FCPRAG: Fusion-Controller Parametric Retrieval-Augmented Generation for Stable Multi-Passage LoRA Injection

Anonymous ACL submission

Abstract

Parametric retrieval-augmented generation (PRAG) injects retrieved evidence into a large language model (LLM) through passage-specific LoRA adapters, reducing reliance on long in-context prompts. When multiple passages are retrieved for the same query, however, *evidence-level fusion* becomes a bottleneck: equal-weight merging can amplify weak or conflicting evidence, and translating retrieval signals into fusion weights often requires fragile global tuning. We propose **FCPRAG**, a fusion-controlled parametric RAG framework that adds a lightweight controller for retrieval-conditioned, sample-level adapter fusion. The controller predicts per-passage fusion scores together with sample-level calibration signals, including a mixing gate and an adaptive temperature, enabling fusion that stays selective under informative retrieval signals and conservative under uncertainty. FCPRAG is trained with merge-aware supervision derived from each adapter's marginal contribution within a multi-adapter merge, using training data only. We further show that a single dataset-level temperature is suboptimal under heteroscedastic retrieval uncertainty, motivating sample-level adaptation. Experiments on HotpotQA, 2WikiMultiHopQA, PopQA, and ComplexWebQuestions (CWQ) across three LLM backbones show that FCPRAG consistently improves F1 over standard RAG and parametric RAG baselines, with gains of up to 4.65% on 2WikiMultiHopQA and 7.55% on CWQ, while also reducing tuning cost and improving robustness under retrieval perturbations.

1 Introduction

Retrieval-augmented generation (RAG) improves the factuality of large language models (LLMs) by retrieving external evidence and conditioning generation on the retrieved content (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021). However, classical in-context RAG increases prompt length with the number and size of retrieved passages, leading to higher inference cost and potential degradation under long or conflicting contexts. Recent RAG research therefore explores stronger retrieval and generation control to reduce redundancy and error propagation (Jiang et al., 2023; Asai et al., 2024), but the fundamental trade-off remains: multi-evidence reasoning often benefits from more context, while longer contexts are expensive and fragile.

Parametric RAG (PRAG) (Su et al., 2025a) offers an alternative: instead of appending passages to the prompt, it *parameterizes* each passage into lightweight LoRA weights (Hu et al., 2022) during an offline stage, and loads these passage-specific parameters at test time to inject knowledge into the model. This design reduces online context length and admits an optional combine setting that also appends retrieved passages to the input prompt (Section 3.1). Recent PRAG-family methods further reduce cost by using translators or routing to reuse a small set of latent experts (Tan et al., 2025; Su et al., 2025b).

While these methods successfully reduce inference cost by moving evidence from the prompt into model parameters, they introduce a new challenge: *multi-evidence fusion*. For many knowledge-intensive queries, retrieval returns multiple relevant passages (often partially relevant, redundant, or mutually inconsistent). In PRAG-family systems, this yields multiple passage-specific LoRA experts that must be merged before generation. Uniform fusion (equal weights) can inject noise from weak passages, while hand-tuned fusion heuristics (e.g., globally set softmax temperature or blending factors) often fail to generalize across queries, tasks, and model scales. This sensitivity affects both efficiency and robustness: in PRAG-style parametric injection, the fusion weights directly control which evidence is written into the model's effective parameters at test time, so even small calibra-

Table 1: Capability comparison across RAG, PRAG-family systems, and FCPRAG.

Method	Parametric evidence injection	Fusion in parameter space	Sample-level calibrated fusion	Uniform-prior fallback
RAG (Lewis et al., 2020)	×	×	×	×
PRAG (Su et al., 2025a)	✓	✓	×	×
DyPRAG (Tan et al., 2025)	✓	✓	×	×
Poly-PRAG (Su et al., 2025b)	✓	✓	×	×
FCPRAG (Ours)	✓	✓	✓	✓

tion mismatches can over- or under-inject passage updates (Figure 2(a)). As a result, PRAG-family pipelines frequently rely on expensive grid search over global temperature and blending coefficients to reach strong performance.

To understand why this grid search is difficult to eliminate, we identify three properties that make this fusion bottleneck particularly challenging in parametric RAG. **First, evidence utility is highly sample-dependent.** Even under a fixed retriever, the same query can surface a mix of strongly supportive, weakly relevant, redundant, or distracting passages, and the *marginal contribution* of each passage-specific adapter (estimated by removing one adapter from the merge and measuring the F1 drop) can vary substantially across instances and subtasks. Uniform fusion therefore tends to either dilute a dominant adapter (when a single passage is decisive) or over-inject weak adapters (when retrieval is noisy). For example, in bridge-style multi-hop QA, one retrieved passage may contain the bridge entity needed to answer the question, while others are distractors. **Second, parametric injection amplifies fusion mistakes.** Unlike in-context conditioning, where the model may ignore or down-weight unhelpful tokens at generation time, parametric RAG fuses evidence through weight-space updates: an over-weighted passage adapter directly alters the model's effective parameters for the current query, making conflicts and noise harder to "wash out" via prompting. **Third, mapping scores to weights is itself a calibration problem.** Even if a scoring model ranks passages reasonably well, the optimal sharpness of the resulting distribution (temperature) and the degree of deviation from uniform fusion (blending) can change across backbones, tasks, and retrieval quality, which is why grid search is often needed to stabilize performance. This motivates learning *per-sample* calibration so that fusion can be selective when confident and conservative when uncertain, while avoiding costly per-model tuning (Figures 2 and 3).

At a high level, this fusion problem resembles model and adapter merging (Wortsman et al., 2022; Ilharco et al., 2022; Yadav et al., 2023; Pfeiffer et al., 2021; Ainsworth et al., 2022; Jin et al., 2022; Stoica et al., 2023; Izmailov et al., 2018; Matena and Raffel, 2022). However, PRAG-family fusion is fundamentally *retrieval-conditioned*: the adapters to be merged are tied to passages retrieved for a specific query, and the appropriate confidence in a peaked vs. conservative weighting can change from sample to sample. This motivates treating fusion as a learned, calibrated decision rather than a fixed global heuristic, analogous in spirit to classical calibration of predictive distributions (Guo et al., 2017; Platt et al., 1999; Niculescu-Mizil and Caruana, 2005; Kuleshov et al., 2018).

To solve the above challenges, we propose FCPRAG which introduces a learned *Fusion Controller* into the PRAG inference pipeline. The controller predicts sample-level fusion weights for the retrieved passages and jointly learns per-sample calibration signals to stabilize the mapping from raw fusion scores to final merge weights. Importantly, FCPRAG is designed as a plug-in: it does not change the underlying PRAG offline parameterization, retrieval, or generation, and can replace the merge step in PRAG/DyPRAG with minimal engineering. We empirically validate that this learned calibration can replace manual hyperparameter tuning (Figure 3).

Why fusion control is the right interface. A key shift in parametric RAG is that evidence is injected through *parameters* rather than *tokens*. When retrieval returns N passages, PRAG-family systems produce N passage-LoRA adapters that must be fused into a single injected update. This fusion step becomes the main place where evidence relevance, redundancy, and conflict must be resolved. Our design therefore targets the fusion interface directly: we learn (i) evidence-level weights and (ii) per-sample calibration, so the system can be both selective when confident and conservative when uncertain. To summarize, in this study, we made the following contributions.

- **FCPRAG architecture.** We introduce a learned Fusion Controller for PRAG-family inference that performs retrieval-conditioned, sample-level LoRA fusion and calibration, adding $< 1\%$ trainable parameters relative to the backbone LLM.
- **Theoretical motivation.** We prove (Proposi-

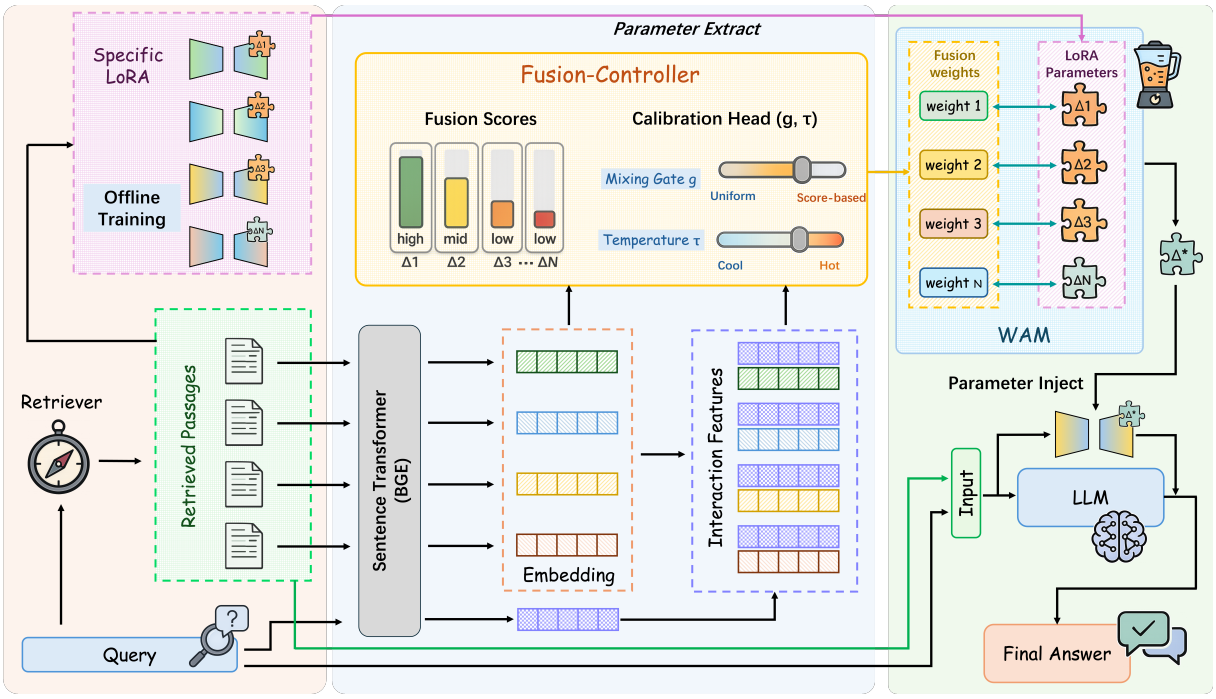

Figure 1: Overview of FCPRAG. Given a query, retrieve multiple passages and load their passage-specific LoRA adapters (offline-trained). A Fusion Controller predicts per-passage weights and per-sample calibration, then a weighted adapter merger (WAM) merges the adapters for parametric injection (optionally combined with in-context passages).

tion 1) that under heteroscedastic retrieval uncertainty, any single dataset-level fusion temperature incurs strictly positive calibration regret, formally motivating sample-level (g, τ) adaptation.

- **Merge-aware supervision under no-leak.** We propose a training-signal construction that supervises fusion weights using each passage adapter's marginal contribution within a multi-LoRA merge, while enforcing a no-leak train/eval separation.

- **Empirical gains and key ablations.** On HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020), FCPRAG improves F1 across three LLM backbones (Llama-3.2-1B (Touvron et al., 2023), Qwen2.5-1.5B (Qwen et al., 2025), Llama-3-8B (Touvron et al., 2023)) and achieves the best Avg. Rank in Table 2. A random controller ablation is substantially worse, validating the learned fusion mechanism.

2 Related Work

Unlike in-context RAG, which injects retrieved evidence by concatenating passages to the prompt, parametric RAG injects retrieved knowledge as parameter updates (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022; Su et al., 2025a). This avoids long-context overhead, but under multi-passage retrieval it introduces a different online bottleneck: multiple evidence-conditioned updates must be combined before generation.

PRAG (Su et al., 2025a) converts each retrieved passage into passage-specific LoRA weights via offline fine-tuning (Hu et al., 2022; Houlsby et al., 2019; Li and Liang, 2021; Han et al., 2024). DyPRAG (Tan et al., 2025) reduces per-passage training and storage by dynamically generating parametric knowledge at inference time, while Poly-PRAG (Su et al., 2025b) routes queries to a small set of latent LoRA experts. Despite these differences in offline parameterization, PRAG-family systems share the same online fusion problem: when several passages are retrieved, the model must merge multiple passage-conditioned parametric updates, often under noisy or partially conflicting evidence.

Existing fusion methods mainly operate either in token/prediction space, such as Fusion-in-Decoder and RePlug (Izacard and Grave, 2021; Shi et al.,

2024), or in task-level parameter composition, such as model soups, task arithmetic, TIES-Merging, and AdapterFusion (Wortsman et al., 2022; Ilharco et al., 2022; Yadav et al., 2023; Pfeiffer et al., 2021). Our setting is different: we study *retrieval-conditioned evidence-level fusion in parameter space*, where useful merge weights vary across queries and over-weighting weak evidence can inject spurious knowledge. FCPRAG addresses this gap by learning both per-query fusion scores and calibration that controls how confidently the model departs from uniform fusion. Extended background is provided in Appendix B.2.

3 Methodology

3.1 Problem Formulation and PRAG Baseline

We consider knowledge-intensive QA where, given a query q, a retriever returns a set of passages $\mathcal{P} = \{p_1, \ldots, p_N\}$. PRAG (Su et al., 2025a) parameterizes each passage p_i into a lightweight passage-LoRA adapter Δ_i by offline fine-tuning on synthetic QA-style supervision derived from p_i. At test time, PRAG loads the offline-trained adapters $\{\Delta_i\}_{i=1}^N$ corresponding to the retrieved passages $\mathcal{P}$ and merges them into a single adapter $\Delta(\mathcal{P})$, which is injected into the LLM to generate an answer. We denote this parametric-only inference as **Parametric-only** (**Param.**), and an optional variant **Combine** where the same retrieved passages are also provided to the LLM in-context in addition to parametric injection (following the "PRAG + in-context" combination described in (Su et al., 2025a)).

Data augmentation (training-side). Following the PRAG-family pipeline, we construct synthetic supervision for each passage p_i using an instruction-tuned LLM: (i) a rewritten variant of p_i and (ii) a small set of QA pairs whose answers are supported by p_i. These augmentation artifacts define a passage-specific training set D_i.

Passage-LoRA training (offline parameterization). We then fine-tune a passage-specific LoRA adapter Δ_i on QA-style prompts built from D_i, optimizing the standard causal language modeling loss while keeping the backbone LLM frozen. To encourage parametric encoding rather than relying only on in-context evidence, training mixes *passage-conditioned* examples (the prompt includes p_i or its rewrite) and *question-only* examples (the prompt omits the passage but still supervises the answer). Appendix B.8 lists the training-side prompt templates used for augmentation and passage-LoRA fine-tuning.

3.2 Evidence-level Multi-LoRA Fusion Setting

In the multi-passage setting, each passage yields a passage-LoRA adapter Δ_i and the system must fuse $\{\Delta_i\}_{i=1}^N$ before generation. Uniform fusion (equal weights) treats all passages as equally useful, which can be suboptimal when passages have different relevance or when some passages introduce conflicting knowledge. We aim to learn a *fusion policy* that assigns weights $w_i(q, \mathcal{P})$ per query so that the merged adapter emphasizes helpful evidence and suppresses noise.

3.3 FCPRAG: Fusion Controller with Learned Calibration

Figure 1 shows the high-level pipeline. FCPRAG augments PRAG-family inference with a lightweight Fusion Controller f_θ. Given a query and its retrieved passages, the controller predicts retrieval-conditioned fusion scores for the corresponding passage-LoRA adapters, and outputs per-sample calibration signals to stabilize the mapping from scores to weights. We denote the downstream weighted merge block as a *Weighted Adapter Merger* (WAM), which applies controller weights to retrieved adapters and outputs one merged adapter for injection. Inference follows a simple sequence: encode the query and retrieved passages, predict fusion scores and sample-wise calibration, map them to merge weights, merge the retrieved adapters, and generate in either Parametric-only or Combine mode. Appendix B.3 gives pseudocode.

Inputs. For controller inputs, we construct interaction features from dense query and passage embeddings produced by a sentence embedding model. Let $\mathbf{e}_q \in \mathbb{R}^d$ be the query embedding and $\mathbf{E}_p = [\mathbf{e}_{p_1}, \ldots, \mathbf{e}_{p_N}] \in \mathbb{R}^{N \times d}$ be passage embeddings (normalized). We form interaction features for each passage:

$$\mathbf{x}_i = [\mathbf{e}_q;\ \mathbf{e}_{p_i};\ \mathbf{e}_q \odot \mathbf{e}_{p_i};\ |\mathbf{e}_q - \mathbf{e}_{p_i}|]\,. \quad (1)$$

Outputs and weight mapping. The controller predicts a scalar score $s_i \in (0, 1)$ for each passage via an MLP, and a calibration head predicts: (i) a gate $g \in (0, 1)$ and (ii) a temperature $\tau \in [\tau_{\min}, \tau_{\max}]$. We map scores to a probability distri-

bution and blend with a uniform prior:

$$\pi_i = \mathrm{softmax}\left(\frac{s_i}{\tau}\right), \tag{2}$$

$$w_i = g \cdot \pi_i + (1-g) \cdot \frac{1}{N}. \tag{3}$$

Intuitively, τ controls how sharply scores are converted into weights, while g controls how far the controller deviates from uniform fusion. This mirrors the theoretical role of calibration below: heteroscedastic retrieval uncertainty motivates sample-level sharpness, and the uniform-prior interpolation in Eq. 3 implements shrinkage toward conservative fusion when score signals are unreliable. The theory supports this calibration choice rather than claiming optimality of the full controller architecture.

Weighted Adapter Merger (WAM). Given retrieved adapters $\{\Delta_i\}_{i=1}^N$ and controller outputs $\{w_i\}_{i=1}^N$ from Eq. 3, we first apply merge-time normalization:

$$\tilde{w}_i = \frac{N\, w_i}{\sum_{j=1}^N w_j + \epsilon}. \tag{4}$$

WAM then applies the normalized weights as scaling factors and composes the scaled adapters into a single merged adapter:

$$\Delta^\star = \mathcal{M}_{\mathrm{WAM}}\big(\{\tilde{w}_i \cdot \Delta_i\}_{i=1}^N\big). \tag{5}$$

Finally, we inject $\Delta^\star$ into the frozen backbone model and denote the adapted model by $\mathcal{M}_{\Delta^\star}$:

$$\mathcal{M}_{\Delta^\star} \leftarrow \textsc{Inject}(\mathcal{M}, \Delta^\star), \quad \hat{a} \leftarrow \textsc{Generate}(\mathcal{M}_{\Delta^\star}, \mathrm{prompt}). \tag{6}$$

3.4 Merge-aware Training Signal and No-leak Protocol

Training the Fusion Controller requires supervision that reflects how each passage-LoRA adapter contributes under *multi-LoRA fusion* (as opposed to isolated single-adapter behavior). We therefore construct *merge-aware* marginal contribution labels using training-side data only: for each training instance $(q, \mathcal{P}, a)$, we evaluate the system with all N adapters merged under uniform fusion to obtain F_1^{all}, then remove the i-th adapter and re-evaluate to obtain $F_1^{\setminus i}$; the difference measures the marginal effect of that adapter under fusion. We define

$$\delta_i = F_1^{\mathrm{all}} - F_1^{\setminus i}, \tag{7}$$

where we use δ_i to denote a scalar marginal contribution (to avoid confusion with LoRA deltas Δ_i). We convert $\{\delta_i\}$ into a soft target distribution

$$y_i = \mathrm{softmax}\left(\frac{\delta_i}{T_{\mathrm{label}}}\right). \tag{8}$$

We then train the controller to match this target distribution via KL divergence (below); when contributions are flat (no preference), the target becomes uniform and we down-weight these samples to prevent learning a trivial uniform mapping. A schematic illustration and label-source analysis are deferred to Appendix B.5. Label construction and controller training use training data only, while evaluation is performed on a held-out split.

Objective. Let $\mathbf{w}_\theta(q, \mathcal{P})$ be the controller's predicted distribution from Eq. 3. For sample b, we define

$$\ell_{\mathrm{KL}}^{(b)} = \sum_{i=1}^N y_i^{(b)} \log \frac{y_i^{(b)}}{w_i^{(b)} + \epsilon}, \tag{9}$$

and optimize the weighted batch objective

$$\mathcal{L}_{\mathrm{KL}} = \frac{\sum_b \omega^{(b)} \ell_{\mathrm{KL}}^{(b)}}{\sum_b \omega^{(b)} + \epsilon}, \tag{10}$$

where $\omega^{(b)}$ is the sample weight. When the deltas are flat (no preference), $\mathbf{y}$ becomes uniform; we down-weight these samples to prevent the model from learning a trivial uniform mapping.

3.5 Plug-and-play Integration

FCPRAG is designed to be a drop-in fusion module: **PRAG** uses uniform weights in the merge; FCPRAG replaces those weights with $\mathbf{w}_\theta$. **DyPRAG** additionally uses translator-generated deltas; FCPRAG can be applied by replacing mean pooling/averaging in the aggregation stage with a weighted sum driven by $\mathbf{w}_\theta$. **Poly-PRAG** performs routing over latent experts; we discuss how a calibrated fusion controller could complement routing decisions in future work, but do not claim full plug-and-play integration.

3.6 Efficiency and Overhead

The Fusion Controller is a small MLP operating on retriever embeddings. Compared to PRAG-family inference, the additional overhead is a single embedding forward pass for the query and passages plus lightweight MLP computation. Controller

Table 2: Main results (F1, %) on 2WikiMultiHopQA, HotpotQA, PopQA, and ComplexWebQuestions (CWQ). Suffix "-Combine" denotes results under the Combine setting. All baseline methods (Vanilla, Standard RAG, PRAG, DyPRAG) were manually reproduced under the same experimental setup for fair comparison. **Avg. Rank** is the mean rank (lower is better) over the 8 non-Avg columns within each backbone block. Best results are highlighted in **bold** and second-best results are underlined.

Base LLM	Method	2WikiMultiHopQA					HotpotQA		PopQA	CWQ	Avg. Rank
		Compare	Bridge	Inference	Compose	Avg.	Bridge	Compare			
Llama-3.2-1B	Vanilla	36.48	34.23	13.80	5.60	22.53	9.85	37.84	14.56	28.52	8.00
	Standard RAG	41.22	39.31	15.57	7.22	25.83	11.42	39.39	17.15	31.78	6.38
	PRAG	**46.83**	**44.32**	17.16	8.55	29.22	12.99	41.84	19.16	32.09	3.62
	PRAG-Combine	41.03	41.26	19.37	8.21	27.47	21.35	42.73	32.94	31.82	4.31
	DyPRAG	41.53	37.98	18.67	8.21	26.60	10.84	42.91	18.69	31.92	5.31
	DyPRAG-Combine	44.87	41.42	19.31	9.42	28.76	11.35	45.67	21.14	33.87	3.25
	FCPRAG (Our)	44.29	42.14	20.21	5.97	28.15	14.56	**47.49**	17.28	**34.93**	3.12
	FCPRAG-Combine	44.11	41.47	**21.67**	**10.65**	**29.48**	**23.05**	45.60	**38.33**	34.17	**2.00**
Qwen2.5-1.5B	Vanilla	39.53	37.26	15.20	6.40	24.60	13.43	43.23	16.58	31.25	6.62
	Standard RAG	40.89	33.12	18.79	11.10	25.97	16.32	43.96	19.73	31.50	4.88
	PRAG	43.85	35.41	**20.17**	12.77	28.05	17.35	46.22	18.90	32.68	3.62
	PRAG-Combine	40.62	37.95	18.03	7.58	26.04	18.71	39.33	23.01	21.48	5.12
	DyPRAG	43.91	36.54	17.98	9.12	26.89	11.69	43.18	19.92	31.54	5.25
	DyPRAG-Combine	42.23	40.83	18.56	9.75	27.84	12.13	46.41	**32.08**	32.29	3.50
	FCPRAG (Our)	**45.59**	42.89	17.81	**13.18**	**29.87**	18.58	**51.54**	21.30	**35.27**	**2.38**
	FCPRAG-Combine	38.97	**43.73**	19.04	7.78	27.38	**19.98**	40.80	20.77	20.02	4.62
Llama-3-8B	Vanilla	50.88	48.60	21.49	12.68	33.41	37.80	61.00	36.39	39.88	7.25
	Standard RAG	55.46	53.68	20.94	11.80	35.47	41.25	63.43	37.97	41.29	6.50
	PRAG	58.11	56.29	21.34	12.73	37.12	43.49	66.03	39.06	43.08	5.00
	PRAG-Combine	62.47	58.23	29.94	18.92	42.39	75.18	**82.83**	42.29	42.49	2.38
	DyPRAG	55.32	52.87	20.44	11.92	35.14	43.76	62.92	35.87	41.14	7.00
	DyPRAG-Combine	59.81	56.94	27.65	18.82	40.81	71.67	81.12	39.21	42.96	3.38
	FCPRAG (Our)	**67.07**	56.53	25.88	16.16	41.41	43.95	70.64	41.33	**50.63**	3.12
	FCPRAG-Combine	67.04	**62.39**	**32.86**	**25.88**	**47.04**	**76.15**	81.87	**43.79**	48.73	**1.38**

training is fast and does not require backpropagating through the LLM; label construction is the main offline cost but can be parallelized and reused across controller variants.

3.7 Theoretical Motivation

We provide a formal justification for sample-level fusion calibration. The full proof is in Appendix A.

Proposition 1 (Limitation of Dataset-level Fusion Calibration)**.** *Under heteroscedastic retrieval uncertainty, different queries induce different optimal fusion sharpness levels. Consequently, no single dataset-level temperature can be optimal for all queries, and any fixed global temperature incurs positive expected calibration regret under the risk-aware objective analyzed in Appendix A:*

$$\mathrm{Reg}(T) \triangleq \mathbb{E}_{t,\mathbf{s}}\Big[\mathcal{R}_t\big(\mathbf{w}(T_t^\star;\mathbf{s})\big) - \mathcal{R}_t\big(\mathbf{w}(T;\mathbf{s})\big)\Big], \tag{11}$$

In particular, even the best dataset-level temperature

$$T^\star_{\text{global}} \in \arg\max_{T>0} \mathbb{E}_{t,\mathbf{s}}\Big[\mathcal{R}_t\big(\mathbf{w}(T;\mathbf{s})\big)\Big] \tag{12}$$

satisfies $\mathrm{Reg}(T^\star_{\text{global}}) > 0$. *The full assumptions and proof are deferred to Appendix A.*

Intuition. Softmax temperature controls how sharply evidence is fused in parameter space. Under low retrieval uncertainty, sharp fusion is beneficial; under high uncertainty, conservative (near-uniform) fusion is safer. When uncertainty varies across queries, a single dataset-level temperature cannot satisfy both regimes. This directly motivates the calibration head in Eq. 3: τ adapts fusion sharpness, while g allows the learned distribution to shrink back toward the uniform prior under unreliable retrieval signals. Appendix A.3 derives the corresponding oracle-weight form.

4 Experimental Setup

We evaluate on HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), PopQA (Mallen et al., 2023), and ComplexWebQuestions (CWQ) (Talmor and Berant, 2018), using three instruction-tuned backbones: Llama-3.2-1B, Qwen2.5-1.5B, and Llama-3-8B (Touvron

Table 3: Ablation on 2WikiMultiHopQA (F1 scores, Qwen2.5-1.5B and Llama-3-8B). We compare uniform fusion (PRAG), a random (untrained) fusion controller, and our learned fusion controller (FCPRAG) under Parametric-only and Combine settings.

Base LLM	Method	Cmp.	Br.	Inf.	Compst.	Avg.
Qwen2.5-1.5B	PRAG	43.85	35.41	**20.17**	12.77	28.05
	PRAG-Combine	40.62	37.95	18.03	7.58	26.04
	Random Ctrl.	44.39	36.05	16.10	7.95	26.12
	Random Ctrl.-Combine	38.98	41.87	18.32	7.32	26.62
	FCPRAG (Ours)	**45.59**	42.89	17.81	**13.18**	**29.87**
	FCPRAG-Combine	38.97	**43.73**	19.04	7.78	27.38
Llama-3-8B	PRAG	58.11	56.29	21.34	12.73	37.12
	PRAG-Combine	62.47	58.23	29.94	18.92	42.39
	Random Ctrl.	60.18	54.12	22.05	12.11	37.12
	Random Ctrl.-Combine	64.58	59.01	28.47	17.18	42.31
	FCPRAG (Ours)	**67.07**	56.53	25.88	16.16	41.41
	FCPRAG-Combine	67.04	**62.39**	**32.86**	**25.88**	**47.04**

et al., 2023; Qwen et al., 2025). Following PRAG-family evaluation practice (Su et al., 2025a; Tan et al., 2025), we use **F1** (%) as the primary metric. For HotpotQA and 2WikiMultiHopQA, we report subtask-level F1; for 2Wiki we also report macro average over its four subtasks.

We compare Vanilla, Standard RAG (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021), PRAG (Su et al., 2025a), DyPRAG (Tan et al., 2025), and FCPRAG. All methods use the same retriever and the same top-K passages per query; only the knowledge injection and fusion mechanism differ. For parametric methods (PRAG, DyPRAG, and FCPRAG), we report both Parametric-only and Combine settings (Section 3.1). A randomly initialized controller is evaluated only as an ablation (Table 3). Full implementation details are deferred to Appendix B.4. Runtime and implementation overhead are discussed in Appendix B.4.

5 Experiments

5.1 Main Results

Table 2 reports results on 2WikiMultiHopQA, HotpotQA, PopQA, and ComplexWebQuestions (CWQ). Across backbones and benchmarks, FCPRAG generally improves over PRAG's uniform merge, with gains most pronounced on multi-evidence settings where retrieved passages vary in usefulness or contain distractors. These results support our thesis that retrieval-conditioned evidence fusion is a first-class bottleneck in parametric RAG, and that learning both weights and calibration provides a practical upgrade over globally tuned heuristics. A varying-K check is reported in Appendix B.6: performance is not monotonic in the Parametric-only setting, but FCPRAG-Combine improves over PRAG-Combine at $K = 4$ and $K = 5$.

5.2 Ablation: Random Fusion Controller

To verify that gains come from *learned* fusion rather than architectural side effects, we replace the Fusion Controller with a random (untrained) one while keeping the same parametric-fusion pipeline. Table 3 (Qwen2.5-1.5B and Llama-3-8B) shows that the random controller underperforms FCPRAG across subtasks under both Parametric-only and Combine settings. This indicates that improvements are not explained by adding an extra module or changing the merge operator; they come from learning retrieval-conditioned fusion weights (and calibration) that map evidence to meaningful fusion decisions.

Merge-aware supervision. The controller also depends on how fusion labels are constructed. Holding the scoring projector fixed on HotpotQA (Qwen2.5-1.5B, Parametric-only), replacing isolated single-adapter labels with our merge-aware labels improves macro F1 from 32.47 to 33.91, with the largest gain on comparison questions (+2.63). This links the leave-one-out construction in Section 3.4 to downstream fusion quality; Appendix B.5 reports the full label-source ablation, attribution diagnostics, and offline cost.

5.3 Analysis: Where Does Fusion Control Help?

The four 2Wiki subtasks stress different reasoning patterns and evidence usage. In bridge/comparison questions, a single passage can dominate the answer, and uniform fusion may over-inject irrelevant passage-specific updates. In compositional/inference questions, multiple pieces of evidence may be required, and the system must balance contributions without amplifying conflicts. FCPRAG addresses both regimes by (i) learning to emphasize dominant experts when confident (high g and low τ) and (ii) falling back toward uniform fusion when signals are ambiguous (low g), improving robustness.

Calibration replaces grid search. Figures 2–3 show that learned calibration can replace per-model global tuning. The score-only global baseline is sensitive to blend/temperature choices, whereas

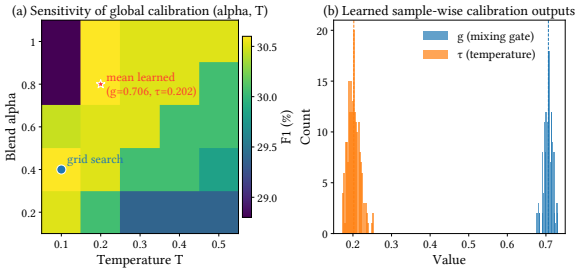

Figure 2: Calibration diagnostics on HotpotQA. **(a)** F1 over global blend α and temperature T for a score-only fusion baseline. **(b)** Learned sample-wise calibration outputs (g, τ) from FCPRAG.

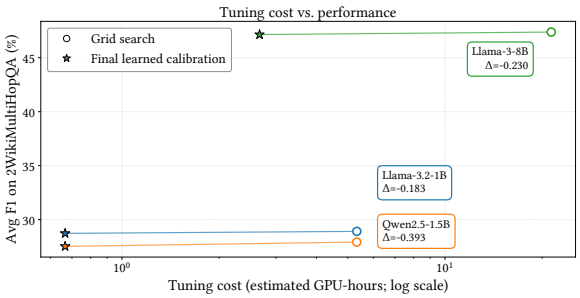

Figure 3: Tuning cost vs. performance on 2WikiMultiHopQA (Combine): each point is macro-average F1 over the four 2Wiki subtasks. Circles denote grid-selected global calibration; stars denote FCPRAG's learned calibration.

FCPRAG predicts sample-wise (g, τ) that avoid degenerate pure-uniform or overly sharp regimes and remain in a stable high-performing basin. This is consistent with Proposition 1: fixed dataset-level calibration is brittle under heteroscedastic retrieval uncertainty. Although the learned mean (g, τ) does not exactly match the single best global (α, T) cell, it stays in the same high-performance region, yielding strong F1 without exhaustive tuning. On 2Wiki in the *Combine* setting, learned calibration achieves comparable performance to grid-selected settings with substantially lower tuning cost, without changing the retriever, merge operator, or offline passage parameterization.

5.4 Robustness to Retrieval Perturbations

Real retrieval systems can return irrelevant or redundant passages, which is especially harmful when evidence is injected through adapter weights. We test retrieval noise at $K = 3$ with two one-passage perturbations: **replace-one** substitutes an irrelevant passage, and **repeat-one** duplicates a retrieved passage. Figure 4 shows that FCPRAG retains more relative performance than PRAG under both perturbations, consistent with conservative fusion when retrieval signals are unreliable. Because each panel is normalized by the method's own unperturbed score, the figure emphasizes relative performance retention rather than absolute score differences. This matches the role of adaptive calibration: noisier retrieved sets should shift fusion toward less aggressive injection instead of applying a fixed global temperature.

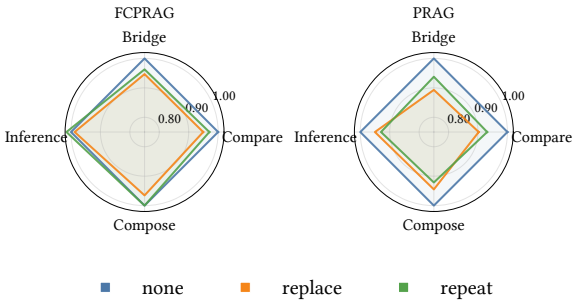

Figure 4: Retrieval perturbation robustness on 2WikiMultiHopQA (Llama-3-8B, Parametric-only; $K = 3$). We report **replace-one** and **repeat-one** perturbations, with each panel normalized by the method's own unperturbed score (none = 1).

6 Conclusion

We presented FCPRAG, a learned fusion controller for stable, retrieval-conditioned multi-passage LoRA fusion in parametric retrieval augmented generation. By predicting per-passage weights and per-sample calibration, FCPRAG improves robustness and reduces the need for expensive global calibration. Across four datasets, 2WikiMultiHopQA, HotpotQA, PopQA, and ComplexWebQuestions(CWQ) and three LLM backbones, FCPRAG improves F1 scores over the baseline methods in most settings, while a random controller performs significantly worse. These results suggest that sample-level calibration is a practical mechanism for adapting evidence fusion to query-dependent retrieval uncertainty, rather than relying on a single global setting across tasks and backbones. Because the method operates at the fusion stage, it can strengthen multi-passage parametric RAG with minimal changes to the underlying retrieval, offline parameterization, and generation pipeline. These results suggest that learned fusion control is a promising and practical direction for parametric RAG systems under multi-passage retrieval and parametric fusion.

Limitations

Offline label cost: merge-aware supervision requires evaluating one all-merge and K leave-one-out merges per training sample (e.g., four forward passes for $K = 3$), increasing offline computation cost and making larger K more expensive. **Embedding dependence:** the controller operates on query/passage embeddings; if embedding similarity is misaligned with actual answer utility, fusion predictions may degrade. **Scaling to larger evidence sets:** we focus on a small, fixed number of passages per query; extending to larger K may require sparse or hierarchical fusion policies.

Ethical Considerations

This work studies a retrieval-augmented QA system using public benchmarks and does not involve new human-subject data collection. The main ethical risk is that parametric injection can make retrieved errors or biased evidence more persistent in the model's effective parameters during inference. FCPRAG mitigates this risk by learning to down-weight weak or conflicting evidence, but it does not guarantee factual correctness or fairness. Deployments should therefore pair the method with source filtering, answer attribution, and task-specific safety evaluation before use in high-stakes settings.

References

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. 2022. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

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

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

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

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Gautier Izacard and Edouard Grave. 2021. 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*, pages 874–880.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 1(2):4.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

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

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2022. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

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

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

D Lepikhin, H Lee, Y Xu, D Chen, O Firat, Y Huang, M Krikun, N Shazeer, and Z Gshard. 2020. Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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

Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M Rush. 2024. A controlled study on long context extension and generalization in llms. *arXiv preprint arXiv:2409.12181*.

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

Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 487–503.

John Platt and 1 others. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. 2023. Zipit! merging models from different tasks without training. *arXiv preprint arXiv:2305.03053*.

Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. 2025a. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1240–1250.

Zhan Su, Fengran Mo, and Jian-yun Nie. 2025b. Parametric retrieval-augmented generation using latent routing of lora adapters. *arXiv preprint arXiv:2511.17044*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Yuqiao Tan, Shizhu He, Huanxuan Liao, Jun Zhao, and Kang Liu. 2025. Dynamic parametric retrieval augmented generation for test-time knowledge enhancement. *arXiv preprint arXiv:2503.23895*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model

soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *arXiv preprint arXiv:2402.14116*.

A Proof of Proposition 1

For a given query q, retrieval returns passages $\mathcal{P} = \{p_1, \ldots, p_N\}$, each associated with a passage-specific LoRA adapter Δ_i. A scoring model assigns per-passage scores $\mathbf{s} = (s_1, \ldots, s_N)$, which may vary across queries. We assume scores are non-degenerate, i.e., $\Pr[\exists\, i \neq j : s_i \neq s_j] > 0$; otherwise $\mathbf{w}(T;\, \mathbf{s})$ is uniform for all T and temperature has no effect. We allow the score distribution $\mathbf{s} \mid t$ to differ across query types. The proof proceeds by expanding the risk-aware objective, showing that the optimal temperature depends monotonically on the noise variance, and concluding that heteroscedastic types force distinct optima that no single temperature can satisfy. Fusion produces a merged update

$$\Delta(T;\, \mathbf{s}) = \sum_{i=1}^{N} w_i(T;\, \mathbf{s})\, \Delta_i, \tag{13}$$

where weights are given by a softmax with temperature $T > 0$:

$$w_i(T;\, \mathbf{s}) = \frac{\exp(s_i/T)}{\sum_{j=1}^{N} \exp(s_j/T)}. \tag{14}$$

Note that even under a *global* (dataset-level) temperature T, the weights $\mathbf{w}(T;\, \mathbf{s})$ are sample-dependent through the scores $\mathbf{s}$. We show that allowing the temperature to also adapt per sample can strictly reduce the expected calibration regret under heteroscedastic uncertainty.

We model the effective utility of adapter i on a given sample as

$$U_i = \mu_i + \varepsilon_i, \tag{15}$$

where μ_i is the mean contribution (independent of query type t) and ε_i is zero-mean noise induced by retrieval uncertainty.

Assumptions.

(A1) *Independence.* ε_i are independent across passages, with $\mathrm{Var}(\varepsilon_i \mid \text{type } t) = \sigma_t^2$.

(A2) *Non-degenerate heteroscedastic mixture.* There exist at least two query types A, B with $\sigma_A^2 \neq \sigma_B^2$, each occurring with positive probability ($0 < \pi < 1$, where $\pi = \Pr[t{=}A]$).

(A3) *Score–utility alignment (local).* For each type t, the expected score–utility interaction $\mathbb{E}_{\mathbf{s}|t}\left[\sum_i \frac{\partial w_i(T;\mathbf{s})}{\partial T}\, \mu_i\right]$ is strictly decreasing in T on an interval containing both $T_A^\star$ and $T_B^\star$; intuitively, increasing T shifts mass from higher-utility to lower-utility passages on average.

(A4) *Interior optimum.*
   (a) For each type t, the expected objective $\mathbb{E}_{\mathbf{s}|t}[\mathcal{R}_t(\mathbf{w}(T;\, \mathbf{s}))]$ admits a unique maximizer $T_t^\star \in (0, \infty)$.
   (b) The derivative term $\mathbb{E}_{\mathbf{s}|t}\left[\sum_i w_i \frac{\partial w_i}{\partial T}\right]$ is non-zero and does not change sign in a neighborhood of $T_t^\star$.

Remark on (A1). The independence assumption simplifies the variance term; allowing correlations adds cross-terms $\sum_{i \neq j} w_i w_j \,\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$ but the variance remains type-dependent through σ_t^2, so the monotonicity argument in §A.1 still applies and Proposition 1 holds. In particular, positive correlations among passage noises can further motivate conservative (near-uniform) fusion under high uncertainty.

We adopt a mean–variance objective as a tractable proxy that captures the key tension between evidence sharpness and noise amplification; the empirical sections validate the conclusion under actual QA metrics. We define the risk-aware objective for fusion as

$$\mathcal{R}_t(\mathbf{w}) = \mathbb{E}\left[\sum_i w_i\, U_i\right] - \lambda\, \mathrm{Var}\left(\sum_i w_i\, U_i\right), \quad \lambda > 0. \tag{16}$$

The per-type optimal temperature maximizes the expected risk-aware objective:

$$T_t^\star \;\in\; \arg\max_{T>0}\; \mathbb{E}_{\mathbf{s}|t}\Big[\mathcal{R}_t\big(\mathbf{w}(T;\, \mathbf{s})\big)\Big], \tag{17}$$

which is unique by assumption (A4a). Note that $T_t^\star$ is a type-level constant (not a function of $\mathbf{s}$).

By linearity of expectation,

$$\mathbb{E}\left[\sum_i w_i\, U_i\right] = \sum_i w_i\, \mu_i. \tag{18}$$

By independence (A1),

$$\mathrm{Var}\left(\sum_i w_i\, U_i\right) = \sigma_t^2 \sum_i w_i^2. \tag{19}$$

Thus,

$$\mathcal{R}_t(\mathbf{w}) = \sum_i w_i\, \mu_i - \lambda\, \sigma_t^2 \sum_i w_i^2. \tag{20}$$

A.1 Effect of Temperature and Retrieval Uncertainty

The softmax temperature T controls the concentration of $\mathbf{w}(T;\,\mathbf{s})$:

- As $T \downarrow 0$, $\mathbf{w}(T;\,\mathbf{s})$ becomes increasingly peaked (approaching one-hot), and $\sum_i w_i^2$ increases.
- As $T \uparrow \infty$, $\mathbf{w}(T;\,\mathbf{s})$ approaches the uniform distribution, and $\sum_i w_i^2$ decreases to $1/N$.

This reveals the core trade-off: decreasing T increases both the alignment with high-μ_i passages (higher expected mean) and the variance of the fused signal (larger $\sum_i w_i^2$). The variance penalty in $\mathcal{R}_t$ is scaled by σ_t^2, so:

- When σ_t^2 is small, the penalty is weak and the optimal temperature $T_t^\star$ is small (sharp fusion).
- When σ_t^2 is large, the penalty dominates and $T_t^\star$ shifts upward (flatter fusion).

More precisely, differentiating the expected objective (Eq. (20), with $\mathbf{w} = \mathbf{w}(T;\,\mathbf{s})$) with respect to T and taking the expectation over $\mathbf{s} \mid t$ (justified by dominated convergence, since softmax derivatives are bounded) yields the first-order condition:

$$\mathbb{E}_{\mathbf{s}|t}\left[\sum_i \frac{\partial w_i}{\partial T}\,\mu_i\right] \;=\; 2\lambda\,\sigma_t^2\;\mathbb{E}_{\mathbf{s}|t}\left[\sum_i w_i \frac{\partial w_i}{\partial T}\right]. \tag{21}$$

By the score–utility alignment assumption (A3), the left-hand side is strictly decreasing in T near the optimum. The right-hand side scales linearly with σ_t^2, and the derivative term is non-zero and of constant sign near $T_t^\star$ by (A4b). By the implicit function theorem applied to (21), $T_t^\star$ is a smooth function of σ_t^2. When σ_t^2 increases, the right-hand side of (21) increases at the old $T_t^\star$, violating the first-order condition; since the left-hand side is decreasing in T, the balance point must shift to a larger T. Combined with uniqueness of the interior maximizer (A4a), the balance point $T_t^\star$ is strictly increasing in σ_t^2.

A.2 Impossibility of a Single Dataset-level Temperature

By (A2), there exist two types A and B with $\sigma_A^2 \neq \sigma_B^2$, each with positive probability. By the strict monotonicity established in §A.1, the type-level optimal temperatures (Eq. 17) satisfy

$$T_A^\star \neq T_B^\star. \tag{22}$$

Any dataset-level heuristic fixes a single temperature T for all samples. Since $T_A^\star \neq T_B^\star$ and both types occur with positive probability (A2), any fixed T differs from at least one type optimum. By uniqueness of the interior maximizer (A4a), using any $T \neq T_t^\star$ on type t yields a strict loss; that is, at least one of the following holds:

$$\begin{aligned}\mathbb{E}_{\mathbf{s}|A}\Big[\mathcal{R}_A\big(\mathbf{w}(T;\,\mathbf{s})\big)\Big] &< \mathbb{E}_{\mathbf{s}|A}\Big[\mathcal{R}_A\big(\mathbf{w}(T_A^\star;\,\mathbf{s})\big)\Big]\\ \mathbb{E}_{\mathbf{s}|B}\Big[\mathcal{R}_B\big(\mathbf{w}(T;\,\mathbf{s})\big)\Big] &< \mathbb{E}_{\mathbf{s}|B}\Big[\mathcal{R}_B\big(\mathbf{w}(T_B^\star;\,\mathbf{s})\big)\Big]\end{aligned} \tag{23}$$

(or both), so the expected calibration regret (11) is strictly positive. In particular, this holds for $T^\star_{\text{global}}$ (Eq. 12), so $\mathrm{Reg}(T^\star_{\text{global}}) > 0$. This completes the proof that sample-level temperature adaptation is strictly beneficial under heteroscedastic retrieval uncertainty. □

A.3 Oracle Weights and Connection to FCPRAG

To make the architectural connection explicit, we derive the oracle weights in the $N = 2$ case, relaxing the softmax parameterization to optimize over w directly. The resulting oracle provides a lower bound on what softmax-parameterized fusion can achieve; the first-order expansion below shows that FCPRAG approximates this oracle in the near-uniform regime. Let $w_1 = w$ and $w_2 = 1-w$, with $\Delta\mu = \mu_1 - \mu_2 > 0$. The objective (20) becomes

$$\mathcal{R}_t(w) = w\,\mu_1 + (1-w)\,\mu_2 - \lambda\,\sigma_t^2\big[w^2 + (1-w)^2\big]. \tag{24}$$

Taking the derivative and setting it to zero:

$$\Delta\mu - 2\lambda\,\sigma_t^2\,(2w-1) = 0 \quad\Longrightarrow\quad w_t^\star = \frac{1}{2} + \frac{\Delta\mu}{4\lambda\,\sigma_t^2}. \tag{25}$$

This oracle weight has the form *uniform baseline + utility-proportional deviation that shrinks with noise*, which directly motivates the (g, τ) gating mechanism: g controls the deviation magnitude and τ controls the sharpness, together playing the role of $\Delta\mu/(4\lambda\,\sigma_t^2)$. We can match this to FCPRAG's gated fusion (Eq. 3). For $N = 2$, FCPRAG produces

$$w_i^{\text{FC}} = g \cdot \mathrm{softmax}(s_i/\tau) + (1-g)\cdot \tfrac{1}{2}. \tag{26}$$

Under a first-order expansion around the near-uniform regime ($|s_1 - s_2|/\tau \ll 1$; used here for interpretability, though the learned controller is not restricted to this regime):

$$\mathrm{softmax}(s_1/\tau) \;\approx\; \frac{1}{2} + \frac{s_1 - s_2}{4\tau}, \tag{27}$$

so

$$w_1^{\mathrm{FC}} \approx \frac{1}{2} + \frac{g\,(s_1 - s_2)}{4\tau}. \tag{28}$$

Comparing with the oracle (25): under score–utility alignment, $s_1 - s_2$ is positively correlated with $\Delta\mu$ in expectation, so the oracle suggests setting g/τ to decrease with σ_t^2. Note that λ is a free parameter in the theoretical model with no direct architectural counterpart; the theory therefore provides qualitative guidance (the *direction* of adaptation) rather than quantitative calibration, while the controller learns the appropriate scale from data. In particular, the controller should increase g/τ (sharper, more confident fusion) when σ_t^2 is small and decrease it (conservative, near-uniform fusion) when σ_t^2 is large, which is precisely the adaptive behavior that FCPRAG learns from data.

Remark (observability). In practice, the Fusion Controller does not observe σ_t^2 directly; it infers signals correlated with retrieval uncertainty from the query–passage interaction features (Eq. 1). Proposition 1 establishes that such adaptation is *beneficial in principle*; whether the learned controller captures sufficient signal is an empirical question validated by our experiments. The case studies in Appendix C show that the learned g and τ values vary meaningfully across queries with different evidence quality, suggesting the controller captures relevant uncertainty signals.

A.4 Discussion

This result formally motivates predicting fusion calibration (e.g., temperature and a gate interpolating with a uniform prior) at the sample level. Such a mechanism can adapt to heteroscedastic retrieval uncertainty and avoid the inherent limitations of dataset-level fusion heuristics. We note several simplifications in the theoretical model: the mean–variance objective is a tractable proxy for downstream QA metrics, the two-type mixture is a minimal construction (the argument extends to any finite number of types with distinct noise variances), and the $N = 2$ oracle relaxes the softmax constraint. Despite these simplifications, the qualitative prediction aligns with our empirical findings: FCPRAG's gains over PRAG are most pronounced on subtasks with heterogeneous evidence quality (Table 2, Figure 4). This is exactly the regime where the calibration regret of a global policy is largest: the retrieved set mixes strongly and weakly relevant passages, amplifying the cost of a fixed temperature.

B Additional Details

B.1 Use of Large Language Models

Large language models were used for language polishing and proofreading to improve the clarity and readability of this manuscript.

B.2 Extended Related Work

In-context RAG and document aggregation. Classical RAG pipelines inject retrieved evidence by concatenating passages with the query or by aggregating document information inside the decoder (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022; Khandelwal et al., 2019). This design has been highly effective, but it also makes performance sensitive to context length, retrieval noise, and evidence conflicts, especially in multi-hop reasoning settings where useful and distracting passages may coexist (Jiang et al., 2023; Asai et al., 2024; Yang et al., 2018; Ho et al., 2020; Zhu et al., 2024; Lu et al., 2024). Methods such as RePlug further improve how retrieved documents are used without updating the backbone LLM (Shi et al., 2024). Our setting differs because the retrieved evidence is fused in *parameter space* rather than prompt space: each passage induces a candidate parametric update, and the core question becomes how strongly each update should influence the final merged adapter.

Expert routing and mixture-of-experts. Learning to select or weight components conditioned on an input has a long history in mixture-of-experts models. Sparsely-gated MoE layers (Shazeer et al., 2017) and efficient variants such as Switch Transformers (Fedus et al., 2022) use routing to activate a subset of experts per input (or token) to improve capacity and efficiency; large-scale conditional computation systems such as GShard further highlight the benefits of input-dependent expert selection (Lepikhin et al., 2020). While our Fusion Controller is not an MoE layer inside the backbone LLM, it plays an analogous role at the *evidence level*: it produces retrieval-conditioned weights over passage-specific parametric updates. This difference in granularity matters: evidence-level fusion must decide not only which updates are helpful, but also how *confidently* to deviate from uniform fusion under retrieval noise and evidence conflicts.

Calibration and re-weighting. Calibration methods such as Platt scaling and temperature scaling (Platt et al., 1999; Guo et al., 2017; Niculescu-Mizil and Caruana, 2005; Kuleshov et al., 2018) adjust predictive distributions to better reflect uncertainty. In PRAG-family fusion, mapping controller scores to merge weights often requires careful temperature and blending choices. FCPRAG integrates calibration into the fusion controller by learning a temperature and a mixing gate that blends learned weights with a uniform prior, yielding stable fusion without expensive global tuning. This perspective aligns calibrated fusion with classical calibration: the goal is not only to rank evidence, but to control how *confidently* the system deviates from uniform fusion when score signals are unreliable. More broadly, reducing tuning and selection overhead is a recurring theme in RAG systems: hyperparameters that work well for one backbone or dataset can shift under changes in retrievers, context length, or evidence noise. Our diagnostic analysis therefore treats calibration as an explicit design target rather than an afterthought.

B.3 Inference Pseudocode

Algorithm 1 Fusion-controlled parametric RAG inference (FCPRAG).

Require: Query q; retrieved passages $\mathcal{P} = \{p_i\}_{i=1}^{N}$; passage adapters $\{\Delta_i\}_{i=1}^{N}$; Fusion Controller f_θ; frozen backbone LLM $\mathcal{M}$; weighted adapter merger $\mathcal{M}_{\text{WAM}}(\cdot)$.

Ensure: Generated answer $\hat{a}$.

Compute query embedding $\mathbf{e}_q$ and passage embeddings $\{\mathbf{e}_{p_i}\}$ using the sentence encoder.
for $i \leftarrow 1$ to N **do**
  Build interaction feature $\mathbf{x}_i = [\mathbf{e}_q;\ \mathbf{e}_{p_i};\ \mathbf{e}_q \odot \mathbf{e}_{p_i};\ |\mathbf{e}_q - \mathbf{e}_{p_i}|]$.
end for
Predict fusion scores $\{s_i\}$ and per-sample calibration (g, τ) using f_θ.
$\pi \leftarrow \text{softmax}(\{s_i/\tau\}_{i=1}^{N})$.
for $i \leftarrow 1$ to N **do**
  $w_i \leftarrow g \cdot \pi_i + (1-g) \cdot \frac{1}{N}$.
end for
for $i \leftarrow 1$ to N **do**
  $\tilde{w}_i \leftarrow \frac{N\, w_i}{\sum_{j=1}^{N} w_j + \epsilon}$.
end for
$\Delta^\star \leftarrow \mathcal{M}_{\text{WAM}}(\{\Delta_i\}, \{\tilde{w}_i\})$
Inject $\Delta^\star$ into $\mathcal{M}$ to obtain adapted model $\mathcal{M}_{\Delta^\star}$.
if Combine mode **then**
  Build the in-context prompt with q and passages $\mathcal{P}$.
else
  Build the prompt with q only (**Parametric-only**).
end if
$\hat{a} \leftarrow \textsc{Generate}(\mathcal{M}_{\Delta^\star}, \text{prompt})$.
return $\hat{a}$.

B.4 Reproducibility Details

All experiments were conducted on 2× NVIDIA H100 GPUs. We evaluate three instruction-tuned backbones: Llama-3.2-1B, Qwen2.5-1.5B, and Llama-3-8B (Touvron et al., 2023; Qwen et al., 2025). All compared methods share the same retriever and use top-$K = 3$ passages per query, matching the default setting in PRAG-family evaluations. We use `bge-base-en-v1.5` from Sentence-Transformers for query and passage embeddings. The Fusion Controller consists of two independent networks: a 3-layer scoring network (hidden dimensions: 2048, 1024) that predicts per-passage fusion scores, and a 2-layer calibration network

(hidden dimension: 256) that predicts per-sample calibration parameters (g, τ). Both networks use ReLU activations and dropout rate 0.1. Training uses AdamW optimizer with learning rate 10^{-4}, weight decay 0.01, batch size 32, and 10 epochs. Following PRAG (Su et al., 2025a), we use LoRA rank $r = 2$, scaling factor $\alpha = 32$, and dropout $= 0$. For the multi-hop benchmarks, we follow a fixed-subset evaluation protocol per subtask and report subtask-level F1 together with the 2Wiki macro average. For retrieval perturbation analysis, we corrupt one of the three retrieved passages via **replace-one** or **repeat-one**, and measure relative performance retention in the Parametric-only setting.

Runtime and overhead. The Fusion Controller operates on fixed-size query and passage embeddings and adds only a small MLP forward pass at inference time. Unlike global fusion calibration via grid search, FCPRAG learns calibration parameters directly from data without backpropagating through the backbone LLM.

B.5 Merge-aware Label Construction and Analysis

Figure 5 gives a schematic view of the leave-one-out label construction process described in Section 3.4.

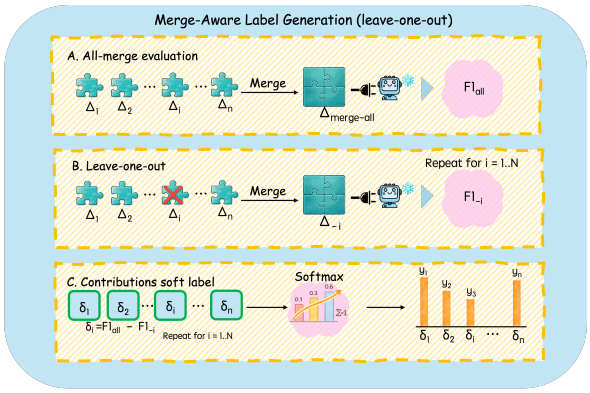

Figure 5: Merge-aware label construction. For each sample, compute F_1 with all adapters merged and with each adapter removed; marginal contributions are softmax-normalized into a soft label distribution y that supervises fusion-weight prediction. This procedure uses training data only.

To test whether merge-aware labels provide useful supervision beyond isolated adapter utility, we compare two label sources while keeping the scoring projector, backbone, and Parametric-only evaluation setting fixed. Table 4 shows that merge-aware labels improve macro F1 over single-adapter labels on HotpotQA.

Table 4: Label-source ablation on HotpotQA (Qwen2.5-1.5B, Parametric-only).

Label source	Bridge	Comparison	Macro
Single-adapter labels	17.71	47.24	32.47
Merge-aware labels	**17.95**	**49.87**	**33.91**
Δ	+0.24	+2.63	+1.43

We also compare isolated single-adapter utility rankings with the merge-aware leave-one-out rankings. Table 5 indicates that isolated adapter behavior is a weak proxy for the full merged ranking, even when the top contributor often agrees. For N retrieved adapters, label construction requires one all-merge evaluation and N leave-one-out evaluations per training sample, i.e., $N + 1$ merged evaluations; this cost is offline, uses training data only, and the resulting labels can be reused across controller variants.

Table 5: Agreement between isolated single-adapter utility and merge-aware leave-one-out labels on HotpotQA (Qwen2.5-1.5B).

Metric	Bridge	Comparison	Overall
Avg. Spearman ρ	0.0098	0.0157	0.0127
Top-1 match (%)	76.00	58.50	67.25

B.6 Varying Number of Retrieved Passages

Table 6 reports a sensitivity check on 2WikiMultiHopQA with Qwen2.5-1.5B when the number of retrieved passages varies. We include this analysis to test behavior beyond the default $K = 3$ setting. Parametric-only performance is not monotonic as K increases, reflecting the difficulty of fusing more passage-specific updates without also injecting more noise. In the Combine setting, however, FCPRAG improves over PRAG-Combine at $K = 4$ and $K = 5$.

Table 6: Varying-K results on 2WikiMultiHopQA with Qwen2.5-1.5B (F1, %).

Method	$K = 3$	$K = 4$	$K = 5$
Standard RAG	25.97	28.58	28.82
PRAG	28.05	27.05	27.01
PRAG-Combine	26.04	31.12	33.90
FCPRAG	**29.87**	26.97	28.21
FCPRAG-Combine	27.38	**33.63**	**35.08**

B.7 Dataset Overview

Table 7. Dataset overview and subtask structure.

Dataset	Domain / Focus	Subtasks
2WikiMultiHopQA	Wikipedia-based multi-hop QA	Bridge-comparison; Comparison; Compositional; Inference
HotpotQA	Wikipedia-based multi-hop QA	Bridge; Comparison
PopQA	Entity-centric open-domain QA	Single task
CWQ	Web-based complex QA	Single task

B.8 Training-side Prompts

This appendix lists the key prompt templates used to construct synthetic supervision and train passage-LoRA adapters offline. Placeholders such as `{passage}` and `{question}` indicate runtime fields.

Passage rewrite (augmentation).

Training prompt: passage rewrite

Rewrite the following passage. While keeping the entities, proper nouns, and key details such as names, locations, and terminology intact, create a new version of the text that expresses the same ideas in a different way. Make sure the revised passage is distinct from the original one, but preserves the core meaning and relevant information. passage

QA-pair generation from a passage (augmentation).

Training prompt: passage-to-QA generation

I will provide a passage of text, and you need to generate three different questions based on the content of this passage. Each question should be answerable using the information provided in the passage. Additionally, please provide an appropriate answer for each question derived from the passage. You need to generate the question and answer in the following format: ["question": "What is the capital of France?", "answer": "Paris", "full_answer": "The capital of France is Paris."] This list should have at least three elements. You only need to output this list in the above format. Passage: passage

Passage-LoRA fine-tuning prompts (offline parameterization). Given QA pairs derived from passage p_i, we fine-tune a passage-specific LoRA adapter Δ_i using a mix of passage-conditioned and question-only prompts:

Training prompt: passage-conditioned and question-only fine-tuning

(A) Passage-conditioned: You should answer the question by referring to the knowledge provided below and integrating your own knowledge. Passage 1: passage
Question: question
(B) Question-only (no passage in context): You should answer the question by referring to the knowledge provided below and integrating your own knowledge.
Question: question

C Case Studies

This section provides one real case per dataset from our main evaluation setting. Each case reports the question, retrieved passages, rendered prompt structure, model output, and fusion signals.

2WikiMultiHopQA (Llama-3-8B): real case (combined inference)

Question.
`Which film came out earlier, Morecambe Church Lads' Brigade At Drill or Little Funny Guy?`

Retrieved passages (top-K).
`Passage 1: ...Church Lads' and Church Girls' Brigade ... origins in 1891 ...`
`Passage 2: ...Church Lads' and Church Girls' Brigade ... historical notes ...`
`Passage 3: ...Boys' Life Brigade ... related youth-brigade timeline ...`

Rendered prompt (combined setting).

```
You should reference the knowledge
provided below and combine it with
your own knowledge to answer the
question. Please follow the format
of the example I provided above.
Here are some examples about how to
answer the questions.
Question: When did the director of
film Hypocrite (Film) die?
Answer: The film Hypocrite was
directed by Miguel Morayta. Miguel
Morayta died on 19 June 2013. So the
answer is 19 June 2013.
Here are some reference.
Passage 1: ...Church Lads' and Church
Girls' Brigade...
Passage 2: ...historical notes on
Church Lads' Brigade...
Passage 3: ...Boys' Life Brigade and
related organizations...
Let's think step by step. Answer the
questions in the same format as above.
```

Question: Which film came out earlier, Morecambe Church Lads' Brigade At Drill or Little Funny Guy?

Model output and evaluation.

Generated text: ...Morecambe Church Lads' Brigade At Drill (1908) vs Little Funny Guy (1912)...
Eval answer: Morecambe Church Lads' Brigade At Drill
Gold answer: Morecambe Church Lads' Brigade at Drill
EM/F1: 1 / 1.0

Fusion signals (projector-enabled run).

scores: [0.5088, 0.5350, 0.5331]
g: 0.8357, tau: 0.0988
weights: [0.8644, 1.0767, 1.0588]

HotpotQA (Qwen2.5-1.5B): real case (combined inference)

Question.

Who has a wider scope of profession, José Echegaray or Graham Swift?

Retrieved passages (top-K).

Passage 1: María Catalina Irigoyen Echegaray ... Spanish Roman Catholic professed religious ...
Passage 2: ...Beatification record of María Catalina Irigoyen Echegaray ...
Passage 3: ...José Echegaray taught mathematics/hydraulics ... also served in government ...

Rendered prompt (combined setting).

You should reference the knowledge provided below and combine it with your own knowledge to answer the question. Please follow the format of the example I provided above.
Here are some examples about how to answer the questions.
Question: Jeremy Theobald and Christopher Nolan share what profession?
Answer: Jeremy Theobald is an actor and producer. Christopher Nolan is a director, producer, and screenwriter. Therefore, they both share the profession of being a producer. So the answer is producer.
Here are some reference.
Passage 1: ...María Catalina Irigoyen Echegaray...
Passage 2: ...beatification context...
Passage 3: ...José Echegaray's teaching and political roles...
Let's think step by step. Answer the questions in the same format as above.
Question: Who has a wider scope of profession, José Echegaray or Graham Swift?

Model output and evaluation.

Generated text: ...José Echegaray was an engineer/politician/teacher, while Graham Swift is a novelist...
Eval answer: José Echegaray
Gold answer: José Echegaray
EM/F1: 1 / 1.0

Fusion signals (projector-enabled run).

scores: [0.4886, 0.4649, 0.4761]
g: 0.8178, tau: 0.0817
weights: [1.1235, 0.8864, 0.9901]

PopQA (Llama-3-8B): real case (combined inference)

Question.

What is Brigitte Bardot's occupation?

Retrieved passages (top-K).

Passage 1: ...Fondation Brigitte Bardot ... animal protection foundation ...
Passage 2: ...Brigitte Bardot Foundation ... created by French actress Brigitte Bardot ...
Passage 3: ...Brigitte Bardot ... French actress, singer, dancer, and fashion model ...

Rendered prompt (combined setting).

You should reference the knowledge provided below and combine it with your own knowledge to answer the question. Please follow the format of the example I provided above.
Here are some examples about how to answer the questions.
Question: When did the director of film Hypocrite (Film) die?
Answer: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is 19 June 2013.
Here are some reference.
Passage 1: ...Fondation Brigitte Bardot...
Passage 2: ...French actress Brigitte Bardot...
Passage 3: ...Brigitte Bardot biography...
Let's think step by step. Answer the questions in the same format as above.
Question: What is Brigitte Bardot's occupation?

Model output and evaluation.

Generated text: ...Brigitte Bardot is a French actress, singer, dancer, and fashion model...
Eval answer: actress
Gold answer: [actor, actress, ...]
EM/F1: 1 / 1.0

Fusion signals (projector-enabled run).
scores: [0.5344, 0.4883, 0.5225]
g: 0.7373, tau: 0.1661
weights: [1.0854, 0.8860, 1.0286]

ComplexWebQuestions (Llama-3-8B): real case (combined inference)

Question.
What countries border the country whose capital is Sucre?

Retrieved passages (top-K).
Passage 1: Sucre ... constitutional capital of Bolivia...
Passage 2: What a Country! ... (distractor passage)
Passage 3: Capital Country ... New South Wales tourism region... (distractor)

Rendered prompt (combined setting).

```
You should reference the knowledge provided below and combine it with your own knowledge to answer the question. Please follow the format of the example I provided above.
Here are some examples about how to answer the questions.
Question: When did the director of film Hypocrite (Film) die?
Answer: The film Hypocrite was directed by Miguel Morayta. Miguel Morayta died on 19 June 2013. So the answer is 19 June 2013.
Here are some reference.
Passage 1: Sucre... constitutional capital of Bolivia...
Passage 2: What a Country!...
Passage 3: Capital Country...
Let's think step by step. Answer the questions in the same format as above.
Question: What countries border the country whose capital is Sucre?
```

Model output and evaluation.
Generated text: ...countries bordering Bolivia are Brazil, Argentina, Paraguay, and Chile.
Eval answer: Brazil
Gold answer: [Brazil, Argentina, Paraguay, Chile, ...]
EM/F1: 1 / 1.0

Fusion signals (projector-enabled run).
scores: [0.5126, 0.5649, 0.5041]
g: 0.7724, tau: 0.1105
weights: [0.8840, 1.2808, 0.8352]