Parenting: Optimizing Knowledge Selection of Retrieval-Augmented Language Models with Parameter Decoupling and Tailored Tuning

Yongxin Xu[1,2*], **Ruizhe Zhang**[1,2*], **Xinke Jiang**[1,2*], **Yujie Feng**[7], **Yuzhen Xiao**[1,2], **Xinyu Ma**[1,2], **Runchuan Zhu**[1,2], **Xu Chu**[1,2,4,5], **Junfeng Zhao**[1,2,6†], **Yasha Wang**[2,3,4†]

[1] School of Computer Science and School of Software & Microelectronics, Peking University
[2] Key Laboratory of High Confidence Software Technologies, Ministry of Education
[3] National Engineering Research Center For Software Engineering, Peking University
[4] Peking University Information Technology Institute (Tianjin Binhai)
[5] Center on Frontiers of Computing Studies, Peking University
[6] Big Data Technology Research Center, Nanhu Laboratory
[7] Department of Computing, The Hong Kong Polytechnic University

{xuyx, nostradamus, xinkejiang}@stu.pku.edu.cn, {zhaojf, wangyasha}@pku.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) offers an effective solution to the issues faced by Large Language Models (LLMs) in hallucination generation and knowledge obsolescence by incorporating externally retrieved knowledge. However, existing methods lack effective control mechanisms for integrating internal and external knowledge. Inspired by human cognitive processes, we propose Parenting, a novel framework that decouples, identifies, and purposefully optimizes parameter subspaces related to adherence and robustness. Specifically, Parenting utilizes a key parameter mining method that combines forward and backward propagation signals to localize subspaces representing different capabilities. Then, Parenting employs a type-tailored tuning strategy, applying specific and appropriate optimizations to different subspaces, aiming to achieve a balanced enhancement of both adherence and robustness. Extensive experiments on various datasets and models validate the effectiveness and generalizability of our method. Our code is available at https://github.com/Nostradamus4869/Parenting.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities, achieving state-of-the-art performance on various tasks (Brown, 2020; Chowdhery et al., 2023; Bubeck et al., 2023; Xu et al., 2025; Ma et al., 2025; Feng et al., 2023, 2025). Despite their success, they still face notable challenges, particularly in generating hallucinations and dealing with outdated knowledge (Gao et al., 2023). Retrieval-Augmented Generation (RAG) has emerged as a promising approach to mitigate these issues (Peng et al., 2023; Ren et al., 2023; Lewis et al., 2020; Jiang et al., 2024a). By integrating external information relevant to a specific query, RAG enhances the generative process with supplementary non-parametric knowledge. However, typical RAG frameworks lack effective control mechanisms for managing internal and external knowledge (Li et al., 2023a), which presents two main challenges: First, the conflicts between external knowledge and the internal memory of LLMs can, in certain cases, prevent the model from effectively following external evidence to produce accurate responses (Wu et al., 2024b; Li et al., 2023a). Secondly, the inherent imperfections of retrieval mechanisms mean that the retrieved contexts might include irrelevant noises (Creswell et al.), which can mislead the LLMs, leading to degraded performance (Fang et al., 2024; Xu et al., 2024a).

To address the above issues, some approaches optimize knowledge selection in the RAG process through carefully designed prompts (Zhou et al., 2023). However, such methods do not fundamentally improve LLMs' ability to integrate external knowledge, leading to suboptimal outcomes in certain situations. Other methods focus on altering the behavioral patterns of LLMs through training techniques such as instruction tuning (Li et al., 2023a; Fang et al., 2024; Xu et al., 2024a). Nevertheless, they lack differentiation in supervisory signals, leading to significant learning variance. On the one hand, an excessive emphasis on adhering to context can lead models to pay attention to noisy information. On the other hand, prioritizing resistance to noise might cause models to overlook critical evidence present within the context (Wu et al., 2024a). In summary, how to establish an effective control

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11643–11662

mechanism for managing both internal and external knowledge within RAG remains an unresolved problem.

The human brain is comprised of multiple functional regions, each tasked with distinct cognitive and physiological roles, with some regions being integral to handling complex tasks (Hawrylycz et al., 2012). For example, the *mirror neuron system* enhances learning through observation and imitation (Rizzolatti and Craighero, 2004), while the *hippocampus* plays a crucial role in retrieving knowledge from stored memories to solve problems (Olton et al., 1979). Inspired by this, a natural question arises:

Can we pinpoint specific subspaces in LLMs' parameter space linked to adhering to contextual evidence (adherence) and resisting contextual noise (robustness), akin to how the human brain localizes different cognitive functions?

To answer this question, inspired by Feng et al. (2024b) on identifying key parameters and integrating knowledge from different tasks to alleviate catastrophic forgetting in continual learning, we introduce a new perspective, as illustrated in Figure 1, that aims to decouple and identify parameter subspaces related to **adherence** and **robustness**. By designing tailored tuning strategies for these different subspaces, we seek to achieve better control over model behavior. Although seemingly straightforward, implementing this intuition faces these challenges: **(C1)** How to precisely quantify the correlation between parameters and two capabilities? **(C2)** How to optimize the entangled subspace that is difficult to decouple?

To address these challenges, we propose a framework named Parenting, which leverages **Par**ameter D**e**coupli**n**g and **T**ailored Tun**ing** to optimize the internal and external knowledge control mechanisms of Retrieval-Augmented Language Models (RALMs), while boosting both adherence and robustness without compromise. Specifically, addressing challenge **C1**, Parenting utilizes a key parameter mining method. By constructing a specialized probing dataset, Parenting measures each parameter's contribution to specific behaviors by combining the activation levels of pre-trained parameters under various inputs with gradient trajectories observed during the backpropagation process. Subsequently, through interactive analysis of the parameter importance distribution for different behaviors, we identify two mutually exclusive parameter subspaces: the **adherence subspace**

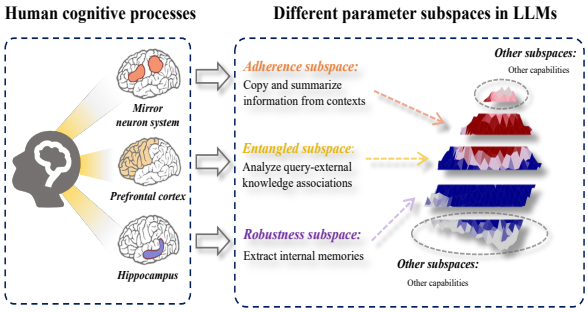

Figure 1: Formal exposition of our critical ideas. Inspired by human cognitive processes, we aim to decouple and localize parameter subspaces linked to distinct abilities, enabling effective behavior control through tailored tuning.

(analogous to the *mirror neuron system*) and the **robustness subspace** (similar to the *hippocampus*), along with an **entangled subspace** that is challenging to decouple. Addressing challenge **C2**, Parenting adopts a type-tailored tuning strategy, applying specific optimizations for different subspaces. Initially, for the entangled subspace, which consists of parameters critical to both adherence and robustness—similar to the *prefrontal cortex* in humans, which is responsible for perceiving the external environment (Puig and Gulledge, 2011)—Parenting employs a novel document extraction task to reinforce training, thereby enhancing both adherence and robustness simultaneously. Concurrently, Parenting employs a boundary-controlled fine-tuning strategy to prevent contradictory supervision signals from contaminating the adherence and robustness subspaces. To summarize, we highlight our contributions as follows:

- We propose a novel insight from the perspective of parameter space to achieve fine-grained decoupling and fine-tuning optimization of adherence and robustness in RALMs.
- We propose a key parameter mining method and a type-tailored tuning strategy, which effectively harmonize and enhance the supervisory signals representing adherence and robustness.
- We conduct extensive experiments across multiple models and datasets, and the results demonstrate that our approach not only achieves a balance and enhancement of model performance, but also enhances its adaptability and robustness when facing out-of-distribution (OOD) data.

2 Related Work

2.1 Retrieval-Augmented Language Models

RALMs integrate external knowledge bases, allowing the generation process of LLMs to be adjusted based on the most up-to-date and relevant documents or knowledge (Guu et al., 2020). The classic "Retrieve-Read" framework leverages the initial input as a query to retrieve relevant information from an external corpus. The retrieved knowledge is then integrated into the input and incorporated into the model's generation process (Gao et al., 2023; Fan et al., 2024b). For instance, kNN-LM (Khandelwal et al., 2019) retrieves external memory using k-nearest neighbors. To prevent confusion and resource waste caused by excessive retrieval, FLARE (Jiang et al., 2023), Self-RAG (Asai et al.), and DRAGIN (Su et al., 2024) combine LLMs' own information needs to determine when to retrieve and what queries to generate. Although these approaches are effective, they still face challenges such as confusion caused by conflicts between internal and external knowledge, as well as misleading effects from external noise (Mallen et al., 2023).

2.2 Addressing Noisy Contexts and Internal-External Knowledge Conflicts

After retrieving external knowledge, effectively integrating it with the internal knowledge of LLMs is essential for enhancing the quality of the final output response (Fan et al., 2024b). On the one hand, retrieval algorithms cannot achieve complete accuracy, and therefore RALMs will inevitably introduce task-irrelevant noises (Creswell et al.). To address the above issue, SA-RetRobust (Yoran et al.) enhances the noise robustness of LLMs by introducing an additional fine-tuning step. Info-RAG (Xu et al., 2024a) and RAAT (Fang et al., 2024) design unsupervised training and adaptive adversarial training strategies, respectively. On the other hand, models tend to become confused when there is conflict between external knowledge and the internal memory of LLMs (Wu et al., 2024b), particularly when the external factual evidence is incoherent or semantically incomplete (Tan et al., 2024; Xie et al.). To address the aforementioned challenge, prompt-based methods improve LLMs' faithfulness to context through carefully designed prompting strategies (Zhou et al., 2023). IR-CAN (Shi et al., 2024a) identifies and reweights adherence-specific neurons. Decoding-based methods measure differences in output probabilities between conditions with and without context (Shi et al., 2024b; Jin et al., 2024a). KAFT (Li et al., 2023a) focuses on constructing specific instruction-tuning datasets. PH3 (Jin et al., 2024b) mitigates knowledge conflicts by pruning negative attention heads using the path patching technique. KnowPO (Zhang et al., 2025) guides the model to avoid errors in knowledge selection by constructing a knowledge conflict dataset and incorporating preference optimization training.

3 Task Formulation

3.1 Task Definition of RALMs

Given a natural language query q as input, a standard RALMs typically employs a retriever to retrieve a set of documents $D = \{d_1, \ldots, d_k\}$ from an external knowledge base $\mathbb{C}$, where k is the window size of the retrieval process. Then, during the inference process, the retrieved context D is concatenated with the query q and is fed into LLM Θ to generate the answer ans_Θ, denoted as $ans_\Theta = \Theta(q \| D)$. Besides, LLMs will use their parametric knowledge $\alpha = \Theta(q)$ to answer such query without retrieved context. We refer to the evidence that can assist in answering q as d_{golden}. Next, we define a relevance label y to indicate whether D contains this evidence (i.e., if $d_{golden} \in D$, then $y = 1$; otherwise, $y = 0$).

Following Li et al. (2023a), our goal is to enable the fine-tuned LLM Θ' to derive answers based on retrieved context when it contains valid evidence, even if this evidence conflicts with its internal memory (we refer to such situations as **conflicting contexts**.). Conversely, when the retrieved context contains only semantically similar yet irrelevant noise (consistent with Yoran et al., we define them as **irrelevant contexts**), the model should detect the flaws in the context and rely on its internal knowledge to respond. This not only effectively prevents the model from confidently producing incorrect answers due to noisy input, but is also more user-friendly than consistently refusing to answer:

$$ans_{\Theta'} = \begin{cases} ans_{golden}, & \text{if } y = 1 \\ \alpha, & \text{if } y = 0 \end{cases}, \tag{1}$$

where ans_{golden} is the correct answer from d_{golden}.

3.2 Definition of Parameter Units

Knowledge organization within LLMs can be described as independent, modular regions, where

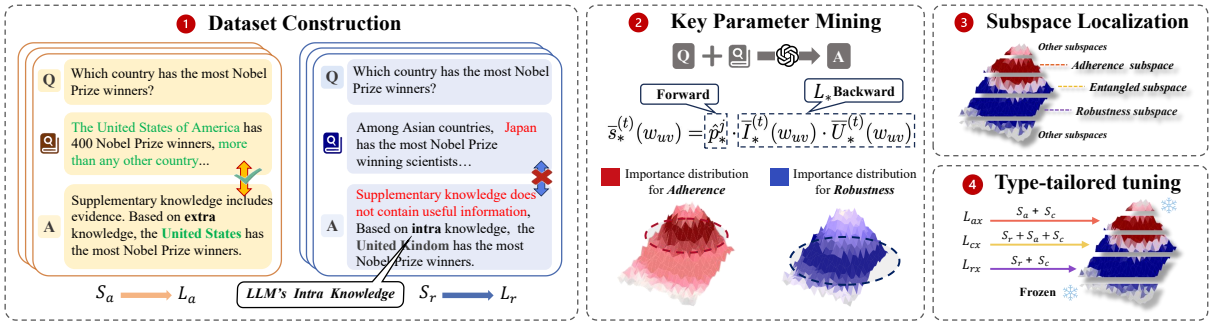

Figure 2: The overview of our proposed Parenting.

distinct matrices are typically regarded as fundamental units for knowledge storage (Wang et al., 2024a, 2023a). Therefore, in this paper, we define the fundamental structural elements that store skills, termed "parameter units", as individual matrices within the model. During full parameter fine-tuning, the matrices of the Multi-Head Self-Attention (MHA) and the Feed-Forward Neural Network (FFN) are regarded as basic parameter units. When employing Parameter-Efficient Fine-Tuning (PEFT) techniques (Ma et al., 2024; Lin et al., 2024) to accelerate training, we consider matrices A and B in LoRA (Hu et al.) as separate basic parameter units. A subspace is composed of multiple parameter units. The model Θ with n parameter units is denoted as $\mathcal{E} = \{e_1, \ldots, e_n\}$. The notations used throughout this paper are provided in Appendix A.2.

4 Methodology

Parenting includes four components:

- **Dataset Construction**, which constructs two supervised fine-tuning (SFT) datasets to stimulate the adherence and robustness of LLMs.
- **Key Parameter Mining**, which measures parameter importance by combining forward activations with gradient trajectories.
- **Subspace Localization**, which identifies and locates different types of subspaces.
- **Type-Tailored Tuning**, which designs specific and appropriate fine-tuning strategies for each type, aiming to achieve a balanced enhancement of adherence and robustness.

Figure 2 presents a detailed overview of our proposed Parenting framework, with subsequent subsections detailing each component.

4.1 Dataset Construction

To elicit adherence and robustness of LLMs, we design two datasets derived from SQuAD2.0 dataset (Rajpurkar et al., 2018), a reading comprehension dataset encompassing multiple general domains, with a substantial corpus of documents and associated question answering (QA) tasks. We first extract the LLMs' parametric knowledge α using probes for each SQuAD2.0 question. Notably, SQuAD2.0 employs manual annotations to indicate whether a document provides an answer to a specific question. Based on these annotations, we construct a SFT dataset S_a that promotes adherence and another dataset S_r that enhances robustness.

For dataset S_a, we begin by selecting ground truth answers and corresponding evidence documents from SQuAD2.0 that conflict with the parametric knowledge α. Following Li et al. (2023a), we further introduce fictional answers and related documents to effectively expand the dataset. S_a can be represented as $S_a = \{(q_a^i, D_a^i, ans_a^i) \mid i = 1, 2, \ldots, m_a\}$, where m_a denotes the number of samples in S_a. The cross-entropy loss on this dataset is denoted as $\mathcal{L}_a$. For dataset S_r, regarding the context, we select the manually annotated irrelevant noise documents from the dataset. We expect the model to indicate that the context lacks the clues necessary to answer the question and to generate its original parametric knowledge α, the cross-entropy loss on which is denoted as $\mathcal{L}_r$. Similarly, S_r can be represented as $S_r = \{(q_r^i, D_r^i, ans_r^i) \mid i = 1, 2, \ldots, m_r\}$.

4.2 Key Parameter Mining

The original pre-trained parameters across different layers store and process information in distinct ways, which influences their functionality and the types of tasks they are best suited for (Chuang et al., 2023; Meng et al., 2022). Therefore, we propose

using the layer-wise activation levels of the original pre-trained parameters across various inputs to inform traditional gradient-based sensitivity calculations, in order to assess the overall contributions of different parameters to adherence and robustness, as follows:

Forward activation probability computation. In the Transformer architecture, the FFN layer functions as a key-value unit capable of storing knowledge (Meng et al., 2022) and is closely tied to the model's responses to different inputs (Gurnee et al., 2024), which can be formalized as follows:

$$h^j = \left(\text{Swish}(\tilde{h}^j W_1^j) \otimes \tilde{h}^j V^j\right) \cdot W_2^j, \quad (2)$$

where $W_1^j \in \mathbb{R}^{d\times 4d}$, $V^j \in \mathbb{R}^{d\times 4d}$, and $\boldsymbol{W}_2^j \in \mathbb{R}^{4d\times d}$ are parameters, $\tilde{h}^j \in \mathbb{R}^d$ and $h^j \in \mathbb{R}^d$ represent the hidden states of a specific token at the j-th layer ($j \in \{1, \ldots, l\}$) after processing through MHA and FFN, respectively. d is the dimension of the hidden states. SwiGLU is an activation function that has been widely adopted in recent LLM architectures (Touvron et al., 2023; Bai et al., 2023). In SwiGLU, the positive activation output of the Swish function indicates that the corresponding neurons are activated, thereby influencing the final prediction outcome through forward propagation (Nair and Hinton, 2010). Inspired by Tang et al. (2024), we evaluate the sensitivity of the original pre-trained parameters to different abilities by calculating the expected positive activation outputs of neurons when faced with different inputs, i.e., $C_r = \left\{\left(q_r^i, D_r^i\right) \mid \left(q_r^i, D_r^i, ans_r^i\right) \in S_r\right\}$ and $C_a = \left\{\left(q_a^i, D_a^i\right) \mid \left(q_a^i, D_a^i, ans_a^i\right) \in S_a\right\}$:

$$p_{*,k}^j = \mathbb{E}\left(\mathbb{I}\left(\text{Swish}(\tilde{h}^j W_1^j)_k > 0\right) \mid C_*\right), \quad (3)$$

where $\text{Swish}\left(\tilde{h}^j W_1^j\right)_k$ represents the activation value of the k-th neuron in the j-th layer. $\mathbb{I}$ is the indicator function, and the $*$ **denotes either adherence (symbolized by a) or robustness (symbolized by r)**. Next, we calculate the average activation probability $\hat{p}_*^j$ of all FFN neurons within a single layer to measure the sensitivity of the original pre-trained parameters across different layers concerning adherence and robustness:

$$p_*^j = \frac{1}{4d}\sum_{k=1}^{4d} p_{*,k}^j, \quad \hat{p}_*^j = \frac{\exp\left(p_*^j\right)}{\sum_{j=1}^{l}\exp\left(p_*^j\right)}. \quad (4)$$

Gradient-based sensitivity and uncertainty calculation. To quantify the sensitivity of parameters to training loss, we calculate the product of gradients and weights for all trainable parameters (Molchanov et al., 2019; Zhang et al., 2022):

$$I_*(w_{uv}) = |w_{uv} \cdot \nabla_{w_{uv}} \mathcal{L}_*|. \quad (5)$$

This metric approximates the impact on the loss when a specific parameter is set to zero (Molchanov et al., 2019). However, it tends to exhibit high variance due to random sampling and the complex dynamics of the training process. To mitigate this issue, we consider integrating sensitivity smoothing and uncertainty quantification (Zhang et al., 2023):

$$\begin{aligned}\overline{I}_*^{(t)}(w_{uv}) =& \alpha_1 \overline{I}_*^{(t-1)}(w_{uv}) + (1-\alpha_1) I_*^{(t)}(w_{uv}),\\ \overline{U}_*^{(t)}(w_{uv}) =& \alpha_2 \overline{U}_*^{(t-1)}(w_{uv}) + (1-\alpha_2)\\ & \cdot |I_*^{(t)}(w_{uv}) - \overline{I}_*^{(t)}(w_{uv})|,\end{aligned} \quad (6)$$

where $0 < \alpha_1, \alpha_2 < 1$ are smoothing factors, and t is the iteration number.

Importance score calculation and aggregation. Subsequently, we use $\hat{p}_*^j$, obtained from the forward propagation, as a layer-specific clue to guide the sensitivity computation based on the gradient trajectory. Inspired by the computational approach in Zhang et al. (2023), we introduce a multiplicative operation that combines the activation levels of the original pre-trained parameters under different inputs with the smoothed sensitivity and uncertainty. The final importance score for each parameter in terms of adherence or robustness ($* \in \{a, r\}$) is computed as follows:

$$\overline{s}_*^{(t)}(w_{uv}) = \hat{p}_*^j \cdot \overline{I}_*^{(t)}(w_{uv}) \cdot \overline{U}_*^{(t)}(w_{uv}), \quad (7)$$

where j represents the layer containing the parameter w_{uv}. To assess the overall contribution of each parameter unit to model performance, we calculate the average importance of each individual parameter within the parameter unit:

$$\mathcal{I}_*(e) = \frac{1}{d_1 \times d_2}\sum_{u=1}^{d_1}\sum_{v=1}^{d_2} \overline{s}_*\left(w_{uv}\right), \quad (8)$$

where $\mathcal{I}_*(e)$ measures the overall importance of all parameters within each parameter unit, with higher values indicating greater importance. As a result, for model Θ, which consists of n parameter units, we can obtain the distribution of importance scores $\mathcal{I}_*(\mathcal{E}) \in \mathbb{R}^n$ for adherence or robustness.

4.3 Subspace Localization

We then conduct an interactive analysis of importance distribution across different behaviors, decoupling and identifying various types of subspaces.

We employ the Z-score, a common statistical measure (Altman et al., 2017), to standardize importance scores for adherence and robustness:

$$\mathcal{Z}_*(e_i) = \Big(\mathcal{I}_*(e_i) - \mu_*\Big)/\sigma_*, \tag{9}$$

where $\mathcal{Z}_*(e_i)$ is the standardized importance score of the i-th parameter unit for adherence or robustness, μ_* is the mean score, and σ_* is the standard deviation. Based on $\{\mathcal{Z}_*(e_i) \mid i \in \{1, \ldots, n\}\}$, we identify and locate four types of subspaces:

1. Entangled Subspace. This subspace consists of parameter units that are crucial for both adherence and robustness, and we assume it represents the model's ability to perceive and analyze context (queries and external knowledge). Specifically, we select parameter units with Z-scores greater than 1 for both adherence and robustness:

$$\mathcal{E}_c = \{e_i \mid \mathcal{Z}_a(e_i) > 1 \ \wedge \ \mathcal{Z}_r(e_i) > 1; i \in \{1, \ldots, n\}\}. \tag{10}$$

2. Adherence Subspace. This subspace is composed of parameter units that are important for adherence but not for robustness. It relates to the model's ability to solve problems by replicating and summarizing contextual information:

$$\mathcal{E}_{ax} = \{e_i \mid \mathcal{Z}_a(e_i) > 1 \ \wedge \ \mathcal{Z}_r(e_i) < 1; i \in \{1, \ldots, n\}\}. \tag{11}$$

3. Robustness Subspace. This subspace comprises parameter units that are crucial for robustness but less significant for adherence. It is associated with the model's capacity to solve problems by retrieving internal memories:

$$\mathcal{E}_{rx} = \{e_i \mid \mathcal{Z}_r(e_i) > 1 \ \wedge \ \mathcal{Z}_a(e_i) < 1; i \in \{1, \ldots, n\}\}. \tag{12}$$

4. Other Subspaces. These subspaces consist of all remaining parameter units outside the three previously mentioned subspaces, representing the other capabilities of LLMs, denoted as $\mathcal{E}_o$.

4.4 Type-Tailored Tuning

After decoupling and identifying key subspaces related to knowledge control, we develop specific and appropriate fine-tuning strategies tailored to each subspace.

4.4.1 Document Extraction Task

To more effectively optimize the entangled subspace, we design a document extraction dataset $S_c = \{(q_c^i, D_c^i, ans_c^i) \mid i = 1, 2, \ldots, m_c\}$. Specifically, we curate documents from the SQuAD2.0 dataset, gathering three types of documents for each question: relevant documents that contain the question's answer, noise documents that are on the same topic, and noise documents from different topics, simulating a variety of retrieval context scenarios. In this document extraction task, we present the mixed context of the question along with the three types of documents as input and expect the LLMs to identify the specific document types and accurately restate their content. We denote the cross-entropy loss on S_c as $\mathcal{L}_c$. To simultaneously enhance adherence and robustness, we train $\mathcal{E}_c$ using a combination of S_a, S_r, and S_c:

$$\begin{aligned} \mathcal{Z}_*(\mathcal{E}_c) &= \mathbb{E}\left[\mathcal{Z}_*(e_i) \mid e_i \in \mathcal{E}_c\right], \\ \gamma_* &= \frac{\exp\left(\mathcal{Z}_*(\mathcal{E}_c)\right)}{\exp\left(\mathcal{Z}_a(\mathcal{E}_c)\right) + \exp\left(\mathcal{Z}_r(\mathcal{E}_c)\right)}, \\ \mathcal{L}_{cx} &= \delta_1\left(\gamma_a \times \mathcal{L}_a + \gamma_r \times \mathcal{L}_r\right) + (1 - \delta_1)\,\mathcal{L}_c, \end{aligned} \tag{13}$$

where $0 < \delta_1 < 1$ acts as a reweighting factor between the original task and the added task. We determine the weights of the two loss terms adaptively by calculating expected Z-scores for adherence and robustness on $\mathcal{E}_c$.

4.4.2 Boundary-Controlled Fine-Tuning

To prevent contradictory supervision signals from contaminating the adherence and robustness subspaces, we propose a boundary-controlled fine-tuning strategy:

- **Adherence Subspace.** For $\mathcal{E}_{ax}$, we ensure that it remains unaffected by gradients related to robustness, allowing the LLMs' adherence capabilities to be fully optimized:

$$\mathcal{L}_{ax} = \delta_1\mathcal{L}_a + (1 - \delta_1)\mathcal{L}_c. \tag{14}$$

- **Robustness Subspace.** For $\mathcal{E}_{rx}$, we ensure that it remains unaffected by gradients related to adherence, thereby enabling a breakthrough in enhancing robustness:

$$\mathcal{L}_{rx} = \delta_1\mathcal{L}_r + (1 - \delta_1)\mathcal{L}_c. \tag{15}$$

Additionally, we maintain the initialized weights of $\mathcal{E}_o$ to prevent divergence from pre-trained weights, preserving the other capabilities of LLMs. We present the detailed algorithm of Parenting in Appendix A.1.

5 Experiment

5.1 Experimental Setup

Datasets. We construct the Parenting training dataset based on SQuAD2.0 (Rajpurkar et al.,

2018). The test datasets comprise the following three types: (1) **SQuAD2.0-Eval**. Following Li et al. (2023a), we further simulate complex retrieval scenarios using annotation information from SQuAD2.0 to more effectively evaluate the adherence and robustness of RALMs, resulting in the creation of SQuAD2.0-Eval. (2) **Open-source RAG datasets:** RGB (Chen et al., 2024) and KNOT (Liu et al., 2024) are two general-domain QA datasets designed to evaluate LLMs' ability to effectively utilize retrieved information while withstanding various imperfections in the retrieval process. (3) **Domain-specific dataset:** CMB (Wang et al., 2024b) is a multi-task QA dataset in the medical domain, comprising 269,359 questions across various categories and disciplines. Due to quantity constraints, we randomly sample 4,000 questions for testing. Further details about the datasets are provided in the Appendix A.3.

Baselines. We evaluate Parenting against the following widely-used and state-of-the-art RAG baseline methods: ***Base Model (Base)*** responds to queries by leveraging external knowledge, functioning as the fundamental retrieve-read framework (Gao et al., 2023). We choose LLaMA2-7B-Chat (Touvron et al., 2023) and Qwen1.5-14B-Chat (Team, 2024) as the base models and investigate the improvements introduced by Parenting. ***Prompt-based Method (Prompt)*** (Zhou et al., 2023) utilizes a carefully designed prompt strategy to enhance adherence. ***COT-VE*** (Zhao et al., 2023) guides LLMs to edit potentially incorrect or outdated rationales using external knowledge. ***KAFT*** (Li et al., 2023a) enhances the adherence and robustness of LLMs by constructing a specific fine-tuning dataset. ***CAD*** (Shi et al., 2024b) reduces the model's reliance on internal knowledge through contrastive decoding. ***RAAT*** (Fang et al., 2024) improves the noise robustness of LLMs through adaptive adversarial training. ***IRCAN*** (Shi et al., 2024a) enhances the adherence of LLMs by detecting and boosting adherence-related neurons.

It is worth noting that the baseline method KAFT is implemented by training all subspaces using the full training dataset uniformly. To ensure the fairness of experimental comparisons, for the training-based baseline methods KAFT and RAAT, we construct training datasets based on the same source data as Parenting, following the strategies outlined in their original methods. As for the IRCAN method, which relies on a specific dataset to identify context-aware neurons, we likewise use the same dataset to conduct the corresponding neuron detection task.

We also compare Parenting with three reduced variants to conduct ablation studies:

- Parenting$_{l-}$ removes layer-level clues based on forward activations when mining key parameters.
- Parenting$_{b-}$ trains $\mathcal{E}_c$, $\mathcal{E}_{ax}$, and $\mathcal{E}_{rx}$ using three datasets (S_a, S_r, and S_c) simultaneously.
- Parenting$_{e-}$ removes the document extraction task during type-tailored tuning.

Metrics. In line with Li et al. (2023a), we evaluate the adherence and robustness of RALMs using two distinct metrics. For adherence, we supplement LLMs with conflicting contexts from the test set as additional knowledge and measure the proportion of cases where the LLMs follow the conflicting evidence to generate their responses, denoted as R_{Ad}. Regarding robustness, we use irrelevant contexts from the test set as supplementary knowledge and measure the proportion of cases where the model successfully avoids extracting answers from these irrelevant contexts, which is denoted as R_{Ro}. Higher values of R_{Ad} and R_{Ro} indicate better adherence and robustness, respectively.

More implementation details and experimental results are provided in Appendix A.5, A.6, A.7, A.8, A.9, A.10, A.12, and A.13.

5.2 Experimental Results

5.2.1 Main Results

Table 1 displays the performance of Parenting compared to baseline methods across three datasets. Overall, **our proposed Parenting excels on all backbone models and datasets, significantly surpassing current state-of-the-art methods.** This highlights the effectiveness of Parenting in fine-grained decoupling in parameter space and its customized fine-tuning capabilities. From Table 1, we can also observe several insights. Firstly, methods that rely on LLMs' basic interaction capabilities at the prompt level, such as Prompt and CoT-VE, often depend indiscriminately on external knowledge due to LLMs' limited noise recognition capabilities. Secondly, IRCAN and RAAT emphasize adherence and robustness respectively, and each inevitably sacrifices another key capability when integrating knowledge. Additionally, a notable phenomenon is that, compared to the baseline, **Parenting achieves a more balanced improvement in both adherence and robustness.** This is because we avoid

Method	LLM	LLaMA2-7B-Chat						Qwen1.5-14B-Chat					
	Dataset	SQuAD		RGB		KNOT		SQuAD		RGB		KNOT	
	Metric	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
Baselines	Base	44.20	16.40	68.00	29.50	45.09	20.54	59.71	21.37	68.50	34.00	52.73	22.69
	Prompt	46.40	9.50	69.50	17.50	43.74	15.01	63.69	18.29	78.00	32.50	53.19	16.77
	COT-VE	47.45	17.20	68.50	31.00	46.39	21.27	64.23	25.94	66.00	38.00	55.42	24.59
	KAFT	54.15	18.43	71.50	30.50	47.09	22.92	75.55	28.00	82.50	39.50	56.23	25.94
	CAD	45.72	8.80	70.00	16.00	44.14	12.79	64.98	17.20	72.00	28.50	53.14	14.79
	RAAT	39.25	40.73	49.50	41.00	25.09	35.58	60.30	40.61	71.50	42.50	46.87	37.92
	IRCAN	53.17	13.50	72.50	20.00	46.51	16.50	76.24	16.15	82.00	30.50	55.21	17.45
Ours	**Parenting**	**69.24**	**44.85**	**79.50**	**45.50**	**67.42**	**42.82**	**80.47**	**47.58**	**84.50**	**46.50**	**73.27**	**43.34**
Ablation	Parenting$_{l-}$	66.15	39.78	77.00	40.00	64.35	37.63	78.06	42.25	83.50	42.50	67.87	40.91
	Parenting$_{b-}$	55.90	20.70	73.50	31.50	49.26	23.05	77.28	30.92	83.00	41.50	58.73	27.36
	Parenting$_{e-}$	62.57	36.71	75.50	37.50	60.55	34.98	77.59	38.24	83.00	42.50	64.13	37.28

Table 1: Performance comparisons (%) on SQuAD2.0-Eval (SQuAD), RGB, and KNOT. The best performance is in **boldface** and the best results among the baselines are underlined.

Method	LLaMA2-7B		Qwen1.5-14B	
	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
Base	54.28	20.17	59.07	21.39
Parenting	**75.79**	**48.21**	**79.63**	**45.55**

Table 2: Performance comparisons (%) on CMB.

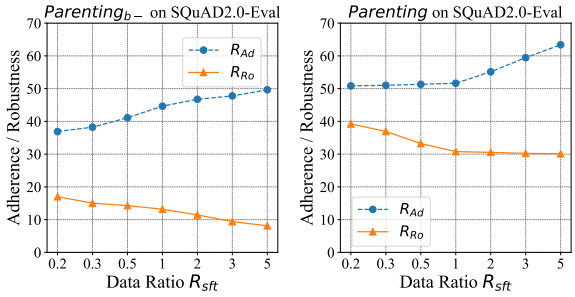

Figure 3: Analysis of the behavioral patterns of Parenting (**Right**) and Parenting$_{b-}$ (**Left**) under varying data ratios (on SQuAD2.0-Eval with LLaMA2-7B-Chat).

contaminating key parameters with conflicting supervisory signals during the fine-tuning process.

5.2.2 Ablation Study

To fully understand the contribution of each component in Parenting to the overall performance, we conducted ablation studies. As shown in Table 1, the performance decline of Parenting$_{l-}$ compared to Parenting indicates that the hierarchical clues provided by forward activations are essential for precisely locating parameters closely related to adherence and robustness. The superior performance of Parenting over Parenting$_{b-}$ and Parenting$_{e-}$ demonstrates that our proposed type-tailored tuning strategy effectively mitigates the negative impact of conflicting supervision signals on the model's behavior, allowing for the full optimization of both adherence and robustness.

5.3 Analysis

5.3.1 Generalization Analysis

To demonstrate the strong generalization capability of Parenting beyond general-domain training sets, we conduct experiments on the medical benchmark CMB (Wang et al., 2024b). By utilizing medical triples and documents from the Huatuo-26M dataset (Li et al., 2023b) to build supplementary medical context, we create a dataset to evaluate the medical knowledge utilization of LLMs, based on 4,000 CMB questions. As shown in Table 2, LLMs trained with Parenting in a general domain continue to demonstrate superior knowledge selection capabilities when transferred to specific vertical domains. These results further **validate the generalization ability and effectiveness of Parenting.**

5.3.2 Behavioral Tendency Exploration

To gain a more intuitive understanding of Parenting's ability to balance adherence and robustness, we explore the behavioral patterns of Parenting and Parenting$_{b-}$ under different ratios of supervised data. Specifically, for the two types of supervision signals that are challenging to balance due to their contradictory nature, we randomly select two training subsets, S_a^{sub} and S_r^{sub}, from S_a and S_r respectively, with $|S_a^{sub}| = |S_r^{sub}| = |S_a|/2$. Within these subsets, we continuously adjust the ratio between these two signals, denoted as $R = |S_a^{sub}|/|S_r^{sub}|$. When $R > 1$, we keep $|S_r^{sub}|$ con-

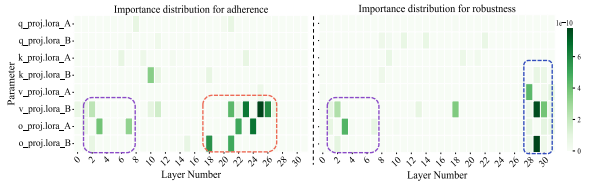

Figure 4: Visualization of parameter importance distributions $\mathcal{I}_a(\mathcal{E})$ and $\mathcal{I}_r(\mathcal{E})$ for adherence (**Left**) and robustness (**Right**) in LLaMA2-7B-Chat.

stant, and when $R < 1$, we keep $|S_a^{sub}|$ constant. As shown in Figure 3, as R increases, the adherence of the LLM fine-tuned with Parenting steadily rises, while its robustness remains stable. Conversely, as R decreases, the robustness of the LLMs improves, with adherence remaining largely unchanged. However, in the case of Parenting$_{b-}$, there is a tug-of-war between adherence and robustness, where an increase in the proportion of one signal results in a significant decline in the other, preventing full optimization of both capabilities. This indicates that separating the conflicting supervision signals allows for a more balanced optimization of adherence and robustness.

5.3.3 Visualization of Parameter Units

Figure 4 visualizes the distribution of parameter unit importance for adherence and robustness in LLaMA2-7B-Chat. We can observe the parameter units (entangled) within the Entangled Subspace (purple boxes). It also highlights the parameter units (adherence-specific) in the Adherence Subspace (the red box) and those (robustness-specific) in the Robustness Subspace (the blue box). Additionally, we observe that adherence-specific units are primarily located in the middle and upper-middle layers. This aligns with findings from studies Chuang et al. (2023) and Dai et al. (2022), which indicate that these layers play a significant role when LLMs copy information from inputs. Robustness-specific units are mainly found in the upper layers, with a few in the middle layers. This is consistent with the study Chuang et al. (2023), suggesting that internal factual knowledge is typically encoded in the higher layers of LLMs. We also note that entangled parameter units are predominantly located in the middle to lower-middle layers, aligning with the study Fan et al. (2024a).

5.4 Noise Recognition Capability

To further validate the effectiveness of our framework in enhancing the context-awareness of LLMs, we construct a noise identification experiment on the SQuAD2.0-Eval and RGB datasets to evaluate the model's ability to analyze the association between queries and external knowledge. In detail, for each question, we select a topic-relevant document based on the dataset's annotation information. This document may either provide supporting evidence or serve as a noise document that does not answer the question. We then prompt LLMs to classify the type of document. The dataset contains a balanced number of labels for both evidence and noise categories, and accuracy is used as the evaluation metric for this task. We focus on the comparison between Parenting and two existing frameworks: KAFT and RAAT. The results in Table 3 show that Parenting stands out among the evaluated frameworks, particularly surpassing RAAT, which is specifically designed to enhance noise robustness. This result highlights the effectiveness of Parenting in enhancing the context-awareness capabilities of LLMs.

Method	SQuAD	RGB
LLaMA2-7B-Chat	55.49	63.50
KAFT	59.24	65.00
RAAT	62.48	67.00
Parenting	**69.89**	**72.50**

Table 3: Performance comparisons of Parenting with other RAG methods (all utilizing the same LLaMA2-7B-Chat backbone) on the noise identification task. All results are reported in accuracy (%).

6 Conclusion

In this work, we propose a novel and versatile RAG knowledge integration framework, dubbed Parenting, which optimizes LLMs' knowledge selection through parameter decoupling and tailored tuning, thereby establishing an effective control mechanism for both internal and external knowledge. Parenting first performs a fine-grained analysis and decoupling of adherence and robustness in the parameter space using a key parameter mining method. It then applies a type-tailored tuning strategy to systematically and thorough optimize behavior-specific parameters. Experimental results across multiple datasets and models consistently demonstrate the effectiveness of Parenting.

Limitations

Despite the promising results obtained in our work, it is important to acknowledge its limitations.

On the one hand, we inevitably encountered some errors where the model failed to adhere to conflicting external knowledge, particularly on the KNOT dataset, which requires knowledge-based reasoning. According to the original setup of the dataset (Liu et al., 2024), we observed that the vast majority of these errors fell under the KNOT-I category, which requires implicit reasoning: the model must independently decompose the question, infer a plausible reasoning path, and derive the answer based on conflicting knowledge. We believe this is primarily due to the current limitations of LLMs in actively integrating internal and external knowledge for reasoning. In the future, we plan to explore the inclusion of more diverse datasets and investigate finer-grained parameter unit partitioning to further enhance the model's reasoning capabilities. First, within the Parenting framework, we treat an individual matrix in LLMs as a parameter unit, yet this granularity may still introduce redundancy within the matrix. There could be a possibility to further refine the model's capability segmentation, such as decomposing the product of two low-rank matrices in LoRA into even finer units (Wang et al., 2023b; Feng et al., 2024a), which might enhance model performance. Additionally, the construction of our training dataset solely relied on resources from SQuAD2.0. Although we have demonstrated the generalized capabilities of the Parenting fine-tuned model across multiple benchmark datasets, there is still room for improvement. In the future, we plan to gather a more diverse array of high-quality datasets, aim to launch a more effective version of Parenting, and establish a broader and more varied benchmark to assess the knowledge integration abilities of RALMs.

On the other hand, although we focus on the reader phase (where LLMs utilize knowledge to answer questions) and enhance model robustness through instruction tuning, which has been empirically validated as effective, the overall performance of RAG systems remains constrained by interference from low-quality documents. This issue is particularly pronounced when content is sourced from external platforms such as the web, where poor-quality materials can significantly degrade generation quality. Addressing this challenge remains non-trivial. To mitigate this issue, we plan to explore training a discriminator capable of identifying low-quality documents and incorporating its results into the prompt. Additionally, we will consider calibrating the uncertainty of LLMs and using it as a clue for evaluating document quality.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62172011).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Edward I Altman, Małgorzata Iwanicz-Drozdowska, Erkki K Laitinen, and Arto Suvas. 2017. Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. *Journal of international financial management & accounting*, 28(2):131–171.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola:

Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. 2024a. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024b. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv preprint arXiv:2405.20978*.

Yujie Feng, Xu Chu, Yongxin Xu, Zexin Lu, Bo Liu, Philip S Yu, and Xiao-Ming Wu. 2024a. Kif: Knowledge identification and fusion for language model continual learning. *arXiv preprint arXiv:2408.05200*.

Yujie Feng, Xu Chu, Yongxin Xu, Guangyuan Shi, Bo Liu, and Xiao-Ming Wu. 2024b. Tasl: Continual dialog state tracking via task skill localization and consolidation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1266–1279.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards llm-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755.

Yujie Feng, Xujia Wang, Zexin Lu, Shenghong Fu, Guangyuan Shi, Yongxin Xu, Yasha Wang, Philip S Yu, Xu Chu, and Xiao-Ming Wu. 2025. Recurrent knowledge identification and fusion for language model continual learning. *arXiv preprint arXiv:2502.17510*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Michael J Hawrylycz, Ed S Lein, Angela L Guillozet-Bongaarts, Elaine H Shen, Lydia Ng, Jeremy A Miller, Louie N Van De Lagemaat, Kimberly A Smith, Amanda Ebbert, Zackery L Riley, et al. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xu Chu, et al. 2024a. Tc-rag: Turing-complete rag's case study on medical llm systems. *arXiv preprint arXiv:2408.09199*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. 2024b. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. *arXiv preprint arXiv:2312.15883*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024a.

Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1193–1215.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Yang Lin, Xinyu Ma, Xu Chu, Yujie Jin, Zhibang Yang, Yasha Wang, and Hong Mei. 2024. Lora dropout as a sparsity regularizer for overfitting control. *arXiv preprint arXiv:2404.09610*.

Yantao Liu, Zijun Yao, Xin Lv, Yuchen Fan, Shulin Cao, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Untangle the knot: Interweaving conflicting knowledge and reasoning skills in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17186–17204.

Xinyu Ma, Xu Chu, Zhibang Yang, Yang Lin, Xin Gao, and Junfeng Zhao. 2024. Parameter efficient quasi-orthogonal fine-tuning via givens rotation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 33686–33729. PMLR.

Xinyu Ma, Yasha Wang, Xu Chu, Liantao Ma, Wen Tang, Junfeng Zhao, Ye Yuan, and Guoren Wang. 2023. Patient health representation learning via correlational sparse prior of medical features. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11769–11783.

Xinyu Ma, Yifeng Xu, Yang Lin, Tianlong Wang, Xu Chu, Xin Gao, Junfeng Zhao, and Yasha Wang. 2025. DRESSing up LLM: Efficient stylized question-answering via style subspace editing. In *The Thirteenth International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

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.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. 2019. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11264–11272.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

David S Olton, James T Becker, and Gail E Handelmann. 1979. Hippocampus, space, and memory. *Behavioral and Brain sciences*, 2(3):313–322.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

M Victoria Puig and Allan T Gulledge. 2011. Serotonin and prefrontal cortex function: neurons, networks, and circuits. *Molecular neurobiology*, 44:449–464.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Giacomo Rizzolatti and Laila Craighero. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27(1):169–192.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024a. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *arXiv preprint arXiv:2406.18406*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Qwen Team. 2024. Introducing qwen1. 5.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, et al. 2024a. Knowledge mechanisms in large language models: A survey and perspective. *arXiv preprint arXiv:2407.15017*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. 2023a. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. 2023b. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, et al. 2024b. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kevin Wu, Eric Wu, and James Zou. 2024a. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Preprint*.

Kevin Wu, Eric Wu, and James Zou. 2024b. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024a. Unsupervised information refinement training of large language models for retrieval-augmented generation. *arXiv preprint arXiv:2402.18150*.

Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023a. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830.

Yongxin Xu, Xinke Jiang, Xu Chu, Rihong Qiu, Yujie Feng, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2025. Dearllm: Enhancing personalized healthcare via large language models-deduced feature correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 941–949.

Yongxin Xu, Xinke Jiang, Xu Chu, Yuzhen Xiao, Chaohe Zhang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2024b. Protomix: Augmenting health status representation learning via prototype-based mixup. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3633–3644.

Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023b. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *IJCAI*, volume 23, pages 4921–4929.

Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5357–5365.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Qingru Zhang, Simiao Zuo, Chen Liang, Alexander Bukharin, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2022. Platon: Pruning large transformer models with upper confidence bound of weight importance. In *International conference on machine learning*, pages 26809–26823. PMLR.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25895–25903.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

A Appendix

Algorithm 1 Training Algorithm of Parenting

Input: Training dataset S_a, S_r, and S_c; total probing iterations T_{prob}; total training iterations T_{tr}; initial LLMs Θ.

for $* \in \{a, r\}$ **do**
  for B in mini-batches of $\mathcal{C}_*$ **do**
    Calculate and store the activation value.
  end for
  Calculate expected activation value $p^j_{*,k}$ via Eq. (3).
  Obtain the sensitivity $\hat{p}^j_*$ through aggregation, for $j \in \{1, \ldots, l\}$ via Eq. (4).
end for
for $* \in \{a, r\}$ **do**
  for $t = 1, \ldots, T_{prob}$ **do**
    Sample a mini-batch from $\mathcal{S}_*$ and compute the gradient $\nabla \mathcal{L}_*$;
    Compute the sensitivity $I_*(w_{uv})$ for each parameter via Eq. (5);
    Update the importance score $\overline{s}^{(t)}_*(w_{uv})$ for each parameter via Eq. (6) and (7);
  end for
  Obtain the importance distribution $\mathcal{I}_*(\mathcal{E})$ through aggregation.
end for
Obtain $\mathcal{E}_c$, $\mathcal{E}_{ax}$, and $\mathcal{E}_{rx}$ respectively using Eq. (10), (11), and (12).
for $t = 1, \ldots, T_{tr}$ **do**
  Update $\mathcal{E}_c$ by optimizing $\mathcal{L}_c$ based on Eq. (13);
  Update $\mathcal{E}_{ax}$ by optimizing $\mathcal{L}_{ax}$ based on Eq. (14);
  Update $\mathcal{E}_{rx}$ by optimizing $\mathcal{L}_{rx}$ based on Eq. (15);
end for

Output: The fine-tuned LLMs Θ'.

A.1 Parenting Algorithm Framework

The detailed implementation of Parenting algorithm can be found in Algorithm 1.

A.2 Notations

The notations in this paper are summarized in Table 4.

A.3 Dataset Details

- **SQuAD2.0** (Rajpurkar et al., 2018) is a reading comprehension dataset encompassing multiple general domains, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment

Notation	Definition
Θ	Initial LLMs
Θ'	LLMs fine-tuned using Parenting
e_i	A single parameter unit in the model
$\mathcal{E}$	The set of parameter units in the model, $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$
α	Internal parametric knowledge of Θ
q	Input natural language query
D	The set of retrieved documents for q, $D = \{d_1, d_2, \ldots, d_k\}$
d_{golden}	Evidence required to answer query q
y	Relevance label indicating whether the retrieved documents contain evidence
$\mathbb{C}$	External knowledge base
ans_{golden}	Correct answer derived from d_{golden}
S_a	SFT dataset promoting adherence
S_r	SFT dataset promoting robustness
S_c	Document extraction dataset
$\mathcal{L}_*, * \in \{a, r, c\}$	Cross-entropy loss on S_*
$\mathcal{L}_{ax}$	Loss for the adherence subspace
$\mathcal{L}_{rx}$	Loss for the robustness subspace
$\mathcal{L}_{cx}$	Loss for the entangled subspace
p_*^j	Expected activation probability of layer j
$\overline{I}_*/\overline{U}_*(w_{uv})$, $* \in \{a, r\}$	Sensitivity / Uncertainty score for parameter w_{uv} based on gradients
$\overline{s}_*(w_{uv})$	Final importance score for parameter w_{uv}
$\mathcal{I}_*(\mathcal{E})$	Distribution of importance scores
$\mathcal{Z}_*(e_i)$	Standardized importance score of the i-th parameter unit
$\mathcal{E}_c$	Entangled subspace
$\mathcal{E}_{ax}$	Adherence subspace
$\mathcal{E}_{rx}$	Robustness subspace
$\mathcal{E}_o$	Other subspaces

Table 4: Notations for Parenting

of text, or span, from the corresponding reading passage, or the question might be unanswerable. Following the methodology outlined in Section 4.1, we construct S_a and S_r using the SQuAD2.0 dataset, with each containing 6,000 entries, and each entry consisting of a single document. For S_a, the document serves as either evidence or as a resource that could help produce a fabricated answer conflicting with the model's internal knowledge. During the process of expanding the dataset, we use GPT-4 to generate fictional answers that deviate from the true answers. We then replace all occurrences of the true answers in the evidence documents with these fictional ones to effectively augment the dataset. This is because the essence of knowledge conflict lies in the conflict itself, rather than in correctness (Li et al., 2023a). For S_r, the document consists of manually annotated irrelevant noise. Following the methodology described in Section 4.4.1, we develop S_c, which comprises 12,000 data entries. Each entry includes four documents: one document relevant to the question's answer (both correct and fabricated), one noise document on the same topic, and two noise documents from entirely different topics.

Following (Li et al., 2023a), we create the evaluation set SQuAD2.0-Eval to simulate complex retrieval scenarios. We construct two types of contexts: conflicting contexts and irrelevant contexts. In line with the typical chunk-size used in RAG tasks (Shi et al., 2023), we set the number of documents in each context to four. For the conflicting context, we select one document related to the same topic and two documents from different topics, all based on semantic similarity (We utilize the widely adopted and advanced bge-reranker-large (Xiao et al., 2023) model.). We ensure that none of these documents can provide an answer to the question. These are then paired with an evidence document that directly conflicts with the internal knowledge of the LLMs. To prevent bias arising from the order of documents, we randomize the order of evidence within the context. For the irrelevant context, we differentiate the documents by difficulty level. The more challenging documents, manually annotated, include two topic-related but unanswerable documents; the less challenging are randomly selected, consisting of two topic-unrelated documents. These four documents are then shuffled to form the irrelevant context.

- **RGB** (Wang et al., 2024b) serves as a benchmark to evaluate LLMs' ability to effectively utilize retrieved information and withstand various retrieval-related flaws, covering both English and Chinese languages. RGB provides question-answer pairs annotated with counterfactual and noisy knowledge. Consistent with our approach in constructing SQuAD2.0-Eval, we build an evaluation dataset with more complex contexts based on the original corpus.
- **KNOT** (Liu et al., 2024) is a benchmark specifically designed for studying the resolution of knowledge conflicts. It categorizes questions into three types based on the reasoning capabilities required to reconcile conflicting information. Consistent with our methodology in constructing SQuAD2.0-Eval, we further enrich this foundation by incorporating noisy contexts, and creating an evaluation dataset with more complex scenarios to test performance.
- **CMB** (Wang et al., 2024b) is a medical open-source query dataset, which are designed for multi-task Q&A, encompass single and multiple-choice questions in the medical field (Jiang et al.,

2024b). The CMB dataset utilizes qualifying exams as a data source in the four clinical medicine specialties of physicians, nurses, medical technicians, and pharmacists, with a total of 269,359 questions. Given the extensive size of the CMB dataset, we randomly sample 4,000 questions for testing. The medical domain concerns human health and life safety (Yang et al., 2023; Xu et al., 2023a; Ma et al., 2023; Xu et al., 2023b, 2024b). Developing language models capable of understanding and applying medical knowledge is of great significance for advancing the intelligence of healthcare services. Consistent with our approach for constructing the SQuAD2.0-Eval, by utilizing medical triples and documents from the Huatuo-26M dataset (Li et al., 2023b) to build supplementary medical context, we create a dataset to evaluate the medical knowledge utilization of LLMs. Specifically, we construct conflicting knowledge and contexts with original questions and answers, as well as irrelevant contexts. We do not require LLMs to complete multiple-choice questions but to directly answer the questions.

A.4 Prompts used in Parenting

In this section, we provide a detailed display of all the prompts used.

- **Extract parameter knowledge** The following prompt extracts world knowledge acquired during the pretraining phase of LLMs.

Prompt A.4.1

This is a question about {*Topic*}. Please answer the question {*Question q*}. Please provide a direct answer without analysis. If you are unsure or do not know the answer, please respond with 'I don't know'.

- **Construct fabricated (counterfactual) answers** The following few-shot prompt is used to generate fabricated answers using GPT-4 that deviates from the realistic answer, which we require to be as plausible as possible.

Prompt A.4.2

Please generate speciously plausible but incorrect answer to the question. Provide only the false answers; do not reiterate the queries.
Question: What is the capital of France? Answer: Paris. Fake answer: Lyon.
Question: What is the highest mountain in the world? Answer: Mount Everest. Fake answer: Lhotse.
. . . 7 more examples . . .
Question: Who is the founder of Microsoft? Answer: Bill Gates. Fake answer: Steve Jobs.
Question: {*Question q*} Answer: {*Realistic Answer* ans_{golden}} Fake answer:

- **Guide LLMs to integrate knowledge** The following prompt guides LLMs to integrate both external retrieved knowledge and internal parameter-based knowledge, serving as the input framework for the training sets S_a and S_r, as well as for all evaluation sets.

Prompt A.4.3

[Instruction] As a knowledge-based QA expert, you will provide professional responses based on user's question, utilizing any supplemental knowledge provided to enhance the quality of your response. If the supplemental information is irrelevant to the question, rely on your own expertise to formulate an answer. If you are unsure about the answer, please respond with 'I don't know'.
[Supplemental Knowledge] {*Context D*}
[Question] {*Question q*}
[Answer]

- **Document Extraction** The following prompt guides LLMs to identify specific document types and accurately restate their content.

Prompt A.4.4

[Instruction] Based on the topic, question, and supplementary knowledge provided below, assess the relevance of the supplementary knowledge to the question and classify it.
Output the document that is relevant to the topic and supports answering the question. *or* [Output the document that is relevant to the topic but does not directly assist in answering the question.] *or* [Output the document that is unrelated to both the topic and the question.]
[Topic] {*Topic*}
[Supplemental Knowledge] {*Context* D}
[Question] {*Question* q}
[Answer]

- **Noise recognition** The following prompt directs LLMs to classify document types, serving as the input framework for the noise identification experiment described in Section 5.4.

Prompt A.4.5

[Instruction] Based on the question and document provided below, assess whether the document assists in answering the question. If the document contains the answer to the question, output the evidence; otherwise, output noise.
[Supplemental Document] {*Document* d_k}
[Question] {*Question* q}
[Evaluation Result]

A.5 Implementation Details

We employ two different parameter-level language model architectures: LLaMA2-7B-Chat (Touvron et al., 2023) and Qwen1.5-14B-Chat (Team, 2024). For both models, we utilize parameter-efficient fine-tuning techniques, specifically LoRA, to expedite the training process. We use PyTorch library to implement all the algorithms based on the open-source HuggingFace transformers (Wolf, 2019) codebase, on an Ubuntu server equipped with 4 NVIDIA A100 GPUs with 80GB memory. For Parenting, we set the hyperparameters α_1 and α_2 in Equation (6) to 0.85, and set δ_1 in Equations (13), (14), and (15) to 0.5.

During the probing phase (i.e., the key parameter mining stage described in Section 4.2), we use a learning rate of 1e-4, a batch size of 16, a cutoff length of 1024, and run the process for 1 epoch.

In the training phase, for LLaMA2-7B-Chat: Training was conducted with a learning rate of 1e-4, a batch size of 16, a cutoff length of 1024, and 3 epochs. Additionally, LoRA adjustments were implemented with settings including a rank of 8, alpha of 16, and a dropout rate of 0.05, specifically targeting the modules [o_proj,q_proj,k_proj,v_proj]. For Qwen1.5-14B-Chat: Training was conducted with a learning rate of 5e-5, a batch size of 8, a cutoff length of 1024, and 3 epochs. LoRA settings were rank = 8, alpha = 16, dropout = 0.05, targeting modules [o_proj,q_proj,k_proj,v_proj].

For testing, settings included temperature = 0.02, top_p = 0, top_k = 1, max new tokens = 512.

Additionally, we carefully tuned the hyperparameters of baselines according to the suggestions in the original paper, in order to achieve their optimal performance.

A.6 General Task Performance

Targeted optimization for RAG scenarios may unintentionally have side effects on the general capabilities of LLMs. To comprehensively evaluate the impact of Parenting, we conducted systematic assessments across six widely-used benchmarks, measuring its performance on various general tasks, including multitask language understanding, logical reasoning, and factuality. These benchmarks include ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al.), Winogrande (Sakaguchi et al., 2021), GSM8K (Cobbe et al., 2021), and TruthfulQA (Lin et al., 2022). Following the setup in Shi et al. (2024a), we also used the Eleuther AI LM Evaluation Harness [1] for evaluation. For the ARC, HellaSwag, MMLU, Winogrande, and GSM8K benchmarks, a 5-shot approach was applied, whereas a zero-shot setup was used for evaluating TruthfulQA. For evaluation metrics, we use acc_norm for ARC and HellaSwag; acc for Winogrande, MMLU, and TruthfulQA; and strict exact_match for GSM8K, aligning with the Open LLM Leaderboard.

The experimental results in Table 5 demonstrate that through key parameter mining and type-tailored tuning strategies, **Parenting can significantly enhance adherence and robustness with**

[1] https://github.com/EleutherAI/lm-evaluation-harness

Models		ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average
LLaMA2-7B-Chat	Original	51.84	77.55	47.99	46.32	71.63	22.16	52.92
	Parenting	51.95	77.57	45.96	46.35	71.66	21.33	52.47
Qwen1.5-14B-Chat	Original	60.39	82.43	64.49	53.14	78.38	68.27	67.85
	Parenting	60.42	82.03	64.58	53.10	78.44	68.12	67.78

Table 5: Results of general abilities of LLMs on six widely-used benchmarks.

α_1,α_2	SQuAD		RGB	
	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
0.15	65.57	40.12	78.00	42.50
0.35	66.49	42.31	79.50	43.50
0.55	67.62	43.28	78.00	44.50
0.85	69.24	44.85	79.50	45.50
0.95	67.01	42.51	78.50	43.00

Table 6: Performance comparisons of Parenting (using LLaMA2-7B-Chat as the backbone model) across SQuAD2.0-Eval and RGB with different α values configured.

δ_1	SQuAD		RGB	
	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
0.1	61.56	36.93	69.50	36.50
0.3	68.12	44.71	80.00	44.00
0.5	69.24	44.85	79.50	45.50
0.7	68.94	43.14	79.50	44.50
0.9	63.57	37.54	75.50	37.50

Table 7: Performance comparisons of Parenting (using LLaMA2-7B-Chat as the backbone model) across SQuAD2.0-Eval and RGB with different δ_1 values configured.

minimal impact on the general capabilities of LLMs. In some cases, Parenting even led to slight performance improvements. These findings provide fresh insights and inspiration for the broader applicability of Parenting in diverse scenarios.

A.7 Sensitivity Analysis for Hyperparameters

Our proposed Parenting framework primarily includes two key hyperparameters: the smoothing hyperparameter α, used in Equation (6) to calculate importance scores during backpropagation, and the reweighting hyperparameter δ_1, applied in Equations (13), (14), and (15) to balance the original tasks with the newly introduced tasks. We aim to explore how different settings of these hyperparameters affect the performance of Parenting on the SQuAD2.0-Eval and RGB datasets, with tests conducted using the LLaMA2-7B-Chat backbone model. As shown in Table 6, we observe that the optimal value for α is 0.85. If the α is set too low, both adherence and robustness experience a decline, indicating that the identification of important parameters is affected by random sampling and the complex dynamics of the training process. Additionally, the results in Table 7 indicate that δ_1 values within a normal range have minimal impact on performance. Conversely, overly small δ_1 values can diminish the effectiveness of tasks aimed at enhancing model adherence and robustness. On the other hand, excessively large δ_1 values may undermine the positive effects of our proposed document extraction task. Overall, the model maintains relatively stable performance under most conditions, indicating a low sensitivity to changes in hyperparameters.

Models	Match Rate
LLaMA2-7B-Chat	1.36%
Qwen1.5-14B-Chat	1.89%

Table 8: The match rate between LLMs' parameterized answers and conflicting answers after training.

A.8 Detection of Unnecessary Memorization

Introducing question-answer pairs and conflicting knowledge contexts (especially fictional knowledge and answers) into the training data presents a potential risk of the model developing unnecessary memory retention. We employ prompts that do not provide supplementary knowledge to extract the parameterized knowledge of LLMs fine-tuned with Parenting. The results in Table 8 demonstrate that the model retains almost none of the conflicting knowledge. This further confirms that **Parenting effectively guides and optimizes the behavior patterns of LLMs across various contexts, without embedding specific knowledge into the model.**

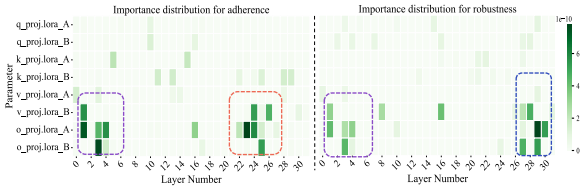

Figure 5: Visualization of parameter importance distributions $\mathcal{I}_a(\mathcal{E})$ and $\mathcal{I}_r(\mathcal{E})$ for adherence (**Left**) and robustness (**Right**) in Qwen1.5-14B-Chat.

A.9 Additional Visualization Analysis

In Figure 5, we visualize the distribution of parameter unit importance for adherence and robustness in Qwen1.5-14B-Chat. Our observations align well with those detailed in Section 5.3.3. Furthermore, in the Qwen1.5-14B-Chat model, we observe an increase in both entangled parameter units and robustness-specific units. This finding aligns with existing research (Jin et al., 2024a; Xie et al.), which indicates that as language models grow in size, their capability to leverage internal memories for problem-solving also improves. Such enhanced reasoning abilities further augment LLMs' capacity to perceive external contexts effectively.

A.10 More Diverse Evaluations

To further validate the generalizability of our approach to more complex real-world tasks, we perform additional experiments on the TempReason (Tan et al., 2023) and MuSiQue (Trivedi et al., 2022) dataset based on LLaMA2-7B-Chat.

- **TempReason** (Tan et al., 2023) contains three levels of time-related questions.It spans a time range from 634 to 2023 and includes a total of 52.8K question-answer pairs. We use this dataset to evaluate whether LLMs can select the most recent knowledge when confronted with conflicting answers from different timestamps. This is crucial for evaluating their performance in scenarios where knowledge bases are continuously updated.
- **MuSiQue** (Trivedi et al., 2022) is a challenging multi-hop question answering dataset designed to test the ability of LLMs to reason across all supporting facts in a given context to arrive at correct answers for complex questions. It contains 24,814 question-answer pairs, with questions involving 2 to 4 hops of reasoning.

The results in Table 9 indicate that our method **effectively supports real-world applications in dynamic scenarios and demonstrates strong generalization capability for more complex tasks.**

Method	TempReason	MuSiQue
Base	20.90	43.71
KAFT	28.70	58.72
Parenting	**31.40**	**61.35**

Table 9: Performance comparisons of Parenting with other RAG methods (all utilizing the same LLaMA2-7B-Chat backbone) on complex real-world tasks. All results are reported in accuracy (%).

Models	SQuAD		KNOT	
	R_{Ad}	R_{Ro}	R_{Ad}	R_{Ro}
Claude-3.5	68.29	29.76	63.85	28.49
GPT-4	69.16	28.21	64.24	27.56
Parenting	**80.47**	**47.58**	**73.27**	**43.34**

Table 10: Performance comparisons of Parenting (using Qwen1.5-14B-Chat as the backbone model) with closed-source LLMs Claude-3.5 Sonnet and GPT-4.

A.11 Comparison with Closed-Source LLMs

To further validate the effectiveness of our approach, we conduct experiments on the SQuAD2.0-Eval and KNOT datasets, using the closed-source LLMs Claude-3.5 Sonnet[2] and GPT-4 (Achiam et al., 2023) for comparison. The experimental results are shown in Table 10. Although these closed-source LLMs demonstrate strong performance, our proposed Parenting (based on Qwen 1.5-14B-Chat) still exhibits superior knowledge control and selection capabilities when faced with complex retrieval contexts.

A.12 Training Overhead and Computational Efficiency

From the perspective of network propagation, the proposed Parenting method mainly consists of two phases: key parameter mining and type-tailored tuning. We provide additional details regarding the workload involved in constructing the training dataset, as well as the training time and memory usage during the probing (key parameter mining) and training (type-tailored tuning) phases, and the number of training tokens.

From results in Table 11, we can observe that, under otherwise identical conditions, Parenting, despite introducing an additional probing stage compared to baseline KAFT, actually results in **shorter**

[2]https://www.anthropic.com/news/claude-3-5-sonnet

Method	Probing phase		Training phase		Number of tokens
	Time	Memory usage	Time	Memory usage	
KAFT	/	/	1.49h	68.3GB	32.7M
Parenting	**154s**	68.3GB	**1.03h**	68.3GB	33.5M

Table 11: Computational cost of Parenting and the baseline KAFT on LLaMA2-7B-Chat.

Question	**Which animal is the fastest on land?**
Documents	[Doc1] The cheetah is considered the fastest land animal, reaching speeds of up to 70 miles per hour over short distances. [Doc2–noise] The pronghorn antelope, native to North America, can reach speeds of 55 miles per hour, making it one of the fastest land animals.
Parenting Answer	Based on extra knowledge, the cheetah is the fastest land animal. **[CORRECT]**
KAFT Answer	Based on extra knowledge, the pronghorn antelope is the fastest land animal. **[WRONG]**

Table 12: Case study comparing the baseline KAFT and Parenting.

training times and faster network convergence. This suggests that our proposed strategy of "decoupling parameters first, then performing tailored tuning" effectively prevents conflicting supervisory signals from interfering with each other, thereby improving training efficiency and accelerating the process. Additionally, the total number of training tokens in our approach is comparable to that of KAFT, with no additional computational overhead. Furthermore, during the inference phase, the trained model performs with no difference in inference speed compared to the original base model.

A.13 Case Study

In this section, we provide a case study to visually demonstrate the effectiveness of Parenting. As shown in Table 12, the two external knowledge documents we provided are semantically highly similar to the question. However, the noise document does not contain a direct answer, as it only describes the ranking of antelope speed using the term "one of", without providing a precise answer.

The baseline KAFT method treats adherence and robustness as analogous instruction-following positive examples. This approach introduces contradictory signals during instruction-following training, leading to increased learning variance and impeding the effective acquisition of both capabilities. As a result, KAFT fails to comprehensively analyze the relevance between the context and the question at a fine-grained level, making it more susceptible to noise interference, ultimately leading to incorrect answers.

In contrast, Parenting effectively addresses the balance between adherence and robustness by decoupling and identifying parameter subspaces associated with each capability. By designing specialized tuning strategies for these distinct subspaces, Parenting ensures that both adherence and robustness are fully optimized. As a result, it can effectively leverage evidence to generate the correct answer.