98.1
Alphabetically_First_3	16.7	31.9	45.2
Alphabetically_First_5	4.8	24.8	88.6
Alphabetically_Last_3	11.4	31.4	48.1
Alphabetically_Last_5	6.7	22.4	42.4
Animal_V_Object_3	10.5	84.3	97.1
Animal_V_Object_5	4.3	94.8	98.1
Choose_First_Of_3	41.4	99.5	99.1
Choose_First_Of_5	8.6	96.2	99.1
Choose_Last_Of_3	5.2	88.1	100.0
Choose_Last_Of_5	1.0	87.1	100.0
Choose_Middle_Of_3	5.7	45.2	96.2
Choose_Middle_Of_5	1.9	41.0	96.7
Color_V_Animal_3	14.3	82.4	100.0
Color_V_Animal_5	1.9	91.4	99.5
Concept_V_Object_3	8.1	88.1	97.6
Concept_V_Object_5	5.2	90.5	99.1
Conll2003_Location	5.6	84.0	95.2
Conll2003_Organization	9.7	79.6	92.8
Conll2003_Person	22.3	89.9	97.7
Fruit_V_Animal_3	13.8	94.3	99.1
Fruit_V_Animal_5	1.0	97.6	100.0
Object_V_Concept_3	15.2	78.6	98.1
Object_V_Concept_5	3.3	84.3	98.1
Squad_Val	60.4	87.5	88.1
Verb_V_Adjective_3	10.0	68.1	96.7
Verb_V_Adjective_5	2.9	79.5	97.6
Average	11.0	<u>75.2</u>	**91.7**

(b) Extractive task results

Table 15: **Task-wise performance on 57 FV tasks using Mistral-7B-Instruct-v0.3.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.4	79.4	86.8
Antonym	0.0	69.9	69.1
Capitalize	5.3	99.6	100.0
Capitalize_First_Letter	10.0	99.2	99.8
Capitalize_Last_Letter	1.2	20.9	94.0
Capitalize_Second_Letter	1.2	32.1	97.4
Commonsense_QA	40.3	72.4	63.5
Country-Capital	4.8	96.0	92.9
Country-Currency	0.0	80.2	81.7
English-French	0.5	81.4	80.5
English-German	1.2	76.1	68.6
English-Spanish	0.2	84.4	83.9
Landmark-Country	0.0	91.1	86.7
Lowercase_First_Letter	0.0	99.6	100.0
Lowercase_Last_Letter	0.0	35.7	95.9
National_Parks	0.0	84.6	80.4
Next_Capital_Letter	0.6	5.5	99.2
Next_Item	2.1	97.2	97.9
Park-Country	0.0	88.8	83.9
Person-Instrument	0.0	85.4	88.5
Person-Occupation	0.0	65.9	80.6
Person-Sport	0.0	95.0	96.5
Present-Past	3.3	99.5	100.0
Prev_Item	2.1	95.7	95.7
Product-Company	0.0	87.8	89.3
Sentiment	0.0	95.9	96.1
Singular-Plural	2.3	99.2	98.4
Synonym	1.8	50.4	52.9
Word_Length	0.0	33.7	83.0
Average	2.7	<u>75.9</u>	**87.7**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	14.3	86.7	99.7
Adjective_V_Verb_5	9.1	86.8	97.9
Alphabetically_First_3	21.9	42.1	56.2
Alphabetically_First_5	16.7	21.1	89.7
Alphabetically_Last_3	16.2	34.4	48.1
Alphabetically_Last_5	10.5	20.8	53.0
Animal_V_Object_3	12.4	81.3	99.2
Animal_V_Object_5	19.1	80.2	98.3
Choose_First_Of_3	52.9	98.9	100.0
Choose_First_Of_5	52.4	98.3	100.0
Choose_Last_Of_3	1.0	97.0	100.0
Choose_Last_Of_5	3.8	95.9	100.0
Choose_Middle_Of_3	2.9	53.7	98.6
Choose_Middle_Of_5	4.3	33.5	91.4
Color_V_Animal_3	16.7	94.9	99.5
Color_V_Animal_5	15.7	91.3	99.2
Concept_V_Object_3	14.3	83.7	99.8
Concept_V_Object_5	17.6	86.0	94.6
Conll2003_Location	21.8	87.7	94.3
Conll2003_Organization	39.3	78.0	91.3
Conll2003_Person	12.4	92.6	97.7
Fruit_V_Animal_3	23.3	81.9	98.6
Fruit_V_Animal_5	10.0	79.8	99.5
Object_V_Concept_3	17.6	96.7	100.0
Object_V_Concept_5	5.7	94.8	98.6
Squad_Val	39.4	85.6	86.7
Verb_V_Adjective_3	11.4	95.1	98.1
Verb_V_Adjective_5	1.9	94.6	98.3
Average	17.3	<u>77.6</u>	**92.4**

(b) Extractive task results

Table 16: **Task-wise performance on 57 FV tasks using Llama-3.1-8B.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.0	77.6	90.0
Antonym	0.4	70.8	71.6
Capitalize	0.6	99.4	99.4
Capitalize_First_Letter	2.4	100.0	100.0
Capitalize_Last_Letter	2.3	49.7	95.3
Capitalize_Second_Letter	1.8	49.7	100.0
Commonsense_QA	71.3	74.0	72.0
Country-Capital	2.4	90.5	90.5
Country-Currency	0.0	81.0	85.7
English-French	0.7	83.1	82.0
English-German	0.7	76.7	70.1
English-Spanish	0.2	84.8	84.3
Landmark-Country	0.0	88.0	82.9
Lowercase_First_Letter	0.0	100.0	99.4
Lowercase_Last_Letter	0.0	60.2	97.1
National_Parks	1.1	86.3	75.8
Next_Capital_Letter	1.2	2.9	99.4
Next_Item	0.0	97.9	97.9
Park-Country	1.3	89.2	84.1
Person-Instrument	0.0	82.2	88.8
Person-Occupation	0.0	65.1	77.3
Person-Sport	0.0	95.5	97.0
Present-Past	1.6	100.0	100.0
Prev_Item	4.3	91.5	95.7
Product-Company	1.8	84.4	84.4
Sentiment	0.0	94.7	97.1
Singular-Plural	4.7	100.0	95.4
Synonym	32.1	53.3	55.1
Word_Length	0.0	74.3	83.6
Average	4.5	<u>79.4</u>	**88.0**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	14.3	87.6	100.0
Adjective_V_Verb_5	13.3	91.0	97.1
Alphabetically_First_3	26.7	37.6	51.4
Alphabetically_First_5	18.6	22.4	92.4
Alphabetically_Last_3	16.2	37.6	49.1
Alphabetically_Last_5	11.0	28.6	74.3
Animal_V_Object_3	21.9	95.7	99.5
Animal_V_Object_5	6.7	96.7	100.0
Choose_First_Of_3	19.5	97.6	100.0
Choose_First_Of_5	9.1	97.1	100.0
Choose_Last_Of_3	29.5	93.3	100.0
Choose_Last_Of_5	26.2	94.3	100.0
Choose_Middle_Of_3	15.2	53.3	99.1
Choose_Middle_Of_5	9.5	33.8	96.7
Color_V_Animal_3	29.5	99.5	100.0
Color_V_Animal_5	8.6	97.6	100.0
Concept_V_Object_3	26.2	91.9	99.5
Concept_V_Object_5	17.1	95.2	97.1
Conll2003_Location	10.2	90.1	94.7
Conll2003_Organization	33.6	80.6	93.7
Conll2003_Person	22.9	93.4	97.5
Fruit_V_Animal_3	4.8	99.5	99.1
Fruit_V_Animal_5	0.5	98.1	99.5
Object_V_Concept_3	26.7	94.8	98.6
Object_V_Concept_5	17.1	94.8	99.5
Squad_Val	76.6	87.8	90.1
Verb_V_Adjective_3	10.5	96.7	98.6
Verb_V_Adjective_5	8.6	97.1	99.1
Average	18.9	<u>81.6</u>	**93.8**

(b) Extractive task results

Table 17: **Task-wise performance on 57 FV tasks using Llama-3.1-8B-Instruct.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.0	77.8	87.7
Antonym	0.0	69.4	66.5
Capitalize	19.4	99.4	98.8
Capitalize_First_Letter	16.5	97.7	100.0
Capitalize_Last_Letter	4.1	24.6	93.0
Capitalize_Second_Letter	11.5	30.3	97.6
Commonsense_QA	42.2	80.8	79.8
Country-Capital	4.8	88.1	88.1
Country-Currency	0.0	81.0	64.3
English-French	0.3	82.3	69.4
English-German	0.7	73.3	57.5
English-Spanish	0.2	84.4	73.9
Landmark-Country	0.0	81.1	77.7
Lowercase_First_Letter	0.0	98.8	100.0
Lowercase_Last_Letter	0.0	60.2	97.7
National_Parks	0.0	79.0	70.5
Next_Capital_Letter	0.6	8.8	92.4
Next_Item	0.0	95.7	95.7
Park-Country	0.0	80.9	70.1
Person-Instrument	0.0	62.6	62.6
Person-Occupation	0.0	46.5	60.5
Person-Sport	0.0	89.6	97.0
Present-Past	1.6	100.0	98.4
Prev_Item	4.3	97.9	93.6
Product-Company	0.0	77.1	80.7
Sentiment	0.0	95.1	95.9
Singular-Plural	4.7	97.7	90.7
Synonym	0.2	49.0	48.3
Word_Length	0.0	67.8	73.7
Average	3.8	<u>75.1</u>	**82.1**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	0.5	77.6	98.1
Adjective_V_Verb_5	0.0	87.1	98.6
Alphabetically_First_3	5.2	31.0	42.4
Alphabetically_First_5	2.4	20.5	89.5
Alphabetically_Last_3	2.9	38.1	42.4
Alphabetically_Last_5	2.4	20.0	80.0
Animal_V_Object_3	1.4	96.7	95.7
Animal_V_Object_5	0.5	98.1	96.7
Choose_First_Of_3	5.7	95.7	100.0
Choose_First_Of_5	1.0	94.8	100.0
Choose_Last_Of_3	2.4	91.4	100.0
Choose_Last_Of_5	1.4	66.2	100.0
Choose_Middle_Of_3	2.9	62.9	99.5
Choose_Middle_Of_5	0.5	31.4	98.1
Color_V_Animal_3	2.9	99.5	100.0
Color_V_Animal_5	4.3	98.6	100.0
Concept_V_Object_3	1.0	91.9	100.0
Concept_V_Object_5	0.0	90.5	95.7
Conll2003_Location	7.7	91.0	95.1
Conll2003_Organization	15.1	83.6	90.7
Conll2003_Person	26.8	94.0	96.6
Fruit_V_Animal_3	6.7	100.0	100.0
Fruit_V_Animal_5	5.2	99.5	99.5
Object_V_Concept_3	0.0	91.4	97.1
Object_V_Concept_5	0.0	93.8	96.7
Squad_Val	38.1	90.7	88.3
Verb_V_Adjective_3	1.9	91.9	99.5
Verb_V_Adjective_5	0.0	97.1	96.2
Average	5.0	<u>79.5</u>	**92.7**

(b) Extractive task results

Table 18: **Task-wise performance on 57 FV tasks using Qwen3-8B.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.1	81.8	88.2
Antonym	8.5	67.1	65.7
Capitalize	4.7	93.5	99.4
Capitalize_First_Letter	6.5	97.7	100.0
Capitalize_Last_Letter	7.0	24.0	95.3
Capitalize_Second_Letter	3.0	29.1	97.0
Commonsense_QA	38.9	86.2	84.5
Country-Capital	7.1	76.2	90.5
Country-Currency	0.0	81.0	81.0
English-French	0.4	80.2	77.2
English-German	0.5	76.1	66.2
English-Spanish	0.0	82.9	80.1
Landmark-Country	0.0	84.0	81.1
Lowercase_First_Letter	0.6	89.5	99.4
Lowercase_Last_Letter	0.6	39.8	95.3
National_Parks	0.0	83.2	79.0
Next_Capital_Letter	2.3	2.9	98.3
Next_Item	0.0	85.1	95.7
Park-Country	0.0	84.1	77.7
Person-Instrument	0.0	56.1	71.0
Person-Occupation	0.0	36.6	69.2
Person-Sport	0.0	88.1	95.5
Present-Past	0.0	83.6	98.4
Prev_Item	0.0	78.7	97.9
Product-Company	0.0	83.5	84.4
Sentiment	2.0	93.9	92.2
Singular-Plural	2.3	97.7	95.4
Synonym	1.2	47.5	50.2
Word_Length	0.0	77.8	73.1
Average	3.0	<u>72.0</u>	**85.5**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	6.2	78.1	99.5
Adjective_V_Verb_5	4.8	81.0	98.1
Alphabetically_First_3	7.1	38.1	97.6
Alphabetically_First_5	3.3	22.9	92.9
Alphabetically_Last_3	5.2	36.2	43.3
Alphabetically_Last_5	2.4	17.6	71.0
Animal_V_Object_3	11.0	97.6	99.1
Animal_V_Object_5	11.0	98.6	98.6
Choose_First_Of_3	14.8	95.7	100.0
Choose_First_Of_5	6.7	95.2	100.0
Choose_Last_Of_3	1.4	92.9	100.0
Choose_Last_Of_5	2.9	85.7	100.0
Choose_Middle_Of_3	1.4	56.2	100.0
Choose_Middle_Of_5	0.0	32.9	98.1
Color_V_Animal_3	4.8	96.2	100.0
Color_V_Animal_5	8.6	97.6	99.5
Concept_V_Object_3	1.9	88.1	99.1
Concept_V_Object_5	1.0	97.1	98.1
Conll2003_Location	2.1	89.4	95.7
Conll2003_Organization	8.7	83.8	92.3
Conll2003_Person	15.1	93.8	96.8
Fruit_V_Animal_3	5.2	100.0	100.0
Fruit_V_Animal_5	6.2	100.0	99.5
Object_V_Concept_3	4.3	96.2	99.5
Object_V_Concept_5	2.9	94.3	98.6
Squad_Val	31.6	90.8	90.5
Verb_V_Adjective_3	0.5	91.4	100.0
Verb_V_Adjective_5	1.4	98.6	99.5
Average	6.2	<u>80.2</u>	**95.3**

(b) Extractive task results

Table 19: **Task-wise performance on 57 FV tasks using Qwen3-32B.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.3	81.5	89.9
Antonym	3.8	70.4	67.3
Capitalize	8.8	99.4	100.0
Capitalize_First_Letter	8.2	97.7	100.0
Capitalize_Last_Letter	0.6	39.2	91.8
Capitalize_Second_Letter	0.0	32.1	95.8
Commonsense_QA	39.5	73.9	61.8
Country-Capital	4.8	90.5	85.7
Country-Currency	0.0	83.3	83.3
English-French	0.2	84.3	82.0
English-German	0.8	78.2	74.4
English-Spanish	0.2	86.4	87.6
Landmark-Country	0.0	90.3	85.1
Lowercase_First_Letter	0.0	93.6	100.0
Lowercase_Last_Letter	0.0	46.2	94.7
National_Parks	1.1	85.3	79.0
Next_Capital_Letter	1.8	5.3	98.3
Next_Item	0.0	97.9	97.9
Park-Country	0.0	91.7	87.9
Person-Instrument	0.0	84.1	87.9
Person-Occupation	0.0	77.9	82.6
Person-Sport	0.0	95.5	98.5
Present-Past	1.6	100.0	100.0
Prev_Item	2.1	97.9	97.9
Product-Company	0.0	89.9	88.1
Sentiment	0.0	96.3	95.1
Singular-Plural	2.3	100.0	100.0
Synonym	0.7	54.6	49.5
Word_Length	0.0	75.4	74.3
Average	2.6	<u>79.3</u>	**87.5**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	26.2	84.3	99.5
Adjective_V_Verb_5	18.1	83.8	97.6
Alphabetically_First_3	28.1	37.1	46.2
Alphabetically_First_5	20.5	22.9	89.1
Alphabetically_Last_3	16.2	39.1	51.0
Alphabetically_Last_5	13.8	20.5	46.2
Animal_V_Object_3	21.4	94.3	97.1
Animal_V_Object_5	24.8	91.9	98.1
Choose_First_Of_3	65.7	99.1	100.0
Choose_First_Of_5	73.3	99.1	100.0
Choose_Last_Of_3	2.9	99.5	100.0
Choose_Last_Of_5	2.4	95.7	99.5
Choose_Middle_Of_3	4.3	50.5	98.1
Choose_Middle_Of_5	1.4	28.1	89.5
Color_V_Animal_3	22.9	97.6	100.0
Color_V_Animal_5	9.1	97.1	99.5
Concept_V_Object_3	18.6	76.2	99.1
Concept_V_Object_5	13.8	86.7	95.7
Conll2003_Location	6.8	88.6	93.7
Conll2003_Organization	12.3	77.2	92.3
Conll2003_Person	8.1	93.8	97.9
Fruit_V_Animal_3	31.0	97.1	99.1
Fruit_V_Animal_5	6.7	97.1	99.5
Object_V_Concept_3	22.9	91.4	99.1
Object_V_Concept_5	12.9	86.7	97.1
Squad_Val	58.9	86.2	87.5
Verb_V_Adjective_3	8.1	70.5	96.7
Verb_V_Adjective_5	4.3	90.0	98.6
Average	19.8	<u>77.9</u>	**91.7**

(b) Extractive task results

Table 20: **Task-wise performance on 57 FV tasks using Mixtral-8x7B-v0.1.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.0	81.3	89.2
Antonym	0.0	72.4	67.7
Capitalize	4.7	99.4	100.0
Capitalize_First_Letter	8.2	99.4	100.0
Capitalize_Last_Letter	0.6	25.7	87.1
Capitalize_Second_Letter	0.6	27.9	92.1
Commonsense_QA	56.8	73.9	67.5
Country-Capital	4.8	90.5	83.3
Country-Currency	0.0	73.8	81.0
English-French	0.0	84.7	82.0
English-German	0.2	77.2	71.5
English-Spanish	0.1	85.8	85.4
Landmark-Country	0.0	92.0	82.3
Lowercase_First_Letter	0.0	97.1	100.0
Lowercase_Last_Letter	0.0	48.0	97.7
National_Parks	2.1	80.0	76.8
Next_Capital_Letter	0.6	5.3	99.4
Next_Item	0.0	97.9	95.7
Park-Country	0.0	90.5	87.9
Person-Instrument	0.0	84.1	88.8
Person-Occupation	0.0	74.4	82.6
Person-Sport	0.0	95.5	98.5
Present-Past	0.0	100.0	100.0
Prev_Item	0.0	93.6	95.7
Product-Company	0.0	91.7	88.1
Sentiment	0.0	94.7	93.9
Singular-Plural	0.0	100.0	100.0
Synonym	0.2	51.7	47.9
Word_Length	0.0	58.5	65.5
Average	2.7	<u>77.5</u>	**86.5**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	30.0	87.6	97.1
Adjective_V_Verb_5	17.6	89.5	98.1
Alphabetically_First_3	21.9	39.1	39.1
Alphabetically_First_5	14.3	24.8	77.6
Alphabetically_Last_3	14.8	30.5	48.6
Alphabetically_Last_5	8.6	24.3	42.9
Animal_V_Object_3	12.9	93.3	97.6
Animal_V_Object_5	8.1	95.7	98.6
Choose_First_Of_3	59.5	98.1	98.6
Choose_First_Of_5	35.7	94.8	99.1
Choose_Last_Of_3	5.7	97.1	100.0
Choose_Last_Of_5	6.7	95.7	100.0
Choose_Middle_Of_3	1.9	53.3	97.1
Choose_Middle_Of_5	2.4	28.1	79.1
Color_V_Animal_3	20.0	98.6	100.0
Color_V_Animal_5	1.0	96.7	99.1
Concept_V_Object_3	16.7	86.2	99.1
Concept_V_Object_5	7.1	92.9	96.7
Conll2003_Location	6.5	89.4	94.1
Conll2003_Organization	7.6	80.6	93.9
Conll2003_Person	20.8	92.6	96.6
Fruit_V_Animal_3	17.1	97.1	97.1
Fruit_V_Animal_5	1.0	97.6	98.6
Object_V_Concept_3	13.8	92.4	97.1
Object_V_Concept_5	6.7	87.6	96.2
Squad_Val	59.3	84.8	87.3
Verb_V_Adjective_3	14.8	71.4	97.6
Verb_V_Adjective_5	2.4	85.7	100.0
Average	15.5	<u>78.8</u>	**90.2**

(b) Extractive task results

Table 21: **Task-wise performance on 57 FV tasks using Mixtral-8x7B-Instruct-v0.1.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.4	84.3	91.0
Antonym	16.7	71.6	71.4
Capitalize	0.0	99.4	100.0
Capitalize_First_Letter	0.6	100.0	100.0
Capitalize_Last_Letter	0.0	35.1	97.7
Capitalize_Second_Letter	0.0	37.6	98.2
Commonsense_QA	31.1	78.7	73.9
Country-Capital	4.8	92.9	92.9
Country-Currency	0.0	78.6	83.3
English-French	0.3	85.5	85.6
English-German	1.0	81.5	80.0
English-Spanish	0.3	89.8	89.2
Landmark-Country	0.0	89.1	84.0
Lowercase_First_Letter	0.0	98.8	100.0
Lowercase_Last_Letter	0.0	42.7	99.4
National_Parks	20.0	81.1	75.8
Next_Capital_Letter	0.6	9.4	100.0
Next_Item	4.3	95.7	95.7
Park-Country	48.4	91.7	86.0
Person-Instrument	0.0	79.4	83.2
Person-Occupation	0.0	66.9	83.7
Person-Sport	0.0	97.0	98.5
Present-Past	1.6	100.0	100.0
Prev_Item	2.1	97.9	97.9
Product-Company	1.8	90.8	88.1
Sentiment	0.0	98.0	96.3
Singular-Plural	2.3	100.0	97.7
Synonym	2.7	55.6	60.3
Word_Length	0.0	77.2	87.1
Average	4.8	<u>79.5</u>	**89.5**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	29.1	89.1	100.0
Adjective_V_Verb_5	17.1	86.7	100.0
Alphabetically_First_3	30.0	37.1	98.6
Alphabetically_First_5	22.4	30.5	96.7
Alphabetically_Last_3	23.3	34.8	62.9
Alphabetically_Last_5	12.9	23.8	93.8
Animal_V_Object_3	22.4	98.6	97.6
Animal_V_Object_5	21.9	97.1	99.1
Choose_First_Of_3	81.9	100.0	100.0
Choose_First_Of_5	88.6	100.0	100.0
Choose_Last_Of_3	0.0	94.8	100.0
Choose_Last_Of_5	0.0	99.1	100.0
Choose_Middle_Of_3	0.5	68.1	98.6
Choose_Middle_Of_5	0.0	36.2	98.1
Color_V_Animal_3	26.7	99.5	100.0
Color_V_Animal_5	13.3	98.6	100.0
Concept_V_Object_3	20.5	93.3	99.1
Concept_V_Object_5	18.1	91.9	97.6
Conll2003_Location	20.6	92.3	96.8
Conll2003_Organization	36.9	85.1	93.6
Conll2003_Person	13.7	95.0	98.7
Fruit_V_Animal_3	17.6	100.0	99.5
Fruit_V_Animal_5	6.7	99.5	99.5
Object_V_Concept_3	21.4	98.6	99.1
Object_V_Concept_5	15.7	89.1	99.1
Squad_Val	48.8	88.4	90.4
Verb_V_Adjective_3	22.4	90.5	100.0
Verb_V_Adjective_5	12.9	96.7	99.5
Average	23.0	<u>82.7</u>	**97.1**

(b) Extractive task results

Table 22: **Task-wise performance on 57 FV tasks using Llama-3.1-70B.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

Task Name	0-shot	10-shot	Ours
AG_News	0.4	83.6	91.4
Antonym	23.4	70.6	71.0
Capitalize	4.1	100.0	99.4
Capitalize_First_Letter	3.5	100.0	100.0
Capitalize_Last_Letter	0.6	54.4	97.7
Capitalize_Second_Letter	0.6	41.8	98.8
Commonsense_QA	76.0	80.8	77.6
Country-Capital	7.1	90.5	88.1
Country-Currency	0.0	76.2	85.7
English-French	0.3	86.2	86.3
English-German	1.4	81.2	80.4
English-Spanish	0.5	88.6	88.2
Landmark-Country	1.7	88.0	84.6
Lowercase_First_Letter	0.0	100.0	100.0
Lowercase_Last_Letter	0.0	77.8	98.3
National_Parks	6.3	83.2	76.8
Next_Capital_Letter	1.2	18.1	98.8
Next_Item	6.4	97.9	95.7
Park-Country	3.8	89.8	84.1
Person-Instrument	0.0	76.6	81.3
Person-Occupation	0.0	63.4	72.7
Person-Sport	0.0	97.0	98.5
Present-Past	3.3	93.4	100.0
Prev_Item	8.5	97.9	97.9
Product-Company	1.8	87.2	88.1
Sentiment	0.0	94.3	95.9
Singular-Plural	4.7	100.0	100.0
Synonym	9.4	55.0	57.1
Word_Length	0.0	87.1	83.6
Average	5.7	<u>81.4</u>	**88.9**

(a) Abstractive task results

Task Name	0-shot	10-shot	Ours
Adjective_V_Verb_3	26.7	94.8	99.5
Adjective_V_Verb_5	8.1	95.7	99.5
Alphabetically_First_3	18.1	37.6	96.7
Alphabetically_First_5	6.2	25.7	96.2
Alphabetically_Last_3	14.3	38.6	93.3
Alphabetically_Last_5	0.5	21.0	78.6
Animal_V_Object_3	5.2	99.5	99.1
Animal_V_Object_5	1.0	100.0	99.5
Choose_First_Of_3	27.6	98.6	100.0
Choose_First_Of_5	17.6	98.6	100.0
Choose_Last_Of_3	5.2	94.8	100.0
Choose_Last_Of_5	2.9	92.4	100.0
Choose_Middle_Of_3	1.4	60.5	100.0
Choose_Middle_Of_5	1.4	31.9	99.5
Color_V_Animal_3	3.8	99.1	100.0
Color_V_Animal_5	0.5	98.1	100.0
Concept_V_Object_3	3.8	96.2	100.0
Concept_V_Object_5	1.4	97.1	99.1
Conll2003_Location	24.3	92.1	97.1
Conll2003_Organization	40.9	81.4	94.2
Conll2003_Person	21.9	96.9	98.1
Fruit_V_Animal_3	5.2	100.0	99.5
Fruit_V_Animal_5	1.4	100.0	99.5
Object_V_Concept_3	13.3	99.1	99.1
Object_V_Concept_5	4.8	94.8	100.0
Squad_Val	66.8	88.9	91.4
Verb_V_Adjective_3	11.9	94.3	99.1
Verb_V_Adjective_5	4.3	97.1	99.5
Average	12.2	<u>83.0</u>	**97.8**

(b) Extractive task results

Table 23: **Task-wise performance on 57 FV tasks using Llama-3.1-70B-Instruct.** Our method is evaluated along with 0-shot and 10-shot ICL baselines. (a) Results on 29 abstractive tasks. (b) Results on 28 extractive tasks. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

C Extended results on optimized soft head-selection values

C.1 Optimized soft head-selection values for all 57 FV tasks

In this section, we present extended results of Figure 4 in Section 5.1, showing the optimized values of the soft head-selection parameters for all 57 FV tasks using Llama-3.1-8B. The full set of results is provided in Figures 7-9

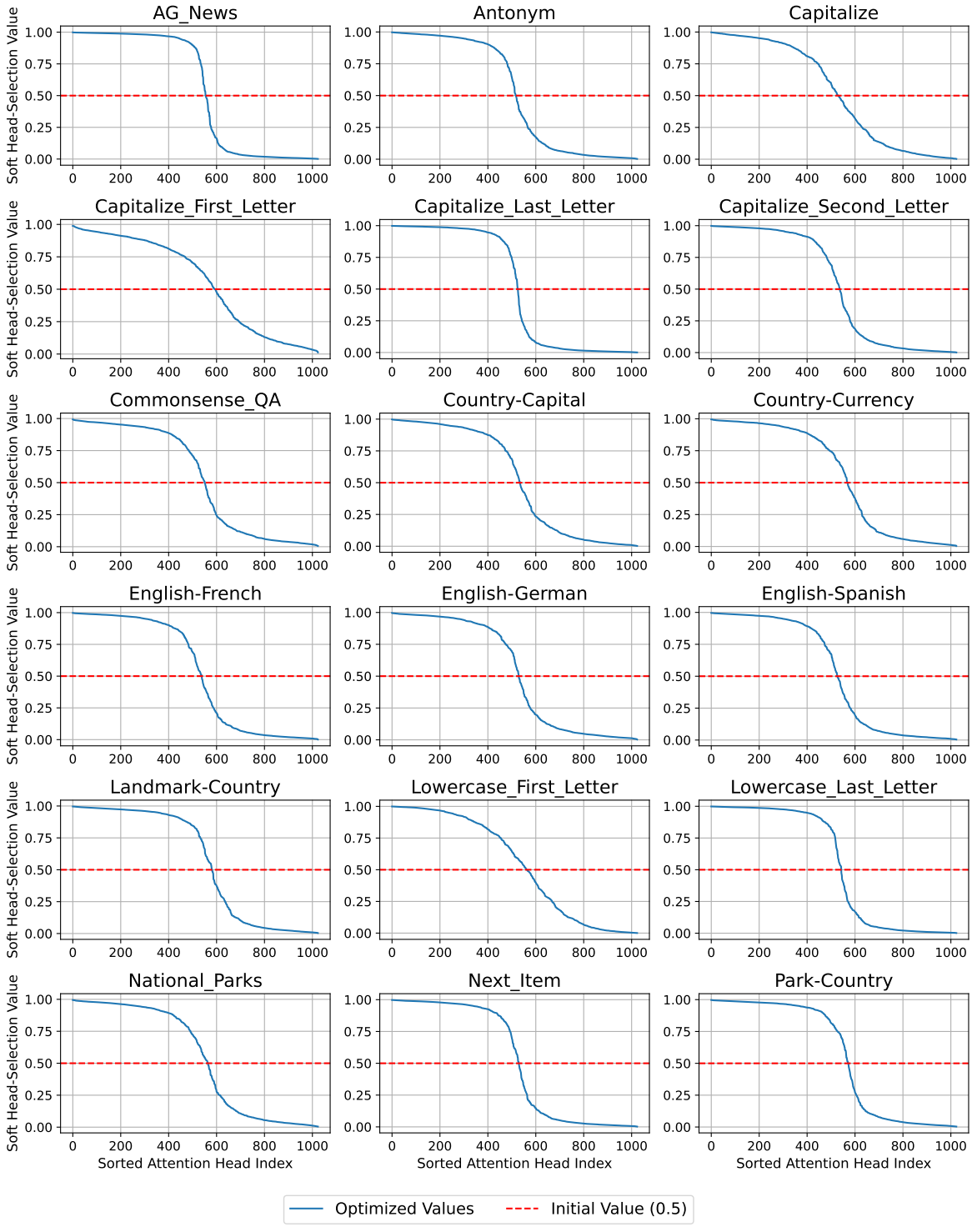

Figure 7: **Optimized values of the soft head-selection parameters for 57 FV tasks (Part 1 of 3).** Each plot shows the optimized values of the soft head-selection parameters for all 1024 attention heads in Llama-3.1-8B, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all selection parameters at the start of training. Plots for the remaining tasks are provided in Figures 8-9.

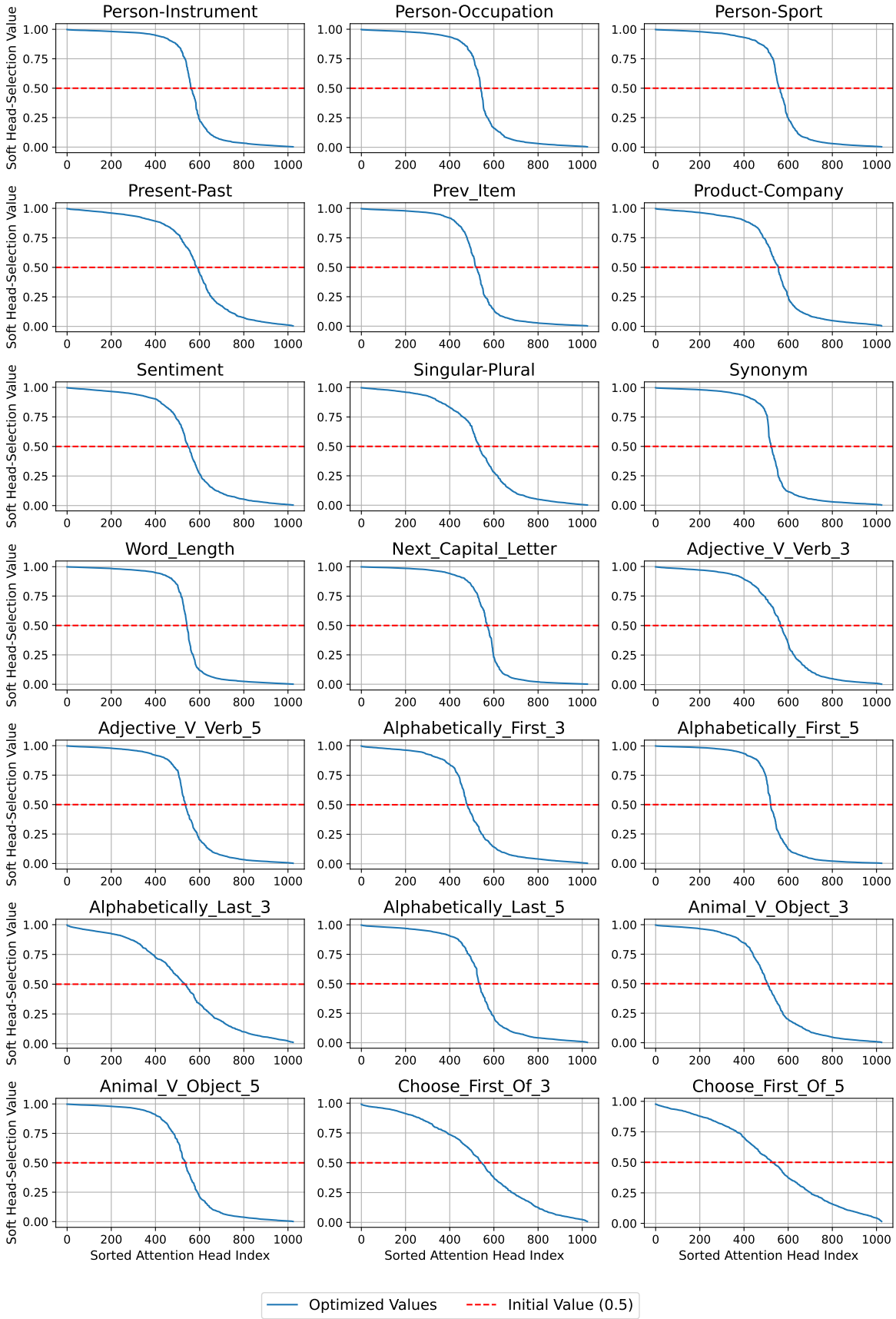

Figure 8: **Optimized values of the soft head-selection parameters for 57 FV tasks (Part 2 of 3).** This figure continues from Figure 7. Each plot shows the optimized values of the soft head-selection parameters for all 1024 attention heads in Llama-3.1-8B, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all selection parameters at the start of training. Plots for the remaining tasks are provided in Figure 9.

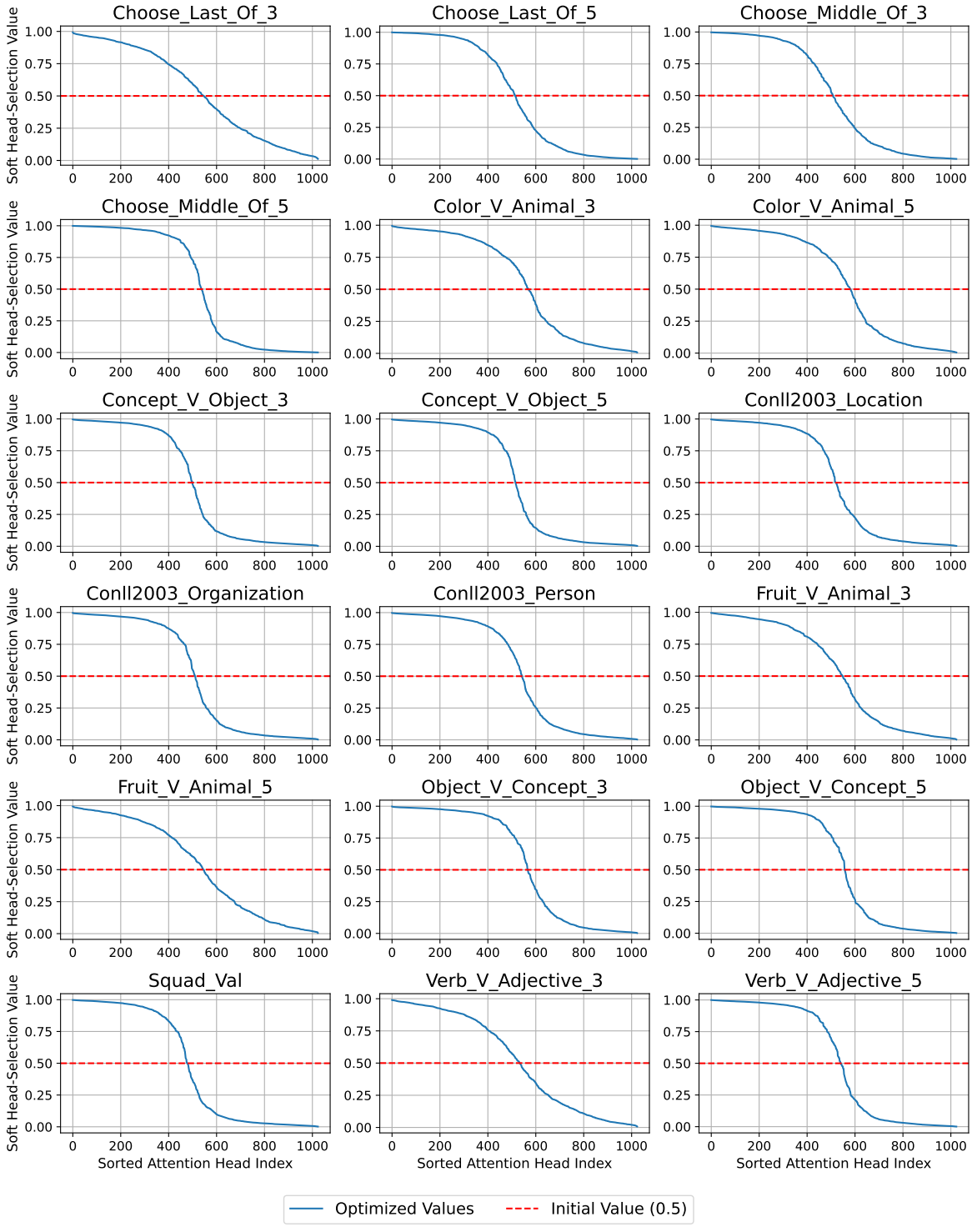

Figure 9: **Optimized values of the soft head-selection parameters for 57 FV tasks (Part 3 of 3).** This figure concludes the series from Figures 7-8. Each plot shows the optimized values of the soft head-selection parameters for all 1024 attention heads in Llama-3.1-8B, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all selection parameters at the start of training.

C.2 Optimized soft head-selection values for larger language models

Figures 10-12 present the optimized values of the soft head-selection parameters for the larger models Qwen3-32B, Mixtral-8x7B-v0.1, and Llama-3.1-70B across six selected tasks from the FV benchmark. The plots show consistent overall patterns, closely matching those observed for Llama-3.1-8B in Section 5.1.

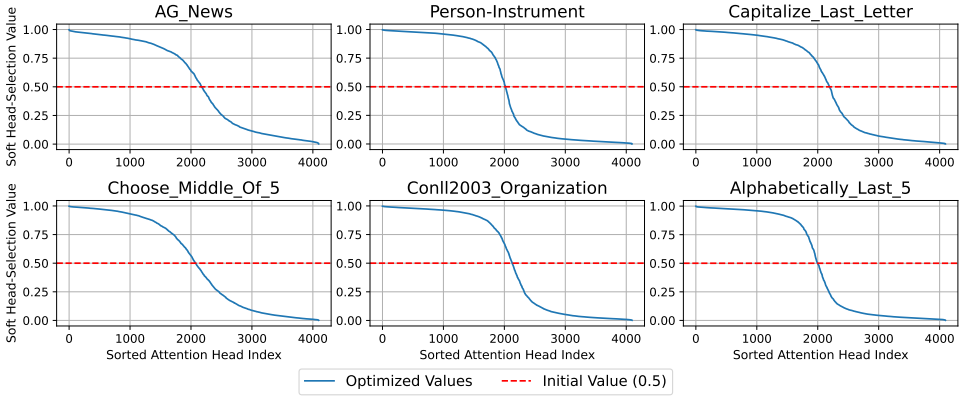

Figure 10: **Optimized values of the soft head-selection parameters for six FV tasks using Qwen3-32B.** Each plot shows the optimized values of the soft head-selection parameters for all 4096 attention heads in Qwen3-32B, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all selection parameters at the start of training.

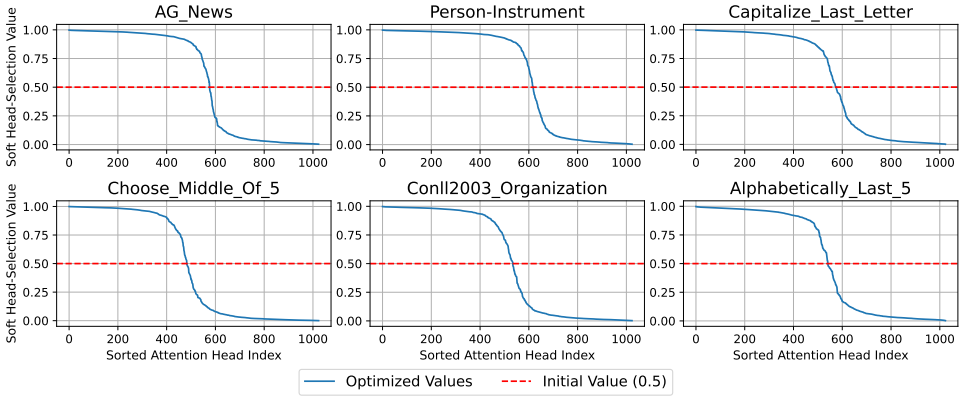

Figure 11: **Optimized values of the soft head-selection parameters for six FV tasks using Mixtral-8x7B-v0.1.** Each plot shows the optimized values of the soft head-selection parameters for all 1024 attention heads in Mixtral-8x7B-v0.1, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all selection parameters at the start of training.

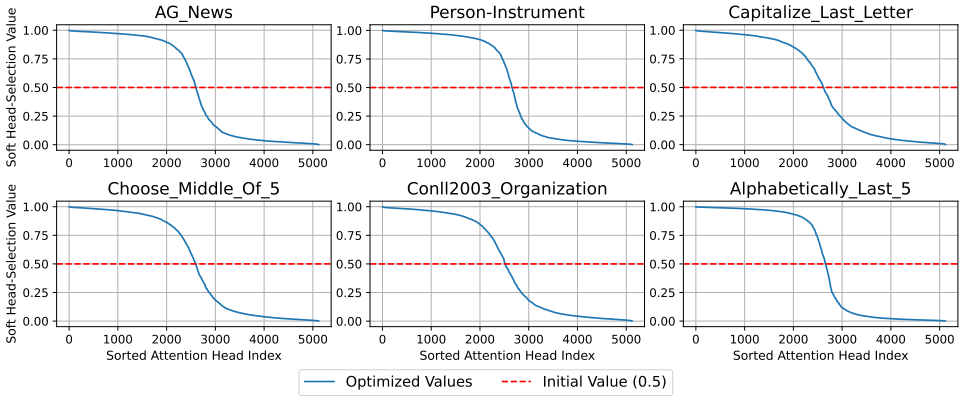

Figure 12: **Optimized values of the soft head-selection parameters for six FV tasks using Llama-3.1-70B.** Each plot shows the optimized values of the soft head-selection parameters for all 5120 attention heads in Llama-3.1-70B, sorted in descending order. Dashed lines indicate the initial value of 0.5 assigned to all selection parameters at the start of training.

D Additional results on inter-task activation patching analysis

D.1 Additional results on inter-task activation patching analysis for Llama-3.1-8B

In Table 24, we present the results of the inter-task activation patching analysis for 12 additional FV tasks using Llama-3.1-8B, following the procedure described in Section 5.2. As explained in Section 5.2, we fix both the task of zero-shot inference and the task embeddings used for activation patching, and vary only the head-selection parameters across tasks. Specifically, we choose a single *evaluation task*, from which both the zero-shot inference input and the task embeddings are derived, and apply head-selection parameters learned from different tasks in the FV benchmark. For each evaluation task, we report the top-3 and bottom-3 head-selection tasks ranked by post-patching performance. The overall trends are consistent with those reported in Table 3 of Section 5.2.

Evaluation Task	Task Description	Top-3 Head-Selection Tasks (Accuracy, %)	Bottom-3 Head-Selection Tasks (Accuracy, %)
Adjective_V_Verb_3	Select the adjective from a list of 3 words (1 adjective, 2 verbs)	Adjective_V_Verb_5 (99.0) Adjective_V_Verb_3 (97.1) Fruit_V_Animal_5 (89.0)	Verb_V_Adjective_3 (1.0) Verb_V_Adjective_5 (7.6) Squad_Val (15.7)
Verb_V_Adjective_3	Select the verb from a list of 3 words (1 verb, 2 adjectives)	Verb_V_Adjective_5 (99.5) Verb_V_Adjective_3 (97.1) Color_V_Animal_5 (80.5)	Adjective_V_Verb_3 (0.5) Adjective_V_Verb_5 (4.8) Synonym (14.8)
Alphabetically_First_3	Select the word that comes first in alphabetical order from a list of 3 words	Alphabetically_First_5 (86.7) Alphabetically_First_3 (52.9) Next_Item (45.2)	Person_Occupation (23.3) Alphabetically_Last_5 (23.3) Alphabetically_Last_3 (25.2)
Alphabetically_Last_3	Select the word that comes last in alphabetical order from a list of 3 words	Alphabetically_Last_5 (50.0) Alphabetically_Last_3 (46.2) Concept_V_Object_5 (36.7)	Alphabetically_First_5 (1.0) Alphabetically_First_3 (15.7) Person_Occupation (22.4)
Concept_V_Object_3	Select the concept from a list of 3 words (1 abstract concept, 2 concrete entities)	Concept_V_Object_3 (99.0) Concept_V_Object_5 (98.1) Animal_V_Object_5 (71.9)	Object_V_Concept_5 (0.5) Object_V_Concept_3 (2.9) Park_Country (13.3)
Concept_V_Object_5	Select the concept from a list of 5 words (1 abstract concept, 4 concrete entities)	Concept_V_Object_3 (95.2) Concept_V_Object_5 (91.9) Animal_V_Object_5 (73.3)	Object_V_Concept_5 (2.4) Park_Country (6.2) Object_V_Concept_3 (6.2)
Object_V_Concept_3	Select the concrete entity from a list of 3 words (1 concrete entity, 2 abstract concepts)	Object_V_Concept_3 (100.0) Object_V_Concept_5 (99.0) Color_V_Animal_3 (95.7)	Concept_V_Object_3 (5.7) Concept_V_Object_5 (8.6) Squad_Val (15.2)
Object_V_Concept_5	Select the concrete entity from a list of 5 words (1 concrete entity, 4 abstract concepts)	Object_V_Concept_5 (98.1) Object_V_Concept_3 (96.2) Fruit_V_Animal_3 (92.4)	Concept_V_Object_3 (3.3) Concept_V_Object_5 (3.8) Squad_Val (10.5)
Capitalize_First_Letter	Generate the first letter of a given word in captial form	Capitalize (100.0) Capitalize_First_Letter (100.0) Lowercase_First_Letter (100.0)	Conll2003_Organization (0.0) Choose_Middle_Of_3 (0.0) Next_Item (0.0)
Lowercase_First_Letter	Generate the first letter of a given word in lowercase	Capitalize (100.0) Capitalize_First_Letter (100.0) English_French (100.0)	Conll2003_Organization (0.0) Conll2003_Location (0.0) Person_Occupation (0.0)
Capitalize_Last_Letter	Generate the last letter of a given word in captial form	Capitalize_Last_Letter (87.7) Lowercase_Last_Letter (86.0) Country_Currency (42.7)	Choose_First_Of_5 (0.0) Choose_Middle_Of_3 (0.0) Conll2003_Organization (0.6)
Lowercase_Last_Letter	Generate the last letter of a given word in lowercase	Lowercase_Last_Letter (94.7) Capitalize_Last_Letter (93.6) National_Parks (50.3)	Conll2003_Organization (0.0) Alphabetically_Last_3 (0.0) Prev_Item (0.0)

Table 24: **Inter-task activation patching analysis for 12 additional FV tasks using Llama-3.1-8B.** For each evaluation task, we report performance after patching high-α attention heads, where task embeddings are fixed to the evaluation task and head-selection parameters are derived from different head-selection tasks. Among the 57 FV tasks, we report the top-3 and bottom-3 head-selection tasks ranked by post-patching accuracy.

D.2 Results on inter-task activation patching analysis for larger language models

Tables 25-27 present the results of the inter-task activation patching analysis for larger models: Qwen3-32B, Mixtral-8x7B-v0.1, and Llama-3.1-70B, across 19 FV tasks. The overall trends are consistent with those reported for Llama-3.1-8B in Table 3 of Section 5.2 and Table 24.

Evaluation Task	Task Description	Top-3 Head-Selection Tasks (Accuracy, %)	Bottom-3 Head-Selection Tasks (Accuracy, %)
Adjective_V_Verb_3	Select the adjective from a list of 3 words (1 adjective, 2 verbs)	Adjective_V_Verb_3 (99.5) Adjective_V_Verb_5 (99.0) Animal_V_Object_3 (85.7)	Person_Occupation (3.3) Sentiment (7.6) Verb_V_Adjective_3 (9.5)
Adjective_V_Verb_5	Select the only adjective from a list of 5 words (1 adjective, 4 verbs)	Adjective_V_Verb_5 (98.1) Adjective_V_Verb_3 (97.1) Animal_V_Object_3 (86.2)	Person_Occupation (5.2) Park_Country (10.0) Sentiment (11.0)
Verb_V_Adjective_3	Select the verb from a list of 3 words (1 verb, 2 adjectives)	Verb_V_Adjective_3 (99.5) Verb_V_Adjective_5 (95.7) Singular_Plural (81.9)	Ag_News (1.4) Person_Occupation (2.9) Adjective_V_Verb_3 (5.2)
Verb_V_Adjective_5	Select the only verb from a list of 5 words (1 verb, 4 adjectives)	Verb_V_Adjective_5 (99.5) Verb_V_Adjective_3 (99.0) Concept_V_Object_5 (91.4)	Person_Occupation (3.8) Ag_News (5.7) Sentiment (6.2)
Alphabetically_First_3	Select the word that comes first in alphabetical order from a list of 3 words	Alphabetically_First_3 (96.7) Alphabetically_First_5 (94.8) Animal_V_Object_3 (41.0)	Person_Occupation (0.5) Sentiment (2.4) Park_Country (6.2)
Alphabetically_First_5	Choose the word that comes first in alphabetical order from a list of 5 words	Alphabetically_First_5 (88.1) Alphabetically_First_3 (85.2) Antonym (27.6)	Sentiment (1.9) Person_Occupation (6.2) Next_Capital_Letter (7.1)
Alphabetically_Last_3	Select the word that comes last in alphabetical order from a list of 3 words	Alphabetically_Last_5 (44.8) Alphabetically_Last_3 (42.9) Color_V_Animal_3 (38.6)	Alphabetically_First_5 (1.0) Alphabetically_First_3 (1.4) Person_Occupation (3.3)
Alphabetically_Last_5	Choose the word that comes last in alphabetical order from a list of 5 words	Alphabetically_Last_5 (39.5) Alphabetically_Last_3 (29.0) Choose_Middle_Of_5 (25.2)	Alphabetically_First_3 (0.0) Alphabetically_First_5 (0.5) Sentiment (5.7)
Concept_V_Object_3	Select the concept from a list of 3 words (1 abstract concept, 2 concrete entities)	Concept_V_Object_3 (99.0) Concept_V_Object_5 (97.1) Fruit_V_Animal_3 (82.9)	Sentiment (3.8) Ag_News (4.8) Person_Occupation (7.6)
Concept_V_Object_5	Select the concept from a list of 5 words (1 abstract concept, 4 concrete entities)	Concept_V_Object_3 (97.1) Concept_V_Object_5 (97.1) Fruit_V_Animal_3 (82.9)	Sentiment (7.6) Person_Occupation (11.4) Ag_News (11.9)
Object_V_Concept_3	Select the concrete entity from a list of 3 words (1 concrete entity, 2 abstract concepts)	Object_V_Concept_3 (100.0) Object_V_Concept_5 (98.6) Country_Capital (90.5)	Person_Occupation (4.3) Park_Country (4.8) Ag_News (7.6)
Object_V_Concept_5	Select the concrete entity from a list of 5 words (1 concrete entity, 4 abstract concepts)	Object_V_Concept_5 (98.1) Object_V_Concept_3 (97.6) Adjective_V_Verb_5 (86.7)	Person_Occupation (1.9) Park_Country (5.7) Concept_V_Object_3 (8.6)
English_French	Translate the given English word into French	English_German (77.4) English_French (77.3) English_Spanish (75.5)	Alphabetically_Last_5 (1.1) Commonsense_Qa (2.7) Person_Instrument (4.9)
English_German	Translate the given English word into German	English_French (67.8) English_German (65.0) English_Spanish (63.3)	Adjective_V_Verb_3 (0.9) Person_Instrument (2.1) Alphabetically_Last_5 (3.0)
English_Spanish	Translate the given English word into Spanish	English_French (83.2) English_German (82.3) English_Spanish (80.2)	Alphabetically_Last_5 (4.5) Person_Instrument (8.0) Commonsense_Qa (15.1)
Capitalize_First_Letter	Generate the first letter of a given word in captial form	Capitalize (100.0) Capitalize_First_Letter (100.0) Country_Capital (100.0)	Person_Occupation (3.5) Prev_Item (4.1) Next_Item (7.6)
Lowercase_First_Letter	Generate the first letter of a given word in lowercase	Capitalize_First_Letter (100.0) English_German (100.0) Lowercase_First_Letter (100.0)	Prev_Item (0.0) Conll2003_Location (0.0) Next_Item (0.0)
Capitalize_Last_Letter	Generate the last letter of a given word in captial form	Capitalize_Last_Letter (90.1) Lowercase_Last_Letter (81.3) Next_Item (49.7)	Choose_Last_Of_3 (0.6) English_French (1.2) Prev_Item (1.2)
Lowercase_Last_Letter	Generate the last letter of a given word in lowercase	Lowercase_Last_Letter (95.9) Verb_V_Adjective_5 (82.5) Capitalize_Last_Letter (81.9)	Prev_Item (0.0) Conll2003_Location (0.0) Next_Item (0.0)

Table 25: **Inter-task activation patching analysis for 19 FV tasks using Qwen3-32B.** For each evaluation task, we report performance after patching high-α attention heads, where task embeddings are fixed to the evaluation task and head-selection parameters are derived from different head-selection tasks. Among the 57 FV tasks, we report the top-3 and bottom-3 head-selection tasks ranked by post-patching accuracy.

Evaluation Task	Task Description	Top-3 Head-Selection Tasks (Accuracy, %)	Bottom-3 Head-Selection Tasks (Accuracy, %)
Adjective_V_Verb_3	Select the adjective from a list of 3 words (1 adjective, 2 verbs)	Adjective_V_Verb_3 (98.6) Adjective_V_Verb_5 (98.1) Conll2003_Organization (84.3)	Verb_V_Adjective_3 (2.9) Verb_V_Adjective_5 (6.7) Product_Company (19.0)
Adjective_V_Verb_5	Select the only adjective from a list of 5 words (1 adjective, 4 verbs)	Adjective_V_Verb_3 (97.6) Adjective_V_Verb_5 (97.6) Animal_V_Object_5 (71.4)	Verb_V_Adjective_3 (1.4) Verb_V_Adjective_5 (5.7) Product_Company (19.5)
Verb_V_Adjective_3	Select the verb from a list of 3 words (1 verb, 2 adjectives)	Verb_V_Adjective_5 (96.7) Verb_V_Adjective_3 (96.2) Fruit_V_Animal_3 (65.2)	Adjective_V_Verb_3 (0.0) Adjective_V_Verb_5 (0.0) Ag_News (12.9)
Verb_V_Adjective_5	Select the only verb from a list of 5 words (1 verb, 4 adjectives)	Verb_V_Adjective_5 (98.1) Verb_V_Adjective_3 (96.2) Animal_V_Object_5 (67.1)	Adjective_V_Verb_3 (0.0) Adjective_V_Verb_5 (1.0) English_French (6.2)
Alphabetically_First_3	Select the word that comes first in alphabetical order from a list of 3 words	Alphabetically_First_5 (76.7) Alphabetically_First_3 (47.1) Lowercase_First_Letter (38.6)	Alphabetically_Last_5 (16.7) Ag_News (23.3) Alphabetically_Last_3 (23.8)
Alphabetically_First_5	Choose the word that comes first in alphabetical order from a list of 5 words	Alphabetically_First_5 (88.6) Alphabetically_First_3 (26.7) Adjective_V_Verb_5 (25.2)	Alphabetically_Last_5 (8.6) Alphabetically_Last_3 (10.5) Animal_V_Object_3 (14.3)
Alphabetically_Last_3	Select the word that comes last in alphabetical order from a list of 3 words	Alphabetically_Last_5 (51.9) Alphabetically_Last_3 (50.5) Squad_Val (41.0)	Alphabetically_First_5 (9.0) Alphabetically_First_3 (25.7) Person_Instrument (26.2)
Alphabetically_Last_5	Choose the word that comes last in alphabetical order from a list of 5 words	Alphabetically_Last_5 (39.0) Alphabetically_Last_3 (33.8) Person_Sport (26.7)	Alphabetically_First_5 (0.5) Alphabetically_First_3 (11.9) National_Parks (12.9)
Concept_V_Object_3	Select the concept from a list of 3 words (1 abstract concept, 2 concrete entities)	Concept_V_Object_3 (99.0) Concept_V_Object_5 (99.0) Fruit_V_Animal_3 (62.4)	Object_V_Concept_3 (2.9) Object_V_Concept_5 (8.6) Person_Instrument (11.0)
Concept_V_Object_5	Select the concept from a list of 5 words (1 abstract concept, 4 concrete entities)	Concept_V_Object_5 (96.2) Concept_V_Object_3 (93.8) Animal_V_Object_5 (71.0)	Object_V_Concept_5 (4.8) Object_V_Concept_3 (6.2) Person_Instrument (10.0)
Object_V_Concept_3	Select the concrete entity from a list of 3 words (1 concrete entity, 2 abstract concepts)	Object_V_Concept_3 (99.0) Object_V_Concept_5 (97.1) Fruit_V_Animal_5 (76.7)	Concept_V_Object_5 (2.4) Concept_V_Object_3 (2.9) Adjective_V_Verb_3 (21.4)
Object_V_Concept_5	Select the concrete entity from a list of 5 words (1 concrete entity, 4 abstract concepts)	Object_V_Concept_5 (96.7) Object_V_Concept_3 (96.2) Fruit_V_Animal_5 (81.0)	Concept_V_Object_3 (1.0) Concept_V_Object_5 (1.9) Adjective_V_Verb_3 (10.0)
English_French	Translate the given English word into French	English_Spanish (82.3) English_French (82.2) Capitalize (80.6)	Conll2003_Organization (2.1) Alphabetically_First_5 (3.0) Object_V_Concept_3 (3.3)
English_German	Translate the given English word into German	English_Spanish (78.4) English_French (75.6) English_German (75.1)	Conll2003_Organization (3.4) Alphabetically_First_5 (6.2) Animal_V_Object_5 (8.3)
English_Spanish	Translate the given English word into Spanish	English_Spanish (87.1) English_German (85.3) English_French (84.0)	Conll2003_Organization (4.9) Alphabetically_First_5 (25.2) Animal_V_Object_5 (28.6)
Capitalize_First_Letter	Generate the first letter of a given word in captial form	Capitalize_First_Letter (100.0) Lowercase_First_Letter (100.0) English_French (99.4)	Ag_News (0.0) Park_Country (0.0) Landmark_Country (0.6)
Lowercase_First_Letter	Generate the first letter of a given word in lowercase	Capitalize (100.0) Capitalize_First_Letter (100.0) English_French (100.0)	Ag_News (0.0) Park_Country (0.6) Conll2003_Organization (1.2)
Capitalize_Last_Letter	Generate the last letter of a given word in captial form	Capitalize_Last_Letter (90.1) Lowercase_Last_Letter (81.3) Present_Past (58.5)	Choose_First_Of_3 (0.0) Choose_First_Of_5 (0.0) Conll2003_Organization (0.6)
Lowercase_Last_Letter	Generate the last letter of a given word in lowercase	Lowercase_Last_Letter (93.0) Capitalize_Last_Letter (90.6) Present_Past (70.2)	Choose_First_Of_3 (0.0) Choose_Middle_Of_5 (0.0) Prev_Item (0.0)

Table 26: **Inter-task activation patching analysis for 19 FV tasks using Mixtral-8x7B-v0.1.** For each evaluation task, we report performance after patching high-α attention heads, where task embeddings are fixed to the evaluation task and head-selection parameters are derived from different head-selection tasks. Among the 57 FV tasks, we report the top-3 and bottom-3 head-selection tasks ranked by post-patching accuracy.

Evaluation Task	Task Description	Top-3 Head-Selection Tasks (Accuracy, %)	Bottom-3 Head-Selection Tasks (Accuracy, %)
Adjective_V_Verb_3	Select the adjective from a list of 3 words (1 adjective, 2 verbs)	Adjective_V_Verb_5 (99.5) Adjective_V_Verb_3 (99.0) Object_V_Concept_3 (88.1)	Verb_V_Adjective_5 (3.8) Conll2003_Organization (10.0) Verb_V_Adjective_3 (10.5)
Adjective_V_Verb_5	Select the only adjective from a list of 5 words (1 adjective, 4 verbs)	Adjective_V_Verb_5 (100.0) Adjective_V_Verb_3 (95.2) Color_V_Animal_3 (79.0)	Choose_First_Of_3 (4.8) Capitalize_Second_Letter (8.1) Verb_V_Adjective_5 (10.0)
Verb_V_Adjective_3	Select the verb from a list of 3 words (1 verb, 2 adjectives)	Verb_V_Adjective_3 (100.0) Verb_V_Adjective_5 (99.5) Concept_V_Object_5 (89.0)	Antonym (7.1) Adjective_V_Verb_3 (8.1) Choose_First_Of_3 (11.9)
Verb_V_Adjective_5	Select the only verb from a list of 5 words (1 verb, 4 adjectives)	Verb_V_Adjective_5 (99.5) Verb_V_Adjective_3 (99.0) Concept_V_Object_5 (92.9)	Antonym (2.4) Choose_First_Of_3 (2.9) Capitalize_Second_Letter (3.8)
Alphabetically_First_3	Select the word that comes first in alphabetical order from a list of 3 words	Alphabetically_First_3 (98.1) Alphabetically_First_5 (95.7) Commonsense_Qa (43.3)	Conll2003_Organization (1.9) Person_Instrument (6.7) Choose_First_Of_3 (9.0)
Alphabetically_First_5	Choose the word that comes first in alphabetical order from a list of 5 words	Alphabetically_First_5 (97.6) Alphabetically_First_3 (96.7) Adjective_V_Verb_3 (31.0)	Alphabetically_Last_5 (1.0) Conll2003_Organization (4.8) Choose_First_Of_3 (6.2)
Alphabetically_Last_3	Select the word that comes last in alphabetical order from a list of 3 words	Alphabetically_Last_5 (85.2) Alphabetically_Last_3 (61.9) Verb_V_Adjective_5 (45.7)	Conll2003_Organization (0.5) Alphabetically_First_3 (0.5) Alphabetically_First_5 (0.5)
Alphabetically_Last_5	Choose the word that comes last in alphabetical order from a list of 5 words	Alphabetically_Last_5 (85.2) Alphabetically_Last_3 (50.0) Country_Currency (33.8)	Alphabetically_First_3 (0.0) Alphabetically_First_5 (0.0) Choose_First_Of_3 (3.3)
Concept_V_Object_3	Select the concept from a list of 3 words (1 abstract concept, 2 concrete entities)	Concept_V_Object_5 (99.5) Concept_V_Object_3 (98.6) Animal_V_Object_5 (83.3)	Conll2003_Organization (0.0) Person_Instrument (10.5) Choose_First_Of_3 (11.0)
Concept_V_Object_5	Select the concept from a list of 5 words (1 abstract concept, 4 concrete entities)	Concept_V_Object_3 (98.6) Concept_V_Object_5 (97.6) Animal_V_Object_5 (82.4)	Capitalize_Second_Letter (4.3) Choose_First_Of_3 (9.0) Word_Length (10.0)
Object_V_Concept_3	Select the concrete entity from a list of 3 words (1 concrete entity, 2 abstract concepts)	Object_V_Concept_3 (99.0) Object_V_Concept_5 (98.6) Conll2003_Location (86.7)	Choose_First_Of_3 (2.9) Conll2003_Organization (3.8) Person_Instrument (9.0)
Object_V_Concept_5	Select the concrete entity from a list of 5 words (1 concrete entity, 4 abstract concepts)	Object_V_Concept_5 (98.6) Object_V_Concept_3 (93.3) Animal_V_Object_5 (90.0)	Choose_First_Of_3 (2.4) Antonym (8.1) Person_Instrument (14.8)
English_French	Translate the given English word into French	English_German (86.2) English_French (85.6) English_Spanish (85.4)	Conll2003_Organization (0.0) Animal_V_Object_5 (23.6) Choose_First_Of_3 (25.6)
English_German	Translate the given English word into German	English_Spanish (80.6) English_French (79.8) English_German (79.2)	Conll2003_Organization (0.0) Choose_First_Of_5 (13.5) Choose_First_Of_3 (17.2)
English_Spanish	Translate the given English word into Spanish	English_Spanish (88.9) English_French (88.7) English_German (88.3)	Conll2003_Organization (0.0) Choose_First_Of_5 (16.1) Choose_First_Of_3 (25.5)
Capitalize_First_Letter	Generate the first letter of a given word in captial form	Capitalize_First_Letter (100.0) Lowercase_First_Letter (100.0) Capitalize (99.4)	Choose_Middle_Of_5 (0.0) Prev_Item (0.0) Person_Instrument (0.0)
Lowercase_First_Letter	Generate the first letter of a given word in lowercase	Capitalize_First_Letter (100.0) Lowercase_First_Letter (100.0) Word_Length (100.0)	Person_Instrument (0.0) Prev_Item (0.0) Next_Item (0.0)
Capitalize_Last_Letter	Generate the last letter of a given word in captial form	Capitalize_Last_Letter (97.1) Lowercase_Last_Letter (94.7) Country_Currency (52.6)	Choose_Middle_Of_5 (0.0) Conll2003_Organization (0.0) Antonym (0.0)
Lowercase_Last_Letter	Generate the last letter of a given word in lowercase	Capitalize_Last_Letter (100.0) Lowercase_Last_Letter (97.7) Present_Past (60.8)	Choose_Middle_Of_5 (0.0) Color_V_Animal_5 (0.0) Antonym (0.0)

Table 27: **Inter-task activation patching analysis for 19 FV tasks using Llama-3.1-70B.** For each evaluation task, we report performance after patching high-α attention heads, where task embeddings are fixed to the evaluation task and head-selection parameters are derived from different head-selection tasks. Among the 57 FV tasks, we report the top-3 and bottom-3 head-selection tasks ranked by post-patching accuracy.

E Extended results on head-selection training dynamics

E.1 Extended results for all 57 FV tasks

In this section, we present extended results corresponding to Figure 5 in Section 6, showing the training dynamics in terms of validation loss and test accuracy for all 57 tasks from the FV benchmark using Llama-3.1-8B. The full set of results is provided in Figures 13-15.

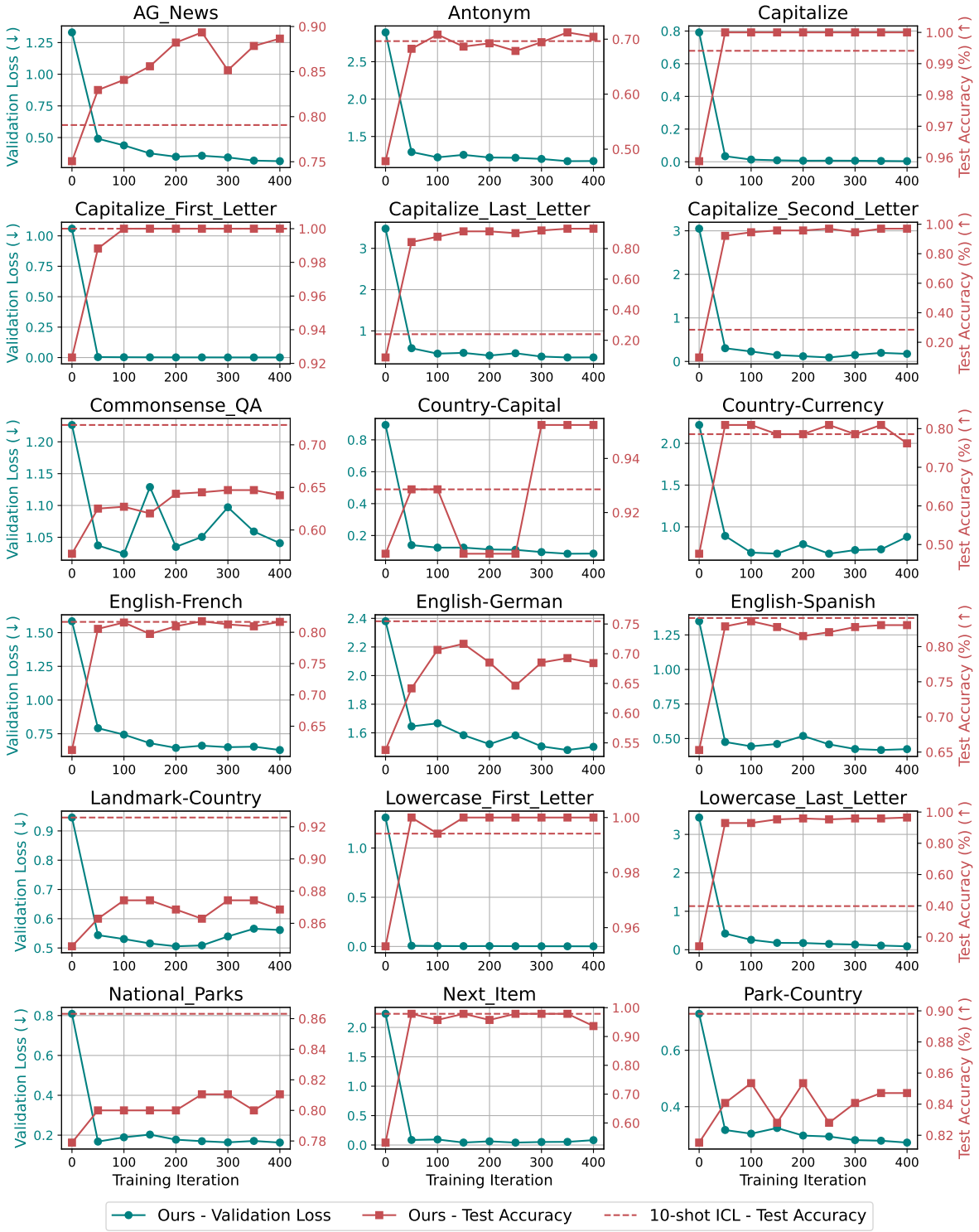

Figure 13: **Training dynamics of soft head-selection parameters for 57 FV tasks (Part 1 of 3).** Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference. The results are based on Llama-3.1-8B. Plots for the remaining tasks are provided in Figure 14-15.

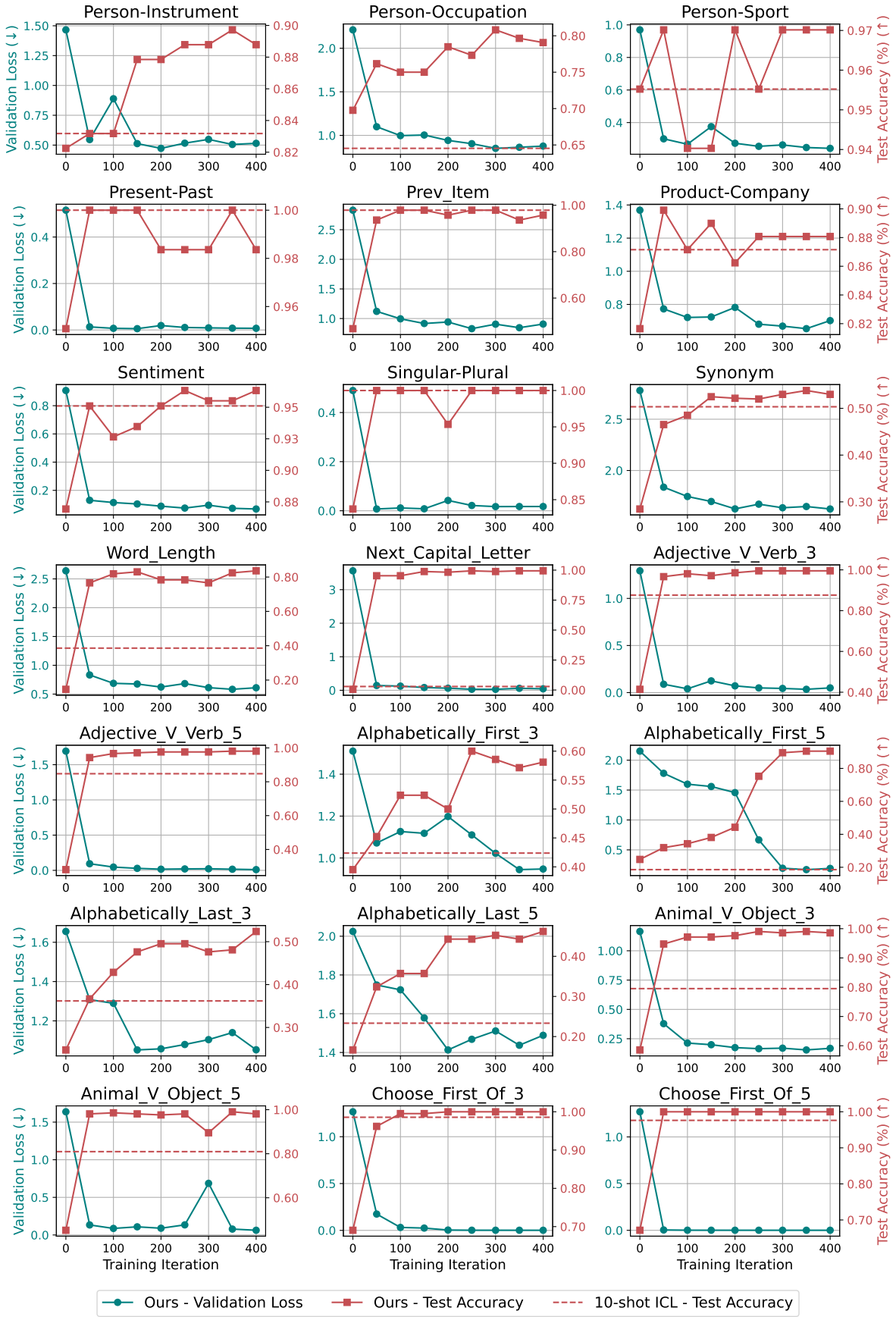

Figure 14: **Training dynamics of soft head-selection parameters for 57 FV tasks (Part 2 of 3).** This figure continues from Figure 13. Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference. The results are based on Llama-3.1-8B. Plots for the remaining tasks are provided in Figure 15.

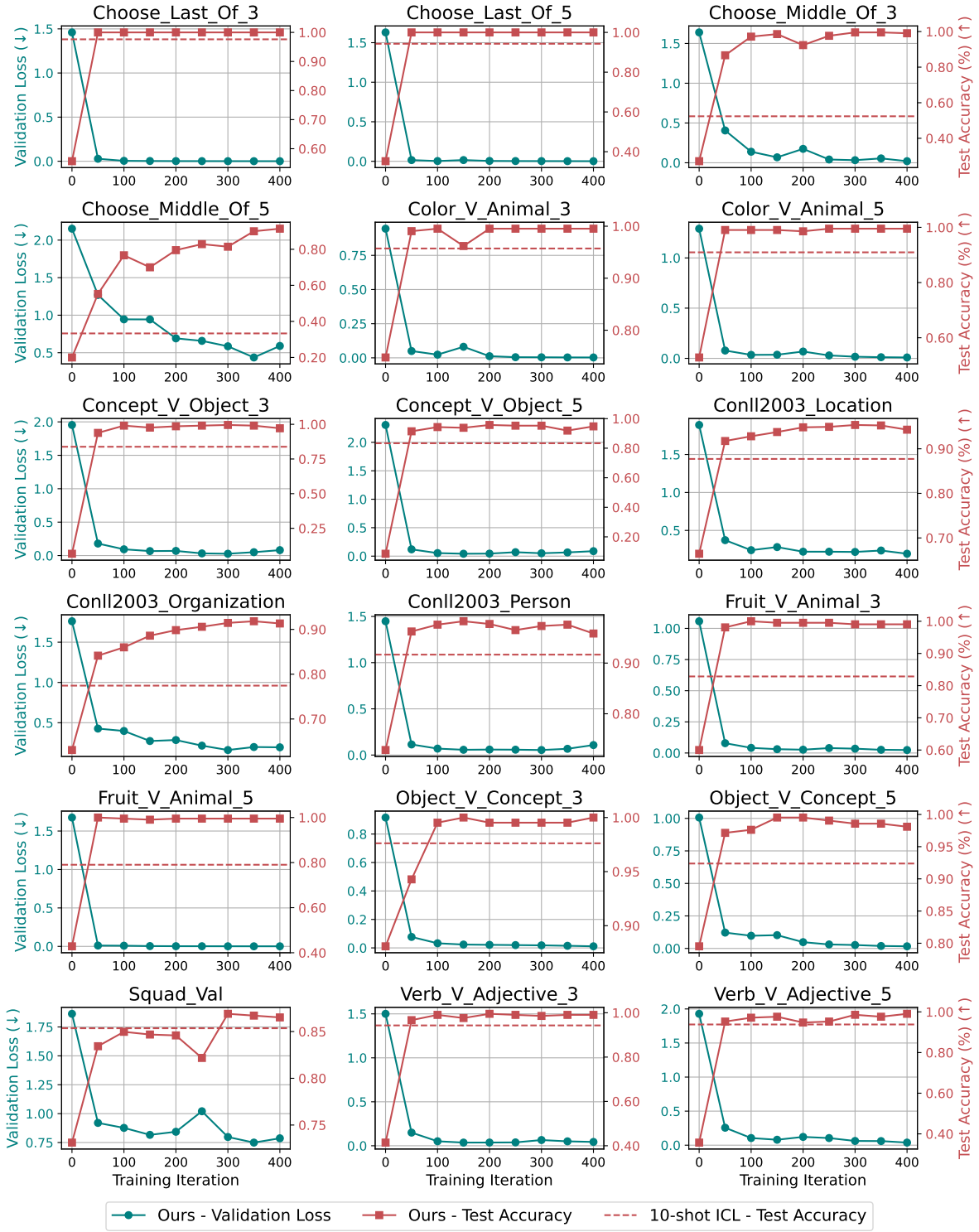

Figure 15: **Training dynamics of soft head-selection parameters for 57 FV tasks (Part 3 of 3).** This figure concludes the series from Figures 13-14. Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference. The results are based on Llama-3.1-8B.

E.2 Results for larger language models

Figures 16-18 present training dynamics for the larger models Qwen3-32B, Mixtral-8x7B-v0.1, and Llama-3.1-70B across six selected tasks from the FV benchmark. The resulting plots exhibit consistent overall trends, similar to those observed with Llama-3.1-8B in Section 6.

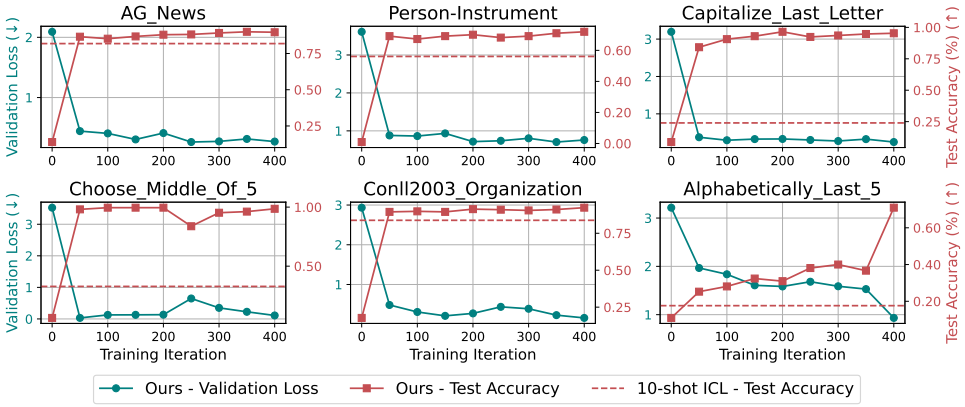

Figure 16: **Training dynamics of soft head-selection parameters for six FV tasks using Qwen3-32B.** Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference.

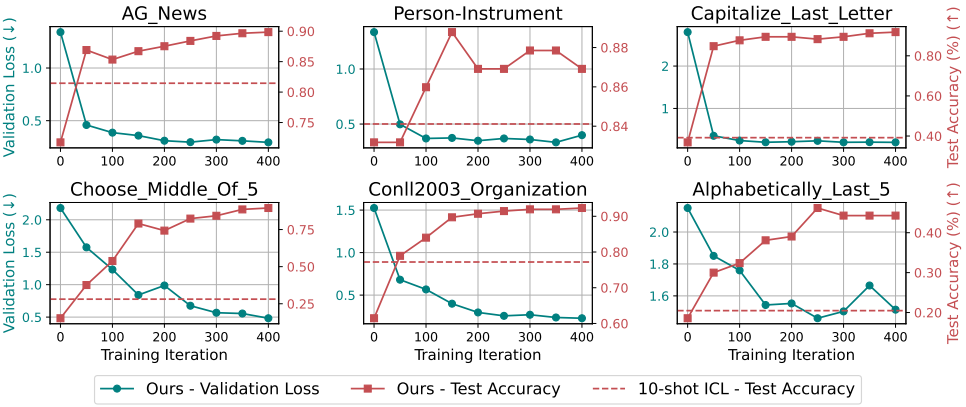

Figure 17: **Training dynamics of soft head-selection parameters for six FV tasks using Mixtral-8x7B-v0.1.** Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference.

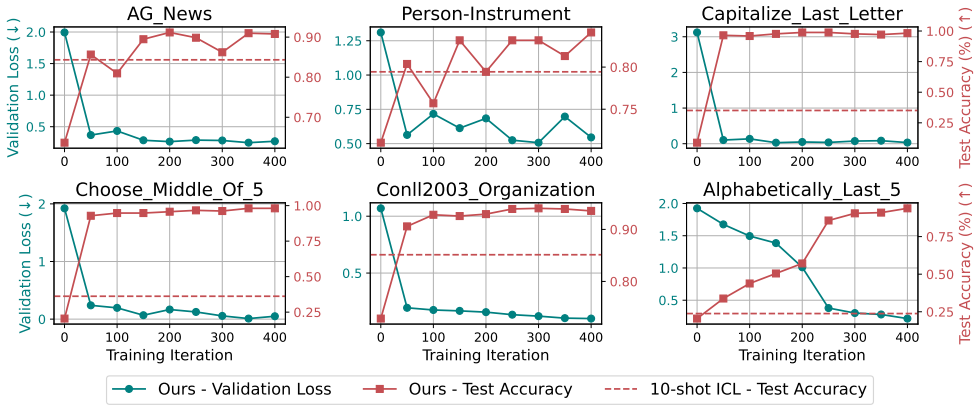

Figure 18: **Training dynamics of soft head-selection parameters for six FV tasks using Llama-3.1-70B.** Validation loss (left y-axis) and test accuracy (right y-axis) are plotted over 400 training iterations. Dashed lines indicate the 10-shot ICL accuracies for reference.

F Use of LLMs in this work

We used chat-based LLMs solely for sentence-level editing to check grammar and improve clarity during paper writing. All edits were reviewed and verified by the authors. No scientific ideas, methods, analyses, or results were produced by LLMs; all conceptual contributions and experimental work are solely by the authors.