SPARK: ADAPTIVE PARAMETRIC KNOWLEDGE PARTICIPATION FOR FAITHFUL RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) enhances the faithfulness of large language models (LLMs) by incorporating external knowledge into the generation process. However, conflicts between retrieved context and parametric knowledge can cause LLMs to rely on inconsistent internal knowledge, hindering their ability to faithfully leverage retrieved evidence. Existing approaches mitigate this issue by editing parametric knowledge to resolve such conflicts. However, locating and modifying knowledge encoded in model parameters is challenging and may inadvertently compromise the backbone LLM's capabilities. In this paper, we introduce SPARK, a novel framework that formulates parametric knowledge control as a latent participation policy learning problem, enabling LLMs to adaptively regulate the contribution of parametric knowledge during generation. Specifically, SPARK introduces a lightweight participation policy that dynamically modulates the contribution of FFN-mediated parametric knowledge while keeping the backbone LLM frozen. The participation policy is optimized with a faithfulness-oriented reward, enabling fine-grained control over parametric knowledge participation during generation. Extensive experiments on six datasets validate the effectiveness of SPARK, which introduces and optimizes only 0.01% additional trainable parameters while keeping the backbone LLM fully frozen, thereby preserving its general capabilities and achieving more faithful generation than competitive baselines. All datasets and code will be released via GitHub.

1 INTRODUCTION

Retrieval-augmented generation (RAG) has become a critical paradigm for improving the factual reliability of Large Language Models (LLMs) by grounding generation in external evidence (Guu et al., 2020; Lewis et al., 2020; Izacard & Grave, 2021; Li et al., 2025d; Zhang et al., 2025c). By incorporating retrieved context into the generation process, RAG enables LLMs to access up-to-date and domain-specific information beyond the static parametric knowledge encoded during pretraining (Izacard et al., 2023; Mallen et al., 2023). A RAG system thus relies on two knowledge sources during inference: *parametric knowledge*, which is encoded in model parameters, and *contextual knowledge*, which is supplied by retrieved documents. When both knowledge sources contradict each other, *knowledge conflict* arises (Xie et al., 2024), under which LLMs may ignore retrieved evidence (Zhang et al., 2025b), overly rely on parametric knowledge (Huang et al., 2025b), or inconsistently blend the two knowledge sources during generation (Choi et al., 2025). These failure modes compromise the factual reliability of RAG systems (Sun et al., 2025), making knowledge conflict mitigation a crucial problem for faithful retrieval-augmented generation.

A central challenge in mitigating knowledge conflict is how to regulate the influence of parametric knowledge when it contradicts retrieved evidence (Longpre et al., 2021; Chen et al., 2022). Prior approaches typically follow a *localize-then-intervene* paradigm: they first identify knowledge-related components and then apply predefined modifications, routing, or suppression strategies to regulate their contribution. However, knowledge localization does not necessarily identify the locations where intervention is most effective (Hase et al., 2023), while predefined intervention strategies may fail to accommodate the evolving generation process (Chen et al., 2025). These limitations motivate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

us to shift our focus from identifying *where* parametric knowledge should be intervened on toward understanding *how* it should participate in generation under knowledge conflict.

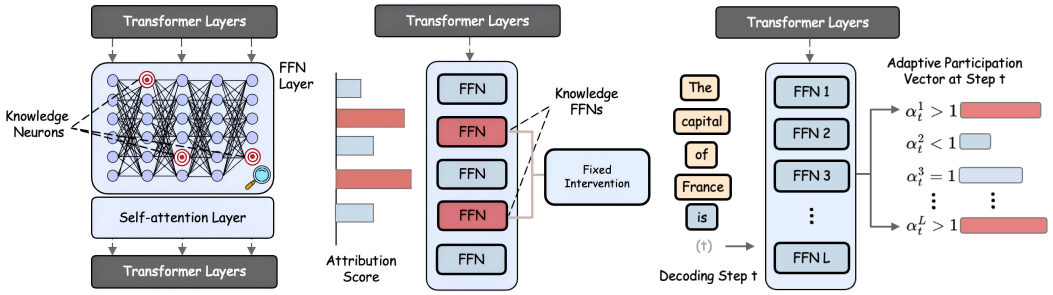
In this paper, we propose SPARK (**S**elective **P**articipation of p**AR**ametric **K**nowledge), which recasts knowledge-conflict mitigation as learning how parametric knowledge should participate during generation. Rather than relying on predefined intervention targets or fixed intervention rules, SPARK models the participation behavior of parametric knowledge during generation as a latent policy. Since such participation behavior is inherently unobservable and cannot be directly annotated, the policy cannot be learned from explicit participation supervision. We therefore keep the backbone LLM frozen and optimize only the participation policy with Group Relative Policy Optimization (GRPO) (Guo et al., 2025), using the faithfulness of generated responses as the learning signal. In this way, SPARK learns latent knowledge participation directly from generation-level feedback, without requiring explicit labels for how parametric knowledge should participate.

Our experiments demonstrate the effectiveness of SPARK in improving the faithfulness of LLM generation, with SPARK outperforming competitive baselines by up to 3.30% on standard RAG tasks. With fewer than 0.01% additional trainable parameters, SPARK effectively modulates LLMs to appropriately leverage contextual knowledge while preserving their general capabilities. Notably, SPARK delivers consistent improvements across diverse knowledge conflict scenarios, demonstrating its ability to learn tailored yet generalizable parametric knowledge participation policies for off-the-shelf LLMs. These policies enable LLMs to adaptively regulate the contribution of parametric knowledge across different scenarios, thereby improving generation faithfulness. Moreover, we investigate the internal activation patterns of LLMs and find that the learned participation policies reshape neuron activations in ways that facilitate knowledge conflict resolution.

2 Related Work

Existing work on faithful retrieval-augmented generation has extensively investigated how to mitigate knowledge conflicts when contextual evidence contradicts a model's parametric memory (Longpre et al., 2022; Wang et al., 2025b). One line of research improves faithfulness by regulating how models incorporate contextual evidence in the presence of such conflicts. For example, prompting-based methods encourage models to prioritize contextual evidence through carefully designed instructions (Zhou et al., 2023) or steer generation by contrasting predictions with and without contextual information (Shi et al., 2024b). Another line optimizes context-following behavior through retrieval-augmented fine-tuning or preference optimization, encouraging models to develop a stronger reliance on external evidence (Lin et al., 2024; Bi et al., 2025a). More recently, adaptive decoding methods have shown that the relative contribution of contextual and parametric knowledge can vary across generation states. For example, AdaCAD (Wang et al., 2025a) dynamically adjusts the strength of contrastive decoding according to the degree of knowledge conflict, while CoCoA (Khandelwal et al., 2025) performs token-level adaptation based on contextual and parametric prediction signals. Despite these advances, existing approaches primarily regulate how external evidence is incorporated or how the output distribution is adjusted, while largely treating the model's internal parametric knowledge as a fixed source. How to selectively control the participation of internal knowledge during generation, particularly at the level of the model's underlying knowledge storage, remains largely unexplored.

A complementary line of work investigates the internal mechanisms through which parametric knowledge is stored and expressed. To understand and enable targeted control over parametric knowledge, early studies attempted to localize factual knowledge at the neuron level (Meng et al., 2023; Dai et al., 2022b). However, accurately localizing such knowledge remains challenging (Geva et al., 2023; Chen et al., 2024), as the resulting interventions are often computationally costly and brittle (Hoelscher-Obermaier et al., 2023; Cohen et al., 2024; Niu et al., 2024). Consequently, recent studies have shifted toward coarser-grained internal structures, such as attention heads and FFN modules, which provide more structured interfaces for knowledge control. Building upon these mechanistic insights, several approaches directly intervene on internal components to mitigate knowledge conflicts. For example, PH3 (Jin et al., 2024), RHIO (Huang et al., 2025a), and JuICE (Li et al., 2025a) regulate generation through head-level pruning, contrastive decoding, or test-time intervention, while ROME (Meng et al., 2023) directly edits FFN-based factual associations. More recently, DoLa (Chuang et al., 2024) identifies intermediate layers and contrasts their



(a) Neuron-level Intervention (b) Layer-level Intervention (c) Our Method - **SPARK**

Figure 1: **Comparison of three paradigms for mitigating knowledge conflict.** (a) Neuron-level intervention requires fine-grained localization of knowledge-related neurons. (b) Layer-level intervention applies fixed rules to a manually selected subset of layers. (c) **SPARK** instead learns an input-dependent participation vector over FFN layers, thereby adapting how parametric knowledge contributes to generation.

logits with those from the final layer, without modifying model parameters. ParamMute (Huang et al., 2025b) identifies FFN modules associated with unfaithful generation and suppresses their contributions, while CoRect (Ma et al., 2026) dynamically identifies suppressive FFN layers and performs hidden-state rectification during inference. Although these methods show that internal components can serve as effective control interfaces, they generally follow a *localize-then-intervene* paradigm: they first identify components associated with undesirable parametric influence and then apply intervention strategies based on these identified components. Furthermore, accurately identifying such components poses significant difficulties, as localization results can vary depending on the input and localization criteria (Geva et al., 2023; Chen et al., 2024; Chang et al., 2024; Chen et al., 2025). More importantly, knowing *where* knowledge is encoded does not by itself determine *how much* the corresponding components should contribute during generation (Hase et al., 2023; Chen et al., 2025). In contrast, **SPARK** reframes knowledge control as a *participation* problem: rather than explicitly localizing and intervening on knowledge-bearing components, we learn a latent participation policy that dynamically determines how much each FFN contributes conditioned on the evolving generation context.

3 Preliminaries: From Knowledge Localization to Knowledge Participation in Faithful Generation

In this work, we study how parametric knowledge should adaptively be expressed in generation to improve faithfulness under knowledge conflict. Our key perspective is to move beyond *knowledge localization* toward *knowledge participation*: rather than merely identifying where knowledge is stored, we investigate whether and to what extent the corresponding components should participate in the generation process. In particular, we focus on feed-forward network (FFN) sublayers, which have been widely recognized as key components for storing and expressing parametric knowledge (Geva et al., 2021; Dai et al., 2022b). This makes FFNs a natural interface for controlling the contribution of parametric knowledge during generation. We first motivate adaptive FFN participation by examining the limitations of predefined FFN-level interventions, which typically apply fixed modification strategies regardless of the evolving generation context. We then formulate FFN participation as a latent policy that dynamically determines whether and how FFN-mediated knowledge should contribute to generation, with the policy learned from response-level faithfulness feedback.

3.1 Adaptive FFN Participation under Knowledge Conflict

In Figure 1, we compare different levels of intervention for improving the faithfulness of LLM generation under knowledge conflict.

First, *neuron-level intervention*, as illustrated in Figure 1(a), typically relies on attribution-based localization methods to identify knowledge neurons that are strongly associated with factual knowledge expression (Dai et al., 2022a). However, precisely localizing such neurons can be challenging

and brittle, as factual knowledge is distributed across a large number of neurons (Dai et al., 2022a), while the identified neuron sets can vary substantially depending on the attribution method and the input query (Chang et al., 2024; Chen et al., 2025), motivating operating at a coarser and more structured level of the model. Recent studies therefore shift toward *layer-level intervention* at the FFN-sublayer level, as shown in Figure 1(b). For example, existing approaches route between original and edited memories at selected FFNs or uniformly suppress a localized subset of knowledge-critical FFNs (Wang et al., 2024; Huang et al., 2025b). These works demonstrate that FFN-level intervention provides a more robust interface for regulating the contribution of parametric knowledge. However, existing FFN-level approaches still follow a *localize-then-intervene* paradigm: they first identify a predefined subset of knowledge-related FFNs and then apply fixed modification, routing, or suppression strategies to the selected components, regardless of the evolving generation context. Such predefined control not only restricts regulation to the localized subset, but may also impair the model's general capabilities by suppressing useful sublayers (Hsueh et al., 2024; Fu et al., 2025). These limitations motivate a shift from deciding *where to intervene* toward determining *when* and *to what extent* each FFN should participate according to the current token representation.

As illustrated in Figure 1(c), our perspective reformulates parametric knowledge control as an *adaptive participation problem*. Rather than committing to a predefined subset of localized FFNs, it considers all FFN sublayers as potential carriers of parametric knowledge within a functional subnetwork (Zhang et al., 2022). This perspective models each FFN sublayer to exhibit an input-dependent participation pattern according to the evolving generation context. Under this formulation, FFN participation can be represented as a latent policy over the degree to which each FFN contributes at each decoding step.

3.2 Learning FFN Participation Policies via Reinforcement Learning

The adaptive participation formulation introduces a latent policy learning problem because the optimal FFN participation coefficients are not directly observable, and reference responses do not provide explicit supervision for their values at individual FFN sublayers and token positions. Therefore, the participation policy must be learned indirectly through its impact on the model's generation behavior.

After applying the FFN participation policy ρ_ψ to the backbone LLM π_θ, we obtain the ρ_ψ-augmented model $\pi_{\theta,\psi}$. During training, the backbone LLM π_θ remains frozen, while only the FFN participation policy ρ_ψ is optimized. A straightforward way to learn this policy is through supervised fine-tuning (SFT):

$$\mathcal{L}_{\mathrm{SFT}}(\psi) = -\sum_{t=1}^{T} \log \pi_{\theta,\psi}\left(y_t^\star \mid x, c, y_{<t}^\star\right), \tag{1}$$

where $y^\star$ is a reference response and parameters θ of the backbone LLM π_θ remain frozen. This objective indirectly optimizes the latent participation variables by maximizing the token-level likelihood of reference responses (Ouyang et al., 2022). However, it learns through next-token prediction under teacher-forced reference prefixes, which creates an objective mismatch with the response-level faithfulness desired for FFN participation. Moreover, since every reference token serves as a supervised target, SFT is sensitive to the quality and diversity of the reference responses (Li et al., 2025c; Bukharin et al., 2024).

To better align policy learning with response-level faithfulness, we instead optimize the FFN participation policy ρ_ψ using response-level rewards. Given the frozen backbone π_θ, the participation policy ρ_ψ induces a generation model $\pi_{\theta,\psi}$, whose objective is

$$\max_{\psi} \ \mathbb{E}_{y\sim\pi_{\theta,\psi}(\cdot|x,c)}\left[R(y; x, c, y^*)\right], \tag{2}$$

where $R(y; x, c, y^*)$ provides response-level faithfulness feedback by evaluating the generated response y against the reference answer y^*. Unlike SFT, this objective evaluates complete responses sampled from the model and directly uses response-level rewards to optimize the latent participation variables (Shao et al., 2024). Reinforcement learning thus provides a natural way to learn latent FFN participation without requiring explicit supervision of participation labels (Shao et al., 2024).

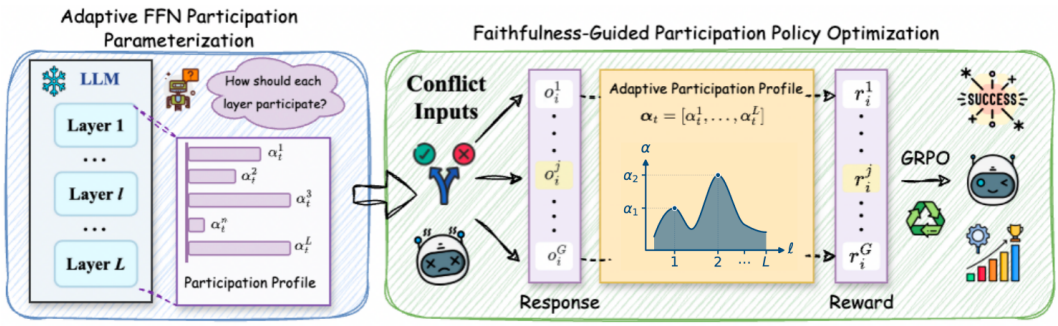


Figure 2: **Overview of the proposed framework SPARK.** The framework consists of two stages: (1) *Adaptive FFN Participation Parameterization*, which introduces a layer-wise participation vector to dynamically regulate the contribution of parametric knowledge during generation; and (2) *Faithfulness-Guided Participation Policy Optimization*, which learns the participation policy by optimizing response-level rewards that encourage faithful knowledge utilization.

4 METHODOLOGY

As illustrated in Figure 2, this section presents SPARK, a framework that learns adaptive knowledge participation policies for faithful generation. The key idea is to formulate parametric knowledge participation as a latent policy learning problem during generation. We introduce a knowledge participation policy ρ_ψ, which observes the current generation state and produces layer-wise participation actions that determine the extent to which FFN-mediated parametric knowledge contributes to the generation process. Specifically, SPARK first parameterizes the knowledge participation policy by introducing layer-wise participation modules into the LLM architecture (Section 4.1). Then, the knowledge participation policy ρ_ψ is optimized with faithfulness-oriented policy optimization, where generation-level rewards provide supervision for learning when and how parametric knowledge should participate during generation (Section 4.2).

4.1 PARAMETERIZING ADAPTIVE FFN PARTICIPATION IN OFF-THE-SHELF LLMS

Motivated by the observation that FFN sublayers serve as interfaces for expressing parametric knowledge, SPARK formulates knowledge participation as a learnable, input-dependent policy that determines how much FFN-mediated parametric knowledge should participate during generation. Instead of modifying the parameters (Huang et al., 2025b), we learn an adaptive participation policy that regulates the contribution of FFN-mediated parametric knowledge in off-the-shelf LLMs.

Specifically, the participation policy ρ_ψ is parameterized by layer-wise participation modules. For each FFN layer ℓ, where $\ell \in \{1, \ldots, L\}$ and L denotes the total number of FFN layers, the associated participation module has parameters $\psi_\ell = (\mathbf{w}_\ell, b_\ell)$. The complete policy parameters are denoted as $\psi = \{\psi_\ell\}_{\ell=1}^{L}$. Given the hidden representation $\mathbf{h}_{i,t}^{\ell} \in \mathbb{R}^d$ entering the FFN sublayer of layer ℓ at decoding step t for the i-th input instance, we treat $\mathbf{h}_{i,t}^{\ell}$ as the current generation state observed by the participation policy. Formally, the participation policy produces a layer-wise participation action:

$$\alpha_{i,t}^{\ell} = \lambda_\alpha \sigma\left(\mathbf{w}_\ell^\top \mathbf{h}_{i,t}^{\ell} + b_\ell\right), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function and $\alpha_{i,t}^{\ell}$ represents the participation action of the policy at layer ℓ and decoding step t. A standard sigmoid output is bounded in $(0, 1)$, allowing the participation policy only to attenuate FFN contributions. However, knowledge conflict does not necessarily indicate that parametric knowledge should be suppressed. When retrieved evidence is insufficient or unreliable, the model's parametric knowledge may instead provide useful complementary information, making stronger FFN contributions desirable. Thus, SPARK introduces a scaling factor λ_α that supports both attenuation and amplification of FFN contributions. This allows the participation policy to adaptively control the degree of parametric knowledge involvement during generation.

Accordingly, the participation action determines the extent to which each FFN sublayer contributes to the residual stream:

$$\mathbf{h}_{i,t}^{\ell+1} = \mathbf{h}_{i,t}^{\ell} + \alpha_{i,t}^{\ell}\,\mathrm{FFN}^{\ell}\left(\mathrm{Norm}(\mathbf{h}_{i,t}^{\ell})\right). \tag{4}$$

When $\alpha_{i,t}^{\ell} < 1$, the corresponding FFN sublayer contributes less to the residual stream, reducing the participation of its associated parametric knowledge. Conversely, when $\alpha_{i,t}^{\ell} > 1$, the FFN contribution is amplified, allowing internal knowledge to play a stronger role. Therefore, **SPARK** provides a continuous and input-dependent mechanism for adaptive FFN participation, rather than relying on predefined neuron selection or fixed layer-level intervention strategies.

To preserve the original generation behavior of the vanilla LLM before policy optimization, we initialize the parameters of the knowledge participation policy by setting $\mathbf{w}_{\ell} = 0$ and initializing the bias as:

$$b_{\ell} = \mathrm{logit}\left(\frac{1}{\lambda_{\alpha}}\right), \quad \forall \ell \in \{1, \ldots, L\}. \tag{5}$$

With this initialization, $\alpha_{i,t}^{\ell} = 1$ for all hidden states, leaving the original FFN computation unchanged. Therefore, the augmented model initially preserves the behavior of the frozen backbone LLM π_{θ}, while subsequent optimization learns the adaptive participation policy ρ_{ψ} through generation-level feedback (Section 4.2).

4.2 Faithfulness-Guided Optimization of the Knowledge Participation Policy

Although existing training data $\mathcal{D}_{\text{training}}$ provides factual correctness annotations for final responses, it lacks explicit supervision for the underlying knowledge participation decisions. In particular, the annotations do not specify when or to what extent parametric knowledge should participate during generation. Since these fine-grained participation behaviors are latent, we formulate adaptive FFN participation as a policy optimization problem and optimize the participation policy ρ_{ψ} using generation-level faithfulness feedback.

After incorporating the knowledge participation policy ρ_{ψ} into the frozen backbone LLM π_{θ}, we obtain the augmented generation model $\pi_{\theta,\psi}$. During optimization, we update only the parameters of the knowledge participation policy ψ, while keeping all backbone parameters θ of π_{θ} frozen. The overall optimization objective is formulated as:

$$\mathcal{L} = -\mathcal{J}_{\text{GRPO}} + \beta_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{regular}}\mathcal{R}_{\text{regular}}, \tag{6}$$

where $\mathcal{J}_{\text{GRPO}}$ is the GRPO objective (Guo et al., 2025) that is used to optimize the knowledge participation policy according to faithfulness-oriented rewards, $\mathcal{L}_{\text{ref}}$ constrains the optimized policy from deviating excessively from the original LLM behavior, and $\mathcal{R}_{\text{regular}}$ regularizes the parameters of the learned participation policy ρ_{ψ}. β_{ref} and λ_{regular} are hyperparameters.

Specifically, given an input instance $(x_i, c_i, y_i^*) \sim \mathcal{D}_{\text{training}}$, where x_i denotes the query, c_i denotes the retrieved context, and y_i^* denotes the reference response, the augmented model $\pi_{\theta,\psi}$ generates a group of G candidate responses:

$$\mathcal{O}_i = \{o_i^1, o_i^2, \ldots, o_i^G\}, \qquad o_i^j \sim \pi_{\theta,\psi}(\cdot \mid x_i, c_i). \tag{7}$$

Each generated response o_i^j is evaluated using a composite reward $R(o_i^j; x_i, c_i, y_i^*)$ comprising a format reward and a faithfulness reward, where the latter is computed by comparing the generated response with the reference answer y_i^*. More description is detailed in Appendix A.5. GRPO then computes the relative advantages among sampled responses and updates the participation policy parameters ψ to increase the probability of responses with higher faithfulness rewards.

To preserve the general generation capability of the original LLM, we adopt the frozen backbone model π_{θ} as the reference policy and introduce a KL-based regularization term:

$$\mathcal{L}_{\text{ref}} = \frac{1}{T}\sum_{t=1}^{T} D_{\text{KL}}\left(\pi_{\theta,\psi}(\cdot|x_i, c_i, o_{i,<t}) \| \pi_{\theta}(\cdot|x_i, c_i, o_{i,<t})\right), \tag{8}$$

where the KL divergence is computed at each decoding step over the token distribution, and T denotes the response length of the LLM response. This constraint encourages the learned knowledge participation policy ρ_{ψ} to improve faithfulness while preserving the behavior of the original LLM.

Table 1: **Performance comparison on the MRQA benchmark.** The best and second-best results are marked in **bold** and underlined, respectively. † and ‡ indicate statistically significant improvements over SFT† and Knowledgeable-R1‡ in F1 score

Methods	HotpotQA	SearchQA	NewsQA	NQ	TriviaQA	SQuAD	Avg.
Qwen-3-1.7B							
No-RAG (Roberts et al., 2020)	8.82	14.77	3.09	9.22	18.33	10.40	10.77
Vanilla-RAG (Lewis et al., 2020)	63.41	53.89	33.08	55.07	62.53	68.15	56.02
CtrlA (Liu et al., 2024)	24.39	30.12	12.92	22.49	34.54	23.78	24.71
CK-PLUG (Bi et al., 2025b)	57.14	58.67	51.31	58.44	**76.71**	76.42	63.12
IRCAN (Shi et al., 2024a)	52.62	49.76	23.19	40.42	50.91	53.64	45.09
SFT (Wei et al., 2021)	64.76	61.80	50.45	60.98	68.93	73.25	63.36
Knowledgeable-r1 (Lin et al., 2026)	59.75	61.86	41.90	59.11	73.05	**77.08**	62.13
SPARK (ours)	**69.30**†‡	**63.86**†‡	**52.91**†‡	**65.55**†‡	74.17†‡	75.99†	**66.96**†‡
Qwen-3-8B							
No-RAG (Roberts et al., 2020)	21.53	49.76	6.22	28.07	56.52	23.70	30.97
Vanilla-RAG (Lewis et al., 2020)	78.81	75.05	60.22	70.99	81.07	84.32	75.08
CtrlA (Liu et al., 2024)	47.28	61.01	43.52	52.55	66.41	59.05	54.97
CK-PLUG (Bi et al., 2025b)	67.90	69.26	55.59	65.56	78.09	80.39	69.47
IRCAN (Shi et al., 2024a)	78.71	74.35	58.34	68.73	81.19	83.52	74.14
SFT (Wei et al., 2021)	73.10	76.39	59.10	69.43	78.19	85.93	75.23
Knowledgeable-r1 (Lin et al., 2026)	78.84	75.71	60.46	70.66	81.78	84.77	75.37
SPARK (ours)	**79.08**†‡	76.27‡	**60.93**†‡	**71.57**†‡	**82.28**†‡	**85.50**‡	**75.94**†‡

Furthermore, since the participation policy is parameterized by layer-wise FFN participation parameters, we impose an L_2 regularization term:

$$\mathcal{R}_{\text{regular}} = \frac{1}{L}\sum_{\ell=1}^{L}\left(\|\mathbf{w}_\ell\|^2 + b_\ell^2\right). \tag{9}$$

This regularization prevents overly aggressive participation adjustments and encourages the learned policy to achieve faithful generation through adaptive FFN participation.

5 Experimental Methodology

This section first describes datasets, evaluation metrics, and baselines. Then, we introduce the implementation details of our experiments. More experimental details are shown in Appendix A.6.

Datasets. To provide a comprehensive evaluation, we conduct extensive experiments on three categories of publicly available datasets: ❶ *In-domain Evaluation:* We evaluate our method on the standard open-domain MRQA benchmark (Fisch et al., 2019), including HotpotQA (Yang et al., 2018), SearchQA (Dunn et al., 2017), NewsQA (Trischler et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and SQuAD (Rajpurkar et al., 2016). ❷ *Out-of-domain Evaluation:* We further evaluate our model on ConfiQA (Bi et al., 2024) to examine its generalization capabilities, including Question Answering (QA), Multi-hop Reasoning (MR), and Multi-Conflicts (MC). ❸ *General Capability Preservation:* We report performance on the MMLU benchmark (Hendrycks et al., 2021) to verify the general capability preservation of our SPARK. Details are provided in Appendix A.2.

Table 2: **Performance comparison on the ConfiQA benchmark.** The best and second-best results are marked in **bold** and underlined, respectively.

Methods	MC	MR	QA
Qwen-3-1.7B			
Vanilla-RAG (Lewis et al., 2020)	7.82	11.10	5.70
SFT (Wei et al., 2021)	7.37	8.85	9.92
Knowledgeable-r1 (Lin et al., 2026)	8.17	11.72	6.70
SPARK (ours)	**10.00**	**14.20**	**10.80**
Qwen-3-8B			
Vanilla-RAG (Lewis et al., 2020)	18.45	21.34	27.17
SFT (Wei et al., 2021)	7.80	11.70	12.60
Knowledgeable-r1 (Lin et al., 2026)	14.72	20.83	19.42
SPARK (ours)	**21.10**	**30.30**	**28.60**

Baselines. We evaluate our proposed SPARK against a range of competitive baselines, which can be categorized into three groups: ❶ *Prompting-based methods*, including Vanilla-RAG (Ram et al.,

2023), CtrlA (Liu et al., 2024), which performs adaptive retrieval by leveraging representation-level control signals. ❷ *Decoding-based methods*, including CK-PLUG (Bi et al., 2025b), which provides plug-and-play control over the model's reliance by modifying token probability distributions. ❸ *Fine-tuning methods*, including Supervised Fine-Tuning (SFT) (Wei et al., 2021) and Knowledgeable-R1 (Lin et al., 2026), which train the model to improve its behavior under contextual knowledge conflicts. Detailed descriptions are provided in Appendix A.4.

Evaluation Metrics. Following Longpre et al. (2021), we evaluate the MRQA benchmark with token-level F1 score. To further evaluate the generalization of SPARK, we also adopt the ConfiQA benchmark with Exact Match (EM), following Bi et al. (2025a).

Implementation Details. To ensure a fair comparison, we use Qwen-3-1.7B and Qwen-3-8B in non-thinking mode as the backbone models throughout our experiments. For SPARK, we set the hyperparameters λ_g to 2, which controls the modulation range in Eq. 3. We set both $\mathbf{w}_l$ and b_l to 0 in Eq. 9 for all layers. Additional details, including training data construction, hyperparameter settings, and training prompt templates, are provided in Appendix A.3, Appendix A.6, and Appendix A.13.

6 Experimental Analysis

In this section, we first show the overall performance of SPARK and conduct ablation studies to show the contribution of different components. We then analyze the behavior of SPARK to explore how it adaptively regulates the participation of parametric knowledge under different knowledge conflict scenarios and investigate the resulting neuron activation patterns within the LLMs. Finally, we present several case studies in Appendix A.12.

6.1 Overall Performance

Table 1 shows the performance of SPARK against the compared methods for the MRQA benchmark on Qwen3-1.7B and Qwen3-8B. Overall, SPARK consistently achieves state-of-the-art performance, outperforming the strongest baseline by up to 3.30%. Compared with No-RAG, Vanilla RAG significantly enhances LLM performance across all datasets, indicating that external knowledge effectively improves the accuracy of generated responses. The intervention methods (Liu et al., 2024; Bi et al., 2025b) such as CtrlA and CK-PLUG yield unstable gains, suggesting that predefined decoding interventions struggle to regulate parametric knowledge. While fine-tuning methods are more competitive on the MRQA benchmark, SFT and Knowledgeable-R1 attain the strongest baseline on Qwen3-1.7B and Qwen3-8B, respectively. Moreover, Table 2 presents the performance of SPARK on the ConfiQA benchmark to evaluate out-of-domain generalization. SPARK consistently outperforms all compared methods across the MC, MR, and QA subsets on both backbones. Overall, the consistent gains on both in-domain MRQA and out-of-domain ConfiQA indicate that SPARK learns an adaptive participation policy that regulates parametric knowledge according to the evolving generation context, rather than simply relying on either knowledge source.

6.2 Ablation Studies

Table 3 presents ablation studies on Qwen3-1.7B and Qwen3-8B to examine how different components contribute to learning an effective knowledge participation policy. Specifically, we compare the full model with the following variants: ❶ w/o L_2 Regulation, which removes the regularization imposed on the knowledge participation policy. ❷ w/o Faithful Reward, which removes the answer-faithfulness reward and retains only the format reward. ❸ w/o Format Reward, which removes the format reward and retains only the faithfulness reward. Overall, the results suggest that these components

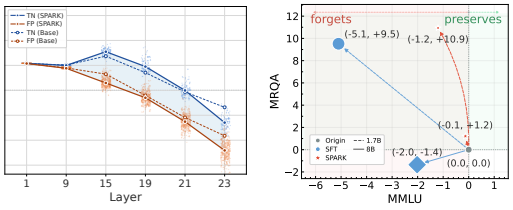
Table 3: **Ablation study on the MRQA benchmark** The best and second-best results are marked in **bold** and underlined, respectively.

Methods	SearchQA	NQ	TriviaQA	Avg.
Qwen-3-1.7B				
w/o L_2 Regulation	59.03	**67.33**	73.98	66.78
w/o Faithful Reward	54.22	66.36	71.85	64.14
w/o Format Reward	60.02	65.38	72.59	66.00
SPARK	**63.86**	65.55	**74.17**	**67.86**
Qwen-3-8B				
w/o L_2 Regulation	75.06	70.56	80.90	75.51
w/o Faithful Reward	75.00	70.85	81.20	75.68
w/o Format Reward	75.15	70.69	81.05	75.63
SPARK	**76.27**	**71.57**	**82.28**	**76.71**

play complementary roles rather than merely providing additive gains. The faithfulness reward is particularly important for Qwen3-1.7B, suggesting that smaller models rely more on explicit response-level supervision to learn effective knowledge participation. In contrast, L_2 regularization improves robustness across datasets by preventing overly aggressive participation adjustments, with a more pronounced effect on Qwen3-8B. The consistent degradation after removing the format reward further indicates that structural supervision complements faithfulness-oriented optimization.

6.3 Characteristics of Adaptive Knowledge Participation in SPARK

In this experiment, we explore whether SPARK could yield consistent improvements across diverse knowledge-conflict scenarios. To quantify this process, we first define two typical knowledge conflict scenarios: ❶ TN represents that the parametric answer is correct, but the context is wrong. ❷ FP represents that the parametric answer is wrong and the context is correct. We then evaluate our SPARK under each scenario and find that on both Qwen3-1.7B and Qwen3-8B, SPARK outperforms or matches the most competitive baseline across all three scenarios on average, detailed in Appendix A.9. To further investigate why SPARK benefits all scenarios, following the linear representation hypothesis (Park et al., 2024; Belrose et al., 2025), we project per-sample residual-stream activations onto the Fisher discriminant direction at each layer. As shown in Figure 3(a), compared with the base model, SPARK produces wider TN/FP separation from layer 15 onward, indicating that SPARK's participation policy strengthens the model's internal differentiation between scenarios where its parametric knowledge is reliable versus unreliable, providing a clearer representational basis for faithful generation. A complementary t-SNE visualization is described in Appendix A.10. In addition to the scenario-specific improvements above, SPARK also preserves general capabilities. To verify this, we further evaluate on MMLU, and Figure 3(b) shows that SPARK remains competitive with the original model while training less than 0.004% of the backbone parameters, whereas full parameter SFT drops up to 5.1.



(a) TN/FP scenario separation across layers of Qwen3-1.7B on NewsQA.

(b) Performance on the MMLU and MRQA benchmarks.

Figure 3: **Characteristics of adaptive knowledge participation.** Marker size in Figure(a) indicates the number of trainable parameters.

6.4 Analysis of Neuron Activation Patterns in SPARK under Knowledge Conflicts

In section 6.3, we have revealed that SPARK's participation policy amplifies scenario separation at the residual-stream level. Inspired by MOEfication (Zhang et al., 2022), which demonstrates that FFN layers can be decomposed into functionally specialized neuron groups, we further explore whether the backbone contains *neuron-level* structures that are sensitive to knowledge conflicts. We find that the backbone already contains scenario-sensitive neurons, but their distribution is highly uneven across layers. More importantly, SPARK's learned participation policy aligns well with this uneven distribution: layers with a higher concentration of conflict-sensitive neurons receive stronger participation adjustment, while layers with few such neurons remain nearly unchanged. This indicates that the RL-trained policy automatically learns to modulate each layer according to its relevance to knowledge conflicts, providing a neuron-level explanation for SPARK's adaptive behavior.

7 Conclusion

In this paper, we introduced SPARK, a lightweight framework that mitigates knowledge conflicts in retrieval-augmented generation through adaptive parametric knowledge participation. By learning per-layer parametric knowledge participation vector via reinforcement learning, SPARK achieves more faithful knowledge arbitration between retrieved context and parametric memory while preserving general capabilities.

Ethics Statement

Our research adheres to the ICLR Code of Ethics. All datasets used in this study are publicly available. The code and related materials will be appropriately released to ensure transparency and reproducibility of our work.

Reproducibility Statement

All information necessary to reproduce our results is documented in the Appendix, including dataset curation and preprocessing, training and inference hyperparameters, and so on.

AI Use Statement

We used a large language model (LLM) solely for writing assistance—specifically, to correct grammar, improve wording, and enhance clarity. The LLM did not contribute to research design, data analysis, modeling, experiments, or interpretation of results. All technical content and conclusions were authored and verified by the authors, who take full responsibility for the manuscript.

References

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-dpo: Aligning language models for context-faithfulness, 2024. URL https://arxiv.org/abs/2412.15280.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, 2025a.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models, 2025b. URL https://arxiv.org/abs/2503.15888.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.195. URL https://aclanthology.org/2024.findings-emnlp.195/.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. Do localization methods actually localize memorized data in LLMs? a tale of two benchmarks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3190–3211, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.176. URL https://aclanthology.org/2024.naacl-long.176/.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2307, 2022.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Chenzheng Zhu, Haofen Wang, Jeff Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. Learning to reason with search for llms via

reinforcement learning. *Advances in Neural Information Processing Systems*, 38:85287–85307, 2026.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17817–17825, 2024.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge localization: Mission not accomplished? enter query localization!, 2025. URL https://arxiv.org/abs/2405.14117.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. Conflict-aware soft prompting for retrieval-augmented generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26981–26995, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1371. URL https://aclanthology.org/2025.emnlp-main.1371/.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models, 2024. URL https://arxiv.org/abs/2309.03883.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

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

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen (eds.), *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5801. URL https://aclanthology.org/D19-5801/.

Zichuan Fu, Xian Wu, Guojing Li, Yingying Zhang, Yefeng Zheng, Tianshi Ming, Yejing Wang, Wanyu Wang, and Xiangyu Zhao. Model merging for knowledge editing. In Georg Rehm and Yunyao Li (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 433–443, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.30. URL https://aclanthology.org/2025.acl-industry.30/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL https://arxiv.org/abs/2012.14913.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali

(eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models, 2023. URL https://arxiv.org/abs/2301.04213.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11548–11559, 2023.

Cheng-Hsun Hsueh, Paul Kuo-Ming Huang, Tzu-Han Lin, Che Wei Liao, Hung-Chieh Fang, Chao-Wei Huang, and Yun-Nung Chen. Editing the mind of giants: An in-depth exploration of pitfalls of knowledge editing in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9417–9429, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.550. URL https://aclanthology.org/2024.findings-emnlp.550/.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yangfan Ye, Weihong Zhong, Yuxuan Gu, Baoxin Wang, Dayong Wu, et al. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16896–16913, 2025a.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Haiyan Zhao, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and Chenyan Xiong. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation, 2025b. URL https://arxiv.org/abs/2502.15543.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880, 2021.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pp. 737–740. ACM, 2025. doi: 10.1145/3701716.3715313. URL https://doi.org/10.1145/3701716.3715313.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 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*, pp. 1193–1215, 2024.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. CoCoA: Confidence- and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6835–6855, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.348. URL https://aclanthology.org/2025.emnlp-main.348/.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pp. 1137–1145. Montreal, Canada, 1995.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

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

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. *arXiv preprint arXiv:2503.10996*, 2025a.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62, 2025b.

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, Maosong Sun, and Chenyan Xiong. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards, 2025c. URL https://arxiv.org/abs/2410.13509.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms, 2025d. URL https://arxiv.org/abs/2507.09477.

Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lyu. Resisting contextual interference in rag via parametric-knowledge reinforcement, 2026. URL https://arxiv.org/abs/2506.05154.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. RA-DIT: retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024. URL https://openreview.net/forum?id=22OTbutug9.

Huanshuo Liu, Hao Zhang, Zhijiang Guo, Jing Wang, Kuicai Dong, Xiangyang Li, Yi Quan Lee, Cong Zhang, and Yong Liu. Ctrla: Adaptive retrieval-augmented generation via inherent control, 2024. URL https://arxiv.org/abs/2405.18727.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering, 2022. URL https://arxiv.org/abs/2109.05052.

Xuhua Ma, Richong Zhang, and Zhijie Nie. Corect: Context-aware logit contrast for hidden state rectification to resolve knowledge conflicts, 2026. URL https://arxiv.org/abs/2602.08221.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Marcus J Min, Yangruibo Ding, Luca Buratti, Saurabh Pujar, Gail Kaiser, Suman Jana, and Baishakhi Ray. Beyond accuracy: Evaluating self-consistency of code large language models with identitychain. *arXiv preprint arXiv:2310.14053*, 2023.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. What does the knowledge neuron thesis have to do with knowledge?, 2024. URL https://arxiv.org/abs/2405.02421.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. Toolrl: Reward is all tool learning needs, 2025. URL https://arxiv.org/abs/2504.13958.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models, 2023. URL https://arxiv.org/abs/2302.00083.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 5418–5426, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons, 2024a. URL https://arxiv.org/abs/2406.18406.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 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)*, pp. 783–791, 2024b.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *International Conference on Learning Representations*, volume 2025, pp. 50250–50279, 2025.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih (eds.), *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2623. URL https://aclanthology.org/W17-2623/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge, 2025a. URL https://arxiv.org/abs/2409.07394.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models, 2024. URL https://arxiv.org/abs/2405.14768.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yilin Wang, Heng Wang, Yuyang Bai, and Minnan Luo. Continuously steering llms sensitivity to contextual knowledge with proxy models, 2025b. URL https://arxiv.org/abs/2508.19720.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947. ISSN 00063444. URL http://www.jstor.org/stable/2332510.

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 *International Conference on Learning Representations*, volume 2024, pp. 35623–35646, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Haozhen Zhang, Tao Feng, and Jiaxuan You. Router-r1: Teaching llms multi-round routing and aggregation via reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation, 2025b. URL https://arxiv.org/abs/2506.08938.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, Ming Zhang, Yangqiu Song, Irwin King, and Philip S. Yu. From web search towards agentic deep research: Incentivizing search with reasoning agents, 2025c. URL https://arxiv.org/abs/2506.18959.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Moefication: Transformer feed-forward layers are mixtures of experts, 2022. URL https://arxiv.org/abs/2110.01786.

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

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 44(1):1–54, 2025.

A Appendix

A.1 License

We are committed to ethical research practices and ensuring the reproducibility of our work. To this end, we present the licensing information for all datasets utilized in our experiments. These include Natural Questions (CC BY-SA 3.0), NewsQA (MIT License), SearchQA (Apache License 2.0), TriviaQA (Apache License 2.0), HotpotQA (CC BY-SA 4.0), and SQuAD (CC BY-SA 4.0). Our use of these datasets is in full compliance with their terms, as all aforementioned licenses expressly permit their application in academic research.

A.2 More Details about the Datasets in Evaluation

In this section, we provide more details about the datasets evaluated in this paper as follows, with dataset statistics presented in Table 4.

A.2.1 In-Domain Datasets

- **HotPotQA** (Yang et al., 2018) is a multi-hop question-answering dataset constructed from Wikipedia. It requires models to reason over multiple supporting documents and identify evidence chains to answer complex questions.
- **NQ** (Kwiatkowski et al., 2019) is a large-scale open-domain question-answering dataset based on real queries issued to the Google search engine. The answers are annotated as short spans or long passages from Wikipedia pages.
- **NewsQA** (Trischler et al., 2017) is a machine reading comprehension dataset built from CNN news articles. Its questions are written by crowdworkers and require models to understand news passages and extract answers from the given articles.
- **SearchQA** (Dunn et al., 2017) is an open-domain question-answering dataset constructed from Jeopardy! questions and search engine snippets. It requires models to infer answers from noisy retrieved evidence rather than relying on a single clean reference passage.
- **SQuAD** (Rajpurkar et al., 2016) is a reading comprehension dataset built from Wikipedia articles. Each question is paired with a context paragraph, and the answer is typically annotated as a text span within the paragraph.
- **TriviaQA** (Joshi et al., 2017) is a large-scale question-answering dataset containing trivia questions paired with independently collected evidence documents. It is designed to evaluate whether models can answer challenging questions by reasoning over distant and potentially noisy evidence.

A.2.2 Out-of-Domain Datasets

- **ConfiQA** (Bi et al., 2024) is a benchmark designed to evaluate faithfulness in retrieval-augmented generation scenarios with knowledge conflicts. It contains questions paired with provided contexts, where the model is expected to generate answers that are faithful to the given context rather than relying on its parametric knowledge.

A.2.3 General Datasets

- **MMLU** (Hendrycks et al., 2021) is a multitask language understanding benchmark covering 57 subjects, including mathematics, history, computer science, law, and other academic or professional domains. It is designed to evaluate models' broad world knowledge and problem-solving ability through multiple-choice questions.

A.3 More Details about the Training Data Construction

In this section, we provide details of the datasets used for training and validation. Since the MRQA benchmark does not provide a standard validation split, we first construct a contrastive dataset from the standard MRQA training sets (Fisch et al., 2019). We then randomly split the constructed dataset

Table 4: **Detailed statistics of datasets used in evaluation.** The first six test datasets are directly drawn from the MRQA Development Data. For ConFiQA, the reported number includes its Question Answering (QA), Multi-hop Reasoning (MR), and Multi-Conflicts (MC) subsets.

Dataset	Category	Task Type	Test Samples
HotPotQA (Yang et al., 2018)	In-domain	Multi-hop QA	5,901
NQ (Kwiatkowski et al., 2019)	In-domain	Open-domain QA	12,836
NewsQA (Trischler et al., 2017)	In-domain	Reading Comprehension	4,212
SearchQA (Dunn et al., 2017)	In-domain	Open-domain QA	16,980
SQuAD (Rajpurkar et al., 2016)	In-domain	Reading Comprehension	10,507
TriviaQA (Joshi et al., 2017)	In-domain	Open-domain QA	7,785
ConFiQA (Bi et al., 2024)	Out-of-domain	Conflict QA	18,000
MMLU (Hendrycks et al., 2021)	General	Multi-task QA	14,042

into training and validation sets, with the corresponding dataset statistics presented in Table 5. To enable LLMs to learn the adaptive knowledge participation policy through `SPARK`, we employ self-consistency to estimate the model's parametric belief and generate two complementary types of examples. Specifically, the context-following type refers to cases where the model's parametric knowledge is incorrect, and the gold context is provided, requiring the model to rely on external evidence. In contrast, the context-resisting type refers to cases where the model's parametric knowledge is correct, but a conflicting context is provided, requiring the model to preserve reliable parametric knowledge against misleading evidence. The specific construction procedure is detailed as follows.

Self-Consistency Filter. For each question q_i, we first estimate the model's dominant parametric belief without providing any retrieved context. Following the self-consistency strategy (Wang et al., 2022; Min et al., 2023), we prompt the model 5 times using only the question and collect its generated answers. If a majority answer appears with a frequency of at least 3, we treat it as the model's dominant parametric answer and denote it as $\hat{r}_i^p$. Questions without a stable majority answer are filtered out to ensure that the constructed conflict examples are based on a reliable estimate of the model's parametric belief (Huang et al., 2025b).

Retrieval of Candidate Contexts. Following previous work (Li et al., 2025b; Chen et al., 2026; Zhang et al., 2025a; Zhu et al., 2025), we adopt the standard retrieval pipeline implemented in FlashRAG (Jin et al., 2025) to retrieve external passages for each question. The retrieved passages serve as candidate contexts for constructing both supportive and conflicting evidence. We denote the original supporting context associated with the gold answer as c_i^+. To construct a counterfactual context, we select a passage from the retrieved candidates that is semantically similar to the query but does not contain the dominant parametric answer $\hat{r}_i^p$. The selected passage is denoted as c_i^-. This construction preserves topical relevance while introducing contextual evidence that conflicts with the model's parametric belief.

Construction of Contrastive Training Data. Based on the relationship between the dominant parametric answer $\hat{r}_i^p$ and the gold answer r_i^c, we construct two types of conflict examples.

- **Parametric Type.** If the dominant parametric answer $\hat{r}_i^p$ matches the gold answer r_i^c, the model's parametric knowledge is considered reliable. We pair the question q_i with the counterfactual context c_i^-, which introduces misleading contextual evidence. The target answer is set to the parametric answer $\hat{r}_i^p$. This type of example requires the model to preserve correct parametric knowledge and avoid being misled by conflicting retrieved context.
- **Contextual Type.** If the dominant parametric answer $\hat{r}_i^p$ does not match the gold answer r_i^c, the model's parametric belief is considered unreliable. We pair the question q_i with the supporting context c_i^+. The target answer is set to the dataset gold answer r_i^c. This type of example requires the model to override its incorrect parametric belief and ground its response in the provided context.

Table 5: **Detailed statistics of the training and validation sets for different models.** Each split contains context-following and context-resisting samples.

Model	Split	Context-Following	Context-Resisting	Total
Qwen3-1.7B	Train	5021 (53.38%)	4386 (46.62%)	9407
	Valid	1255 (53.38%)	1096 (46.62%)	2351
Qwen3-8B	Train	7216 (50.59%)	7049 (49.41%)	14265
	Valid	1803 (50.58%)	1762 (49.42%)	3565

Training format. Each constructed example is converted into the training format used by **SPARK**. For a constructed context c_i and target answer a_i, we build the input prompt as

$$x_i = \mathcal{T}(q_i, c_i), \tag{10}$$

where $\mathcal{T}(\cdot)$ denotes the same prompt template used in the main experiments. For Parametric Type examples, we set $c_i = c_i^-$ and $a_i = \hat{r}_i^p$. For Contextual Type examples, we set $c_i = c_i^+$ and $a_i = r_i^c$. The final training set is the union of the two constructed subsets:

$$\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_c, \tag{11}$$

where each instance is represented as (x_i, a_i). This training data encourages the gate modules to learn both directions of knowledge arbitration: relying on parametric knowledge when it is correct and relying on contextual evidence when the parametric belief is incorrect.

A.4 More Details about the Baselines

In this section, we report the details about the baselines used for comparison in this paper.

- **Vanilla-RAG** (Ram et al., 2023) is a standard retrieval-augmented generation baseline that directly prepends retrieved documents to the input prompt and generates answers conditioned on the augmented context.
- **CtrlA** (Liu et al., 2024) is a prompting-based method that adaptively controls the model's reliance on retrieved information during generation. It aims to improve answer reliability by guiding the model to determine how much external context should be used.
- **CK-PLUG** (Bi et al., 2025b) is a decoding-based method for handling conflicts between contextual and parametric knowledge. It introduces fine-grained control during decoding to better integrate retrieved evidence with the model's internal knowledge.
- **IRCAN** (Shi et al., 2024a) identifies context-aware neurons through attribution analysis and reweights them to enhance the model's sensitivity to contextual knowledge.
- **SFT** (Wei et al., 2022) is a supervised fine-tuning baseline that updates model parameters using labeled training examples. It trains the model to generate target answers directly from the given query-context pairs.
- **Knowledgeable-R1** (Lin et al., 2026) is a fine-tuning-based method designed to improve robustness against contextual interference in retrieval-augmented generation. It trains the model to better resist misleading retrieved contexts while preserving useful knowledge for answering questions. We utilize LoRA-based parameter-efficient fine-tuning and adhere to its standard training recipe to implement this baseline.

A.5 More Details about SPARK

In this section, we provide additional details about the reinforcement learning procedure of **SPARK**, including the reward design and response parsing rule.

Reward design. Following recent work on rule-based rewards for RL-based LLM training (Guo et al., 2025; Zhang et al., 2025a), we employ a composite reward $R = R_{\text{format}} + R_{\text{faithful}}$ with two components.

Algorithm 1: `SPARK` Framework

Input : Frozen backbone LLM π_θ; training dataset $\mathcal{D} = \{(x_i, a_i)\}_{i=1}^{N}$; maximum number of iterations N_{iter}; group size G; learning rate η.

// Stage 1: Adaptive FFN Participation Parameterization;

Initialize the layer-wise participation parameters $\psi = \{(\mathbf{w}_\ell, b_\ell)\}_{\ell=1}^{L}$ with $\mathbf{w}_\ell = \mathbf{0}$, $b_\ell = \text{logit}(1/\lambda_\alpha)$ so that $\alpha_\ell = 1$ initially (Eq. 5);

Insert the participation modules into the frozen backbone to obtain the augmented policy $\pi_{\theta,\psi}$ using Eqs. 3 and 4;

Keep the backbone parameters θ frozen throughout training;

// Stage 2: Faithfulness-Guided Participation Policy Optimization;

for $k \leftarrow 1$ **to** N_{iter} **do**

  Sample a mini-batch $\mathcal{B} \subset \mathcal{D}$;

  foreach $(x_i, a_i) \in \mathcal{B}$ **do**

    Generate a group of G candidate responses $\mathcal{O}_i = \{o_i^j\}_{j=1}^{G} \sim \pi_{\theta,\psi}(\cdot|x_i, c_i)$ (Eq. 7);

    Compute the composite reward $r_i^j \leftarrow R(o_i^j, a_i)$ for each response $o_i^j \in \mathcal{O}_i$;

    Estimate group-relative advantages from $\{r_i^j\}_{j=1}^{G}$;

  Compute the KL regularization $\mathcal{L}_{\text{ref}}$ (Eq. 8);

  Compute the L2 regularization $\mathcal{R}_{\text{regular}}$ (Eq. 9);

  Compute the overall objective $\mathcal{L} = -\mathcal{J}_{\text{GRPO}} + \beta_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{regular}}\mathcal{R}_{\text{regular}}$;

  Update only the participation parameters: $\psi \leftarrow \text{Adam}\big(\psi, \nabla_\psi \mathcal{L}, \eta\big)$;

Output: Optimized participation policy parameters ψ^*.

To encourage adherence to the required structured output, we first define a **format reward**:

$$R_{\text{format}} = \begin{cases} 1.0, & \text{valid format with answer,} \\ 0.5, & \text{valid format without answer,} \\ 0.0, & \text{otherwise.} \end{cases} \tag{12}$$

Inspired by (Guo et al., 2025; Qian et al., 2025), we assign an intermediate score of 0.5 to well-structured but incomplete responses, which provides partial credit and yields a denser learning signal than a purely binary alternative.

We then define a **faithfulness reward** to encourage the final answer to be consistent with the context-supported target answer. Let $\hat{a}_i^j$ denote the answer extracted from the answer field of the j-th sampled response for input x_i, and let $\mathcal{A}_i$ denote the set of acceptable target answers. The faithfulness reward is computed as

$$R_{\text{faithful}} = \mathbf{1}\Big[\max_{a \in \mathcal{A}_i} \text{EM}\big(\text{norm}(\hat{a}_i^j), \text{norm}(a)\big) = 1\Big], \tag{13}$$

where $\text{norm}(\cdot)$ denotes the normalization function and $\text{EM}(\cdot,\cdot)$ denotes exact match after normalization. For datasets with a single reference answer, we set $\mathcal{A}_i = \{a_i\}$. If the response does not contain a valid answer field, we set $R_{\text{faithful}} = 0$.

Response parsing and normalization. We parse each generated response according to the required structured output template. When a valid answer field is detected, its content is extracted as $\hat{a}_i^j$ and used for reward computation. Otherwise, the extracted answer is treated as empty and the response receives no faithfulness reward. During reward computation, we apply the same normalization function $\text{norm}(\cdot)$ to both the extracted answer and the reference answers. Specifically, we lowercase the text, remove punctuation and articles, and collapse consecutive whitespace. This normalization prevents the reward from penalizing superficial surface-form differences while preserving the criterion that the generated answer should match the context-supported target.

Table 6: Hyperparameter settings used for training in SPARK.

Hyperparameter	Value
Training iterations	150
Episodes per iteration	64
Per-device batch size	4
Optimizer	AdamW
Learning rate	1×10^{-4}
Generations per sample G	4
KL coefficient	0.001
Sampling temperature	0.6
Top-p	0.95
Top-k	20
Gate reg. coefficient	0.01
Trainable parameters	Gates only

Table 7: **Performance comparison across three datasets.** Qwen models are evaluated in non-thinking mode, and Llama models are instruction-tuned variants. The **boldfaced** scores represent the **best** results.

Model	Method	Train. Params	NQ			TriviaQA			SQuAD			Avg.		
			EM	ACC	F1	EM	ACC	F1	EM	ACC	F1	EM	ACC	F1
Qwen Models														
Qwen-3-0.6B	Vanilla-RAG	N/A	40.63	51.34	53.56	46.72	57.60	54.11	46.91	63.32	59.30	44.75	57.42	55.66
	SPARK	28.7K (0.0048%)	**47.20**	**58.19**	**62.83**	**60.26**	**71.91**	**68.40**	**58.50**	**72.92**	**71.63**	**55.32**	**67.67**	**67.62**
Llama Models														
Llama-3.2-1B	Vanilla-RAG	N/A	10.95	14.97	14.90	24.37	32.23	28.90	16.76	30.25	23.80	17.36	25.82	22.53
	SPARK	32.8K (0.0027%)	**34.96**	**46.14**	**48.74**	**50.84**	**65.93**	**59.08**	**38.75**	**69.54**	**54.94**	**41.52**	**60.54**	**54.25**
Llama-3.2-3B	Vanilla-RAG	N/A	53.43	**66.94**	69.41	60.42	77.15	69.25	70.67	**86.86**	83.22	61.50	76.98	73.96
	SPARK	86.0K (0.0027%)	**56.80**	65.07	**71.95**	**71.83**	**81.91**	**78.98**	**76.04**	85.75	**86.18**	**68.22**	**77.58**	**79.04**
Llama-3.1-8B	Vanilla-RAG	N/A	52.23	67.60	68.78	64.75	81.03	73.83	67.95	87.71	81.92	61.64	78.78	74.84
	SPARK	131.1K (0.0016%)	**53.54**	**68.18**	**70.59**	**68.25**	**85.48**	**78.11**	**69.64**	**89.30**	**84.33**	**63.81**	**80.99**	**77.68**

A.6 HYPERPARAMETERS

In this section, we provide the hyperparameter settings used in the training stage of SPARK to better ensure reproducibility, which are detailed in Table 6. These training settings are the same for all the models used in our paper.

A.7 PSEUDOCODE OF PROPOSED SPARK

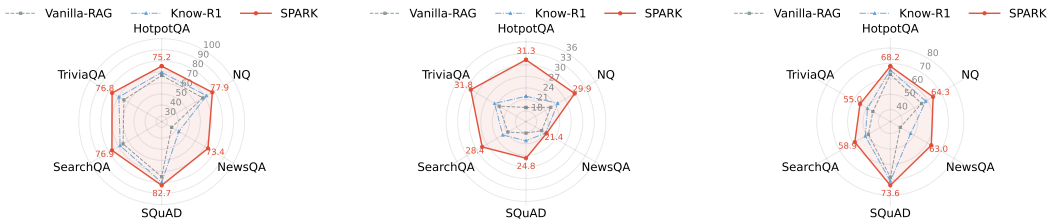
In this section, we present the step-by-step training algorithm of our SPARK in Algorithm 1, which is self-explanatory.

A.8 EXTENDING SPARK TO MORE LLMS

In this section, we provide more experimental results of our proposed SPARK on more LLMs. These LLMs include Qwen3-0.6B, Llama3.2-1B, Llama3.2-3B, Llama3.1-8B. To comprehensively evaluate the performance, we adopt three metrics, including Exact Match (EM), token-level F1 score (F1), and Accuracy (Acc). As shown in Table 7, SPARK observes a consistent improvement across different families and sizes of LLMs, demonstrating the versatility and scalability of SPARK across a wide spectrum of model architectures and sizes. Furthermore, we extend the MMLU evaluation to these LLMs to examine whether SPARK consistently maintains the general capabilities of different model families and sizes reported in Table 8.

Table 8: **Performance comparison on the MMLU benchmark under the 5-shot setting.** Hum. and Soc. denote Humanities and Social Sciences, respectively.

Size	Method	MMLU				
		Hum.	Soc.	STEM	Other	Avg.
Qwen Models						
0.6B	Origin	40.98	54.31	45.64	51.14	47.19
	SPARK	41.38	54.76	44.72	50.72	47.13
1.7B	Origin	52.77	68.54	59.40	64.18	60.24
	SPARK	51.80	67.92	57.72	62.54	59.04
8B	Origin	66.18	83.82	75.42	78.24	74.79
	SPARK	66.06	83.62	75.39	78.08	74.66
Llama Models						
1B	Origin	42.85	51.06	39.84	52.14	46.03
	SPARK	42.98	51.58	40.34	52.08	46.28
3B	Origin	57.45	68.87	51.25	66.59	60.58
	SPARK	57.39	69.52	50.78	66.85	60.65
8B	Origin	64.29	77.25	60.20	73.29	68.20
	SPARK	64.57	77.19	59.91	73.25	68.21



(a) Visualization of the participation vector on Llama-3.2-1B.

(b) Visualization of the participation vector on Llama-3.2-3B.

(c) Visualization of the participation vector on Llama-3.1-8B.

Figure 4: T-SNE visualization of the participation vectors of different LLMs.

A.9 More Details about the Scenario-specific Evaluation on MRQA Benchmark

In this section, we provide a more specific description of scenario-specific evaluation on the MRQA benchmark. First, we describe the three knowledge conflict scenarios used for evaluation and report the data statistics of each scenario. Then, we provide more experimental results using different LLMs.

Data Statistics. To comprehensively evaluate the performance of SPARK, we define three typical knowledge conflict scenarios along two dimensions, including whether the model's parametric knowledge is correct and whether the context is right: ❶ TP represents that both the parametric answer and the context agree on the correct answer. ❷ TN represents that the parametric answer is correct, but the context is wrong. ❸ FP represents that the parametric answer is wrong and the context provides the correct answer. Then, we utilize the same procedure described in Appendix A.3 to elicit the parametric answer of LLMs and to construct the conflict contexts. The corresponding data statistics for Qwen3-1.7B and Qwen3-8B are shown in Table 9. We utilize the first two subsets for scenario-specific evaluation.

Evaluation Results on more LLMs. For a more comprehensive evaluation of SPARK, we evaluate the performance of Qwen3-1.7B illustrated in Figure 4 and Qwen3-8B on three different scenarios shown in Table 10.

Table 9: **Detailed statistics of the MRQA scenario-specific split.** Right-answer indicates that the LLMs could answer the question correctly. Wrong-answer indicates that the LLMs could answer the question, but the answer is wrong. No-answer indicates that the LLMs could not answer the question.

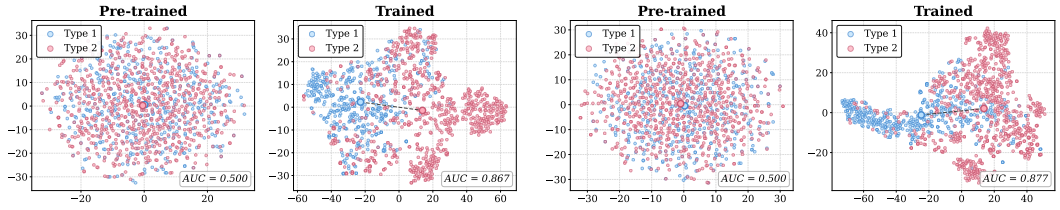
Model	Dataset	Wrong-Answer	Right-Answer	No-Answer	Total
Qwen-3-1.7B	HotpotQA	1088 (72.15%)	413 (27.39%)	7 (0.46%)	1508
	NQ	2286 (63.15%)	1333 (36.82%)	1 (0.03%)	3620
	NewsQA	383 (79.13%)	101 (20.87%)	0 (0.00%)	484
	SQuAD	1905 (68.70%)	865 (31.19%)	3 (0.11%)	2773
	SearchQA	2507 (36.33%)	4377 (63.43%)	16 (0.23%)	6900
	TriviaQA	1790 (44.84%)	2199 (55.09%)	3 (0.08%)	3992
	Total	**9959** (51.66%)	**9288** (48.18%)	**30** (0.16%)	**19277**
Qwen-3-8B	HotpotQA	876 (58.09%)	535 (35.48%)	97 (6.43%)	1508
	NQ	1857 (51.30%)	1696 (46.85%)	67 (1.85%)	3620
	NewsQA	379 (78.31%)	95 (19.63%)	10 (2.07%)	484
	SQuAD	1634 (58.93%)	1045 (37.68%)	94 (3.39%)	2773
	SearchQA	1192 (17.28%)	5574 (80.78%)	134 (1.94%)	6900
	TriviaQA	1100 (27.56%)	2836 (71.04%)	56 (1.40%)	3992
	Total	**7038** (36.51%)	**11781** (61.11%)	**458** (2.38%)	**19277**

Table 10: **F1 score of Qwen3-8B under three knowledge-conflict scenarios.** The **boldfaced** scores represent the **best** results.

Scenario	Method	HotpotQA	NQ	NewsQA	SQuAD	SearchQA	TriviaQA
TP	Origin	84.76	83.78	81.52	89.53	85.42	83.53
	Know-R1	**85.41**	84.19	**81.78**	90.11	85.93	83.78
	SPARK	84.11	**84.84**	81.53	**90.57**	**86.48**	**85.17**
TN	Origin	39.81	42.78	36.77	30.85	44.37	48.48
	Know-R1	40.28	43.65	37.53	32.69	44.83	49.09
	SPARK	**52.11**	**51.94**	**44.98**	**43.49**	**56.73**	**62.79**
FP	Origin	**74.85**	**65.18**	68.48	80.55	59.77	**47.94**
	Know-R1	73.76	65.01	**68.97**	80.34	**61.45**	46.65
	SPARK	74.38	64.56	68.07	**80.81**	61.03	47.78

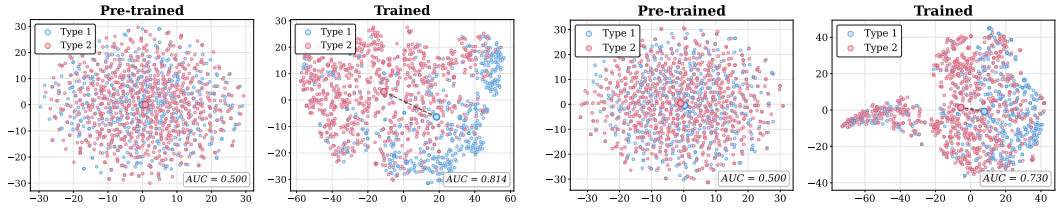
A.10 T-SNE VISUALIZATION OF PARTICIPATION VECTORS OF SPARK

In this section, we exhibit the t-SNE visualizations of the participation vectors to further illustrate the effectiveness of our proposed SPARK. Specifically, following the data construction procedure detailed in Appendix A.3, we sample 1,200 instances from MRQA evaluation subsets (200 per subset), categorized into two types: Type 1 (444 samples), where parametric knowledge is correct but the retrieved context is incorrect, and Type 2 (756 samples), where the retrieved context is correct but parametric knowledge is incorrect. For each instance, we collect the participation vectors from SPARK on Qwen-3-8B and represent them as an L-dimensional vector $\mathbf{g} = [g_1, g_2, \ldots, g_L]$, where g_l denotes participation degree at layer l. As illustrated in Figure 5, compared to the pre-trained model, the trained participation vectors exhibit visible separation between the two conflict types. To quantify this, we further evaluate the linear separability of the gate vectors via 5-fold cross-validated logistic regression (Kohavi et al., 1995). For example, the AUC of Qwen-3-8B increases from 0.500 to 0.707, suggesting that the gate module learns conflict-discriminative representations rather than applying uniform modulation across all inputs. This further suggests that SPARK learns an effective participation policy, providing supporting evidence for its adaptive parametric knowledge participation behavior under different conflict scenarios.



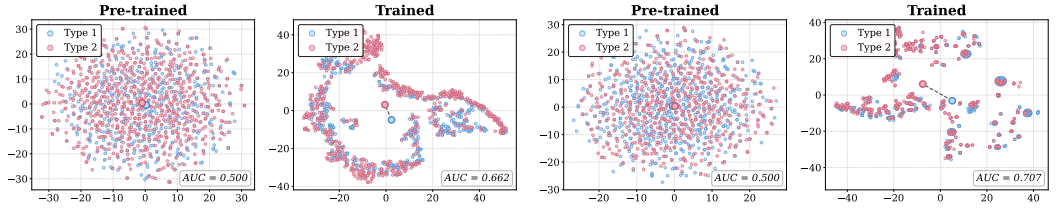
(a) Visualization of the participation vector on Llama-3.2-1B.

(b) Visualization of the participation vector on Llama-3.2-3B.



(c) Visualization of the participation vector on Llama-3.1-8B.

(d) Visualization of the participation vector on Qwen-3-0.6B.



(e) Visualization of the participation vector on Qwen-3-1.7B.

(f) Visualization of the participation vector on Qwen-3-8B.

Figure 5: T-SNE visualization of the participation vectors of different LLMs.

A.11 More Details about the Analysis of Scenario-Sensitive Neurons

In this section, we provide the full experimental details and per-layer results on Qwen3-1.7B and Qwen3-8B for the neuron activation analysis discussed in Section 6.4. We first describe the experimental setup and the statistical procedure used to identify scenario-sensitive neurons. We then examine how these neurons concentrate in deeper layers and compare `SPARK` with the vanilla base model. Finally, we quantify the alignment between the learned participation policy and the layer-wise distribution of scenario-sensitive neurons.

Experimental Setup. We analyze the SwiGLU intermediate activations of each FFN layer at the last input token position, where the dimensionality is d_m. For each model, we collect activations on two conflict scenarios constructed from six MRQA datasets: **TN** represents that the parametric knowledge is correct, but the context is wrong, and **FP** represents that the parametric knowledge is wrong, but the context provides the gold answer. For Qwen3-1.7B, which has L=28 layers and d_m=6,144, we obtain 9,288 TN and 9,959 FP samples. And for Qwen3-8B, which has L=36 layers and d_m=12,288, we obtain 11,781 TN and 7,038 FP samples. Both `SPARK` and the vanilla base model, which is equivalent to setting α=1 for all layers, are evaluated.

Identifying Scenario-sensitive Neurons. For each neuron j at layer l, we perform a two-sided Welch's t-test (Welch, 1947) comparing its activation values across TN and FP samples. To control the family-wise error rate over d_m simultaneous tests per layer, we apply Bonferroni correction with significance level $\alpha_{\text{test}} = 0.05/d_m$. A neuron is deemed *scenario-sensitive* if its corrected p-value falls below this threshold. We also compute Cohen's d as a standardized effect size for each neuron.

Concentration in Deep Layers. Table 11 reports the average proportion of scenario-sensitive neurons across layer groups. In both backbones, scenario-sensitive neurons are heavily concentrated in

the deeper layers, consistent with prior findings that mid-to-deep FFNs serve as primary repositories of parametric knowledge (Dai et al., 2022b).

Table 11: **Average proportion of scenario-sensitive neurons (%) by layer group in the SPARK model.**

Model	Early	Middle	Late
Qwen3-1.7B (L0–9 / L10–17 / L18–27)	0.8	25.4	61.7
Qwen3-8B (L0–11 / L12–23 / L24–35)	0.1	5.2	57.6

Comparison between SPARK and the Base Model. Table 12 presents the per-layer comparison between SPARK and the base model. Overall, SPARK slightly reduces the total proportion of scenario-sensitive neurons, by 1.7 for 1.7B and 0.5 for 8B, while the top-100 neurons ranked by effect size generally exhibit *higher* Cohen's d under SPARK in the 1.7B model, suggesting that SPARK's participation policy concentrates scenario-discriminative signals into fewer but more specialized neurons.

Table 12: **Scenario-sensitive neuron statistics: SPARK vs. Base.** "%sig" denotes the proportion of neurons passing the Bonferroni-corrected threshold. "Δpp" denotes the difference in percentage points between SPARK and the base model. "top$|d|$" is the mean |Cohen's d| of the 100 neurons with the largest effect sizes. Selected layers shown.

	Qwen3-1.7B				Qwen3-8B			
Layer	**Base%**	**SPARK%**	**Δpp**	**top$\|d\|$**	**Base%**	**SPARK%**	**Δpp**	**top$\|d\|$**
0	0.0	0.0	+0.0	1.53	0.0	0.0	+0.0	1.17
5	0.3	0.3	+0.0	1.62	0.0	0.0	−0.0	1.43
10	4.4	4.0	−0.4	1.61	0.2	0.2	+0.0	1.53
12	12.0	14.0	+2.0	1.62	0.3	0.3	+0.0	1.50
15	35.3	40.1	+4.8	1.33	1.4	1.5	+0.0	1.49
18	71.1	64.8	−6.3	1.41	2.9	2.6	−0.3	1.55
20	73.5	66.3	−7.2	1.31	7.6	7.1	−0.5	1.61
21	70.4	60.9	−9.5	1.26	11.3	11.2	−0.0	1.66
25	67.9	59.8	−8.1	1.11	28.5	28.1	−0.4	1.38
27	48.2	45.1	−3.2	1.04	48.8	47.6	−1.2	1.26
30	–	–	–	–	74.4	70.2	−4.2	1.07
35	–	–	–	–	59.6	62.5	+2.9	0.99
Avg	31.2	29.6	−1.7		21.4	20.9	−0.5	

Correlation between Participation Policy and Sensitive-neuron Distribution. To quantify how well the learned participation policy aligns with the knowledge-conflict landscape, we compute the Pearson correlation between the per-layer participation deviation $|\alpha^{(l)}-1|$ and the proportion of scenario-sensitive neurons at layer l. As shown in Table 13, both models exhibit a strong positive correlation, with r=0.677 for 1.7B and r=0.768 for 8B, both significant at p<0.001, confirming that the RL-trained policy applies stronger modulation at layers where knowledge-conflict signals are most prevalent.

Table 13: **Pearson correlation between $|\alpha-1|$ and the proportion of scenario-sensitive neurons (%sig).**

Model	r	p-value
Qwen3-1.7B	0.677	< 0.001
Qwen3-8B	0.768	< 0.001

A.12 Case Studies

In this section, we provide the qualitative case studies to illustrate the effectiveness of our SPARK. Table 14 and Table 15 present representative case studies from Qwen3-1.7B and Qwen3-8B, respectively, under the TN and FP knowledge-conflict scenarios.

Table 14: Case studies from Qwen3-1.7B under two knowledge-conflict scenarios. Correct answers are highlighted in **Red**; incorrect answers are highlighted in **Blue**. In TN, parametric knowledge is correct but the context is misleading; in FP, parametric knowledge is wrong but the context provides the gold answer.

TN Case 1 *(HotpotQA — Parametric knowledge: Manchester United)*
Question: Joe Ford played for which football club based out of Old Trafford?
Context (cf): Peter Ford (footballer) Peter Leslie Ford (born 10 August 1933) is an English former footballer who played in the Football League for Port Vale and Stoke City . . .

Vanilla-RAG: Unknown **SFT:** West Ham United **SPARK:** Manchester United

TN Case 2 *(NewsQA — Parametric knowledge: India)*
Question: What other country is communicating with the Tamil Tigers?
Context (cf): The Tamil Tigers are designated as a terrorist organization by 32 countries, including the United States, Canada, and the member nations of the European Union . . .

Vanilla-RAG: (empty) **SFT:** Malaysia **SPARK:** India

FP Case 1 *(HotpotQA — Parametric knowledge: Western Cape)*
Question: The 1974 South African Grand Prix was a Formula One motor race held at a racing circuit located in which South African Province?
Context: The 1974 South African Grand Prix was a Formula One motor race held at **Kyalami** on 30 March 1974 . . . Kyalami is a racing circuit located in Midrand, **Gauteng province** . . .

Vanilla-RAG: (empty) **SFT:** Midrand **SPARK:** Gauteng province

FP Case 2 *(NaturalQuestions — Parametric knowledge: Harrison Ford)*
Question: Who was originally cast to play Indiana Jones?
Context: Originally, Spielberg suggested Harrison Ford; Lucas resisted the idea . . . and did not want Ford to become his "Bobby De Niro" . . . **Tom Selleck** was cast as Indiana Jones; however, CBS would not release Selleck from Magnum, P.I. . . .

Vanilla-RAG: Harrison Ford **SFT:** Selleck **SPARK:** Tom Selleck

In TN Case 1 of Qwen3-1.7B, the question asks which football club Joe Ford played for, and the model's parametric knowledge correctly points to Manchester United. However, the counterfactual context describes a different player (Peter Ford) who played for entirely different clubs. Vanilla-RAG fails to extract any relevant answer and outputs "Unknown", while SFT is misled into producing "West Ham United". SPARK correctly retains its parametric knowledge and answers "Manchester United". A similar pattern appears in TN Case 2, where the counterfactual context omits the answer "India" entirely; Vanilla-RAG produces an empty response, SFT hallucinates "Malaysia", and only SPARK answers correctly. These TN cases demonstrate that SPARK's adaptive participation policy enables the model to resist misleading context and preserve reliable parametric knowledge when appropriate. In FP Case 1, the model's parametric knowledge incorrectly recalls "Western Cape" as the province where the 1974 South African Grand Prix took place, while the context explicitly states Kyalami is located in Gauteng province. SFT extracts the nearby but incorrect surface span "Midrand" (a city, not a province), and Vanilla-RAG fails to produce any answer. SPARK correctly follows the context and answers "Gauteng province". FP Case 2 is particularly illustrative: the question asks who was originally cast to play Indiana Jones. The context mentions both Harrison Ford (who was suggested by Spielberg) and Tom Selleck (who was actually cast). Vanilla-RAG latches onto the more salient name "Harrison Ford"—which coincides with the model's parametric memory—while SFT produces the incomplete answer "Selleck". SPARK accurately extracts "Tom Selleck" from the context. These FP cases show that SPARK can suppress incorrect parametric knowledge and faithfully follow the retrieved context when it provides the correct answer.

The Qwen3-8B cases in Table 15 exhibit consistent patterns. Notably, in FP Case 2, the context contains a structured WHO table listing Zabol (Iran) as the most polluted city, contradicting the parametric answer "Delhi, India". Vanilla-RAG claims there is "not enough information", SFT hallucinates "Dhaka", and only SPARK correctly extracts "Zabol" from the tabular data. Across both model scales, these cases illustrate that SPARK achieves more faithful knowledge arbitration in both directions: retaining parametric knowledge when the context is unreliable (TN) and following contextual evidence when parametric knowledge is wrong (FP).

Table 15: Case studies from Qwen3-8B under two knowledge-conflict scenarios. Correct answers are highlighted in **Red**; incorrect answers are highlighted in **Blue**. In TN, parametric knowledge is correct but the context is misleading; in FP, parametric knowledge is wrong but the context provides the gold answer.

TN Case 1 *(HotpotQA — Parametric knowledge: House of Commons)*
Question: The constituency which is represented by the 2015 elected Minister for Universities and Science is represented in what wing of the UK parliament?
Context (cf): David Willetts David Linsay Willetts, Baron Willetts, (born 9 March 1956) is an English Conservative Party politician, life peer, and academic. From 1992 to 2015, he was the Member of . . .

Vanilla-RAG: Conservative **SFT:** Conservative Party `SPARK`: House of Commons

TN Case 2 *(NaturalQuestions — Parametric knowledge: Vicente Fox)*
Question: Who was elected president in Mexico in 2000?
Context (cf): . . . a majority in both chambers of Congress—a situation that prevented Fox from implementing his reform pledges. Nonetheless, the transfer of power in 2000 was quick and peaceful . . .

Vanilla-RAG: Felipe Calderón Hinojosa **SFT:** Victor Manuel Mejía `SPARK`: Vicente Fox

FP Case 1 *(HotpotQA — Parametric knowledge: Cruz Azul)*
Question: Rafael de Souza Pereira is a defensive midfielder for what organization that is the top level of the Mexican football league system?
Context: Rafael Carioca Rafael de Souza Pereira (January 16, 1989) . . . is a Brazilian footballer who plays as a defensive midfielder for Liga MX club Tigres UANL . . .

Vanilla-RAG: Liga BBVA Bancomer **SFT:** Liga BBVA Bancomer `SPARK`: Liga MX

FP Case 2 *(NaturalQuestions — Parametric knowledge: Delhi, India)*
Question: Most polluted city in the world according to the World Health Organization?
Context: . . . Position: 1, Country: Iran, Town/City: Zabol, PM2.5: 217 . . .

Vanilla-RAG: Not enough information provided **SFT:** Dhaka `SPARK`: Zabol

A.13 Prompt Templates

In this section, we provide the prompt templates used in our experiments, covering both training and evaluation.

Training Prompt

System Prompt: You are a helpful assistant who answers questions based on the provided background information. The background information may be incorrect, so you should judge whether to believe yourself or to believe the background information. You should think about the reasoning process and then provide the answer based on the given context.
Prompt Template: Background: context
Task Instruction: Answer the question with the given background information above. Provide your final answer in <answer> </answer> tags, for example <answer>Petri Alanko</answer>.
Q: question
A:

Figure 6: Training prompt used for Qwen models in Non-Thinking Mode during GRPO training in `SPARK`.

Eval Prompt

System Prompt: You are a helpful assistant who answers questions based on the provided background information. The background information may be incorrect, so you should judge whether to believe yourself or to believe the background information. You should think about the reasoning process and then provide the answer based on the given context.
Prompt Template: Background: context
Task Instruction: Answer the question with the given background information above. Provide your final answer in <answer> </answer> tags, for example <answer>Petri Alanko</answer>.
Q: question
A:

Figure 7: Eval prompt on Qwen models in Non-Thinking Mode used for evaluation.

Training Prompt

System Prompt: You are a helpful assistant who answers questions based on the provided background information. The background information may be incorrect, so you should judge whether to believe yourself or to believe the background information. You should think about the reasoning process and then provide the answer based on the given context.
Prompt Template: Background: context
Task Instruction: Answer the question with the given background information above. Provide your final answer in <answer> </answer> tags, for example <answer>Petri Alanko</answer>.
Q: question
A:

Figure 8: Training prompt used for Llama models during GRPO training in SPARK.

Eval Prompt

System Prompt: You are a helpful assistant who answers questions based on the provided background information. The background information may be incorrect, so you should judge whether to believe yourself or to believe the background information. You should think about the reasoning process and then provide the answer based on the given context.
Prompt Template: Background: context
Task Instruction: Answer the question with the given background information above. Provide your final answer in <answer> </answer> tags, for example <answer>Petri Alanko</answer>.
Q: question
A:

Figure 9: Eval prompt on Llama models used for evaluation.