LAM-RAG: LoRA Messages for Multi-Document Retrieval-Augmented Generation

Anonymous authors
Paper under double-blind review

Abstract

Multi-agent debate is built to reach consensus. Multi-document RAG often needs the opposite: when evidence is ambiguous or conflicting, different documents support different correct answers, all of which should be kept. Transcript-based debate works against this goal: from the same local answers, transcript exchange lowers answer-set F1 below the no-exchange level as agents copy their peers' answers and supported alternatives disappear. We propose LAM-RAG (LoRA-as-Message RAG), which keeps local answers and final aggregation in text but carries inter-agent influence as transient LoRA messages. A frozen translator maps each answer state to low-rank factors that a reliability-aware controller prunes and scales. Because a LoRA message exposes no copyable peer answers, it couples agents weakly: peer-answer adoption drops sharply and distinct supported answers survive. Yet the messages still carry useful information: on Qwen2.5-1.5B, one round of exchange raises F1 by 1.5–2.5 points over no exchange, and zero, random, or cross-question messages lower it by 1.3–1.9. On RAMDocs, AmbigDocs, and FaithEval with three backbones, LAM-RAG beats the strongest text-exchange or single-pass baseline by 2.86–5.20 average F1. Under added distractors or peers, LoRA messages degrade less than peer text.

1 Introduction

Retrieval-augmented generation (RAG) improves factuality by conditioning generation on external evidence (Lewis et al., 2020; Gao et al., 2023). The setting becomes harder when a question is paired with several retrieved documents that contain duplicate facts, partial answers, entity aliases, distractors, or direct contradictions (Shi et al., 2023; Yoran et al., 2024; Xie et al., 2024; Jin et al., 2024). For ambiguous or conflicting questions, the correct output is often a *set*: each supported answer may come from a different document. A system must therefore keep every supported answer while discarding unsupported ones.

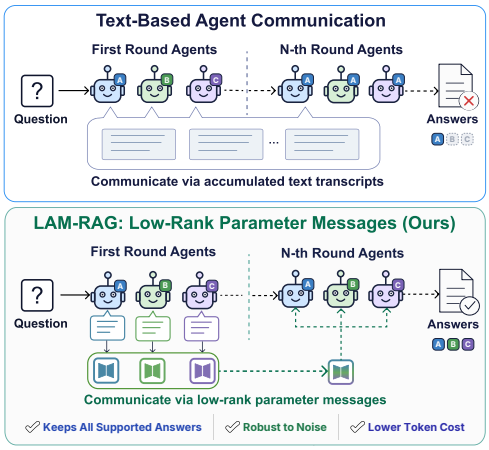

Figure 1: Transcript debate lets agents copy peer answers (tags A/B/C), which collapses supported alternatives. LAM-RAG exchanges transient LoRA messages, keeps final answers in text, and retains every supported answer.

Multi-agent debate looks like a natural fit: assign one agent to each document and let the agents exchange their answers (Du et al., 2024; Liang et al., 2024; Wang et al., 2025b). But debate is designed to reach consensus, and recent analyses find that exchanging arguments alone does not improve expected correctness (Choi et al., 2025b; Smit et al., 2024). In set-valued QA, consensus is the wrong objective. Starting from the same independently generated local answers, transcript exchange on Qwen2.5-1.5B lowers answer-set F1 below the no-exchange level at every depth: 11–66% of agents switch to another agent's answer, and supported alternatives disappear (Figure 1; Section 4.3). Transcripts are a strong coupling channel, because a receiver sees its peers' answers verbatim and can copy them.

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.

This raises a question about the carrier rather than the protocol: can agents inform one another without this pull toward consensus? Our idea is to separate *answer expression* from *inter-agent influence*. Text expresses local answers and the final answer set, while influence travels as transient low-rank parameter messages. A *LoRA message* is predicted from an agent's textual answer state, acts only during the following round, and exposes no copyable peer answers, so it couples agents weakly; it can also be composed and scaled directly without growing the receiver context.

We instantiate this idea in **LAM-RAG** (LoRA-as-Message RAG). Each document agent first answers from its local evidence. A frozen DyPRAG-style Text2LoRA translator (Tan et al., 2025) maps each answer state into a transient low-rank message; messages are composed in LoRA factor space and injected into the next round without modifying the stored backbone. A reliability-aware controller combines local evidence support with answer confidence to prune low-reliability messages and scale retained ones. Final aggregation returns to text, so the answer set remains explicit and auditable.

Across RAMDocs, AmbigDocs, and FaithEval with three instruction-tuned backbones, LAM-RAG attains the highest average F1 among text-exchange and single-pass baselines. A communication-depth analysis on Qwen2.5-1.5B shows why: from the same no-exchange starting point, transcript exchange lowers F1 at every depth, while LoRA-message exchange raises it by 1.5–2.5 points, a gain that content-free messages do not give. Receiver statistics trace the difference to conformity: LoRA exchange keeps 89–95% of round-one gold answers and nearly all distinct answers (62–84% under text exchange), and stress tests with added distractors or peers agree.

Our contributions are:

- We identify conformity as a failure mode of transcript debate in multi-document RAG: across three benchmarks, text exchange raises peer-answer adoption, removes supported answers, and lowers answer-set F1 relative to no exchange.
- We propose LAM-RAG, which carries inter-agent influence as transient LoRA messages composed in factor space and regulated by reliability-aware pruning and scaling, while local and final answers stay textual.
- On three backbones and three benchmarks, this weak-coupling channel still transfers useful information, outperforms every text-exchange baseline, keeps peer adoption low and answer diversity intact, and degrades less than peer text under source-matched stress tests.

2 Related Work

Textual multi-agent debate and RAG filtering. Multi-agent debate improves reasoning by letting LLM instances critique one another (Du et al., 2024; Liang et al., 2024; Chan et al., 2024; Wang et al., 2024). RAG variants instantiate this idea with specialized roles, filters, judges, or response voting, including MADAM-RAG (Wang et al., 2025b), DRAG (Hu et al., 2025), MAIN-RAG (Chang et al., 2025), VOTE-RAG (Xie & Sun, 2026), and PROClaim (Chowdhury et al., 2026). These systems control which text is shown, filtered, or aggregated. Analyses of debate find that its gains come mostly from majority voting and that exchanging arguments alone does not improve expected correctness (Choi et al., 2025b; Smit et al., 2024). These studies consider single-answer reasoning; we show that in set-valued multi-document QA, the pull toward consensus actively removes supported answers, and we change the channel rather than the protocol.

Robust and conflict-aware RAG. RAG degrades under irrelevant or contradictory passages (Shi et al., 2023; Yoran et al., 2024; Xie et al., 2024; Jin et al., 2024). Self-RAG (Asai et al., 2024), CRAG (Yan et al., 2024), Self-Refine (Madaan et al., 2023), Astute RAG (Wang et al., 2025a), and CARE (Choi et al., 2025a) add retrieval gates, self-feedback, knowledge consolidation, or conflict-aware context handling. These methods mainly decide which evidence to retrieve, revise, trust, or expose in context. We instead fix evidence and study how agent states should influence one another.

Latent communication between agents. Several works replace natural-language messages with continuous signals: CIPHER debates through expected token embeddings (Pham et al., 2024), activation communication merges intermediate activations across models (Ramesh & Li, 2025), Cache-to-Cache projects one model's KV cache into another's (Fu et al., 2026), and LatentMAS shares latent working memory (Zou et al., 2026). These methods target reasoning or general-purpose

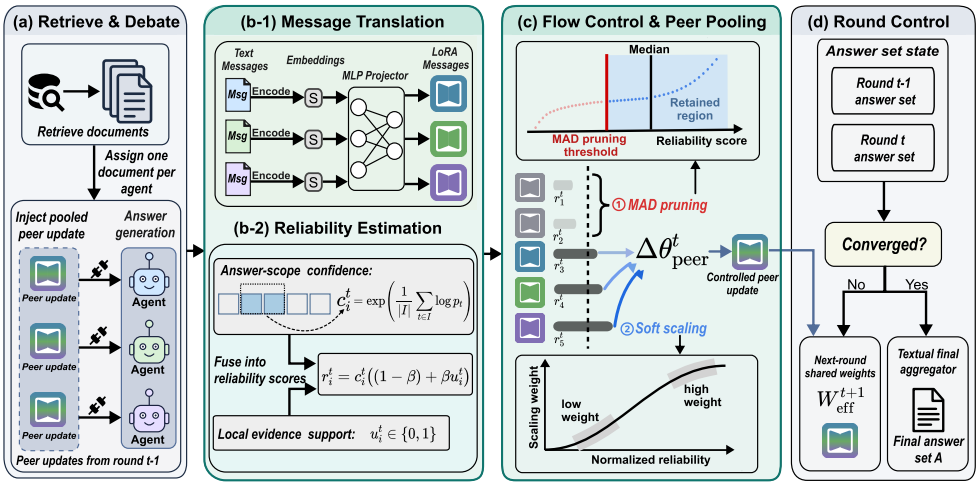

Figure 2: LAM-RAG inference loop. Document-specific agents produce textual answer states, which a frozen Text2LoRA translator maps to transient LoRA factor pairs. Flow control prunes messages and differentially scales retained factors before factor-space pooling into a shared next-round update; final answer enumeration remains textual.

collaboration and mainly argue for bandwidth or efficiency. We study a different property in multi-document RAG: a parameter-space message does not expose copyable peer answers, which reduces conformity when different documents support different answers.

Parametric RAG, LoRA composition, and continuous control. Parametric approaches encode information into model parameters or continuous prompts rather than only visible text. Test-time training, hypernetworks, LoRA, model merging, LoRA composition, and document-level parameterization show that weight-space representations can store or compose task information (Sun et al., 2020; Ha et al., 2017; Hu et al., 2022; Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Huang et al., 2024; Zhang et al., 2023; Wan et al., 2024; Tan et al., 2025). Concurrent work TFlow (Bao et al., 2026) also uses transient LoRA perturbations as a weight-space communication channel, mapping sender hidden states through a learned parameter generator and fusing complete per-sender perturbations at the receiver. LAM-RAG instead focuses on multi-round, multi-document evidence reconciliation with a frozen Text2LoRA interface, factor-space message composition, reliability-aware influence control, and textual final aggregation. The two systems also make different composition choices: TFlow avoids cross-sender products by fusing complete perturbations, whereas our canonical FactorMean averages LoRA factors before multiplication; Appendix D.4 isolates this operator choice without claiming universal superiority.

3 Method

3.1 Problem Setup and Framework Overview

LAM-RAG has four stages: local answer generation, answer-state translation into a LoRA message, reliability-aware flow control, and textual final aggregation. Figure 2 gives the system overview; Appendix A.1 provides the full inference algorithm and fixed configuration. Given a question q and documents $\mathcal{D} = \{d_1, \ldots, d_N\}$, LAM-RAG assigns one agent to each document. At round t, agent i uses W_{eff}^t with (q, d_i) to emit a textual answer state $x_i^t = (a_i^t, e_i^t)$, containing a proposed answer and supporting rationale.

The encoder maps each answer state to LoRA factors, the controller selects and scales messages, and the pooled factors affect only the next-round computation. Canonical receivers share this temporary pooled update while retaining distinct local documents and histories; stored backbone weights are restored after generation. LLM-Debate, MADAM-RAG, and LAM-RAG use the same one-document-per-agent assignment; Appendix A discusses this granularity choice and the final-source rule. The

receiver-specific peers-minus-self construction used in the source-matched carrier diagnostic is a diagnostic variant, not the canonical topology.

3.2 LoRA Messages via Text2LoRA

LAM-RAG separates three objects: the textual *answer state* $x_i^t = (a_i^t, e_i^t)$, the *LoRA message* emitted as module-wise factor pairs, and the *controlled pooled update* injected into the next round. The Text2LoRA translator is a frozen, backbone-matched interface fixed before benchmark evaluation; no translator checkpoint is trained, tuned, early-stopped, or selected on RAMDocs, AmbigDocs, or FaithEval. For adapted module ℓ, it emits $\phi(x_i^t) = \{(A_{\ell,i}^t, B_{\ell,i}^t)\}_\ell$. For the retained set $\mathcal{S}_t = \{i : m_i^t = 1\}$, FactorMean is

$$\bar{A}_\ell^t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} s_i^t A_{\ell,i}^t, \tag{1}$$

$$\bar{B}_\ell^t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} s_i^t B_{\ell,i}^t, \tag{2}$$

$$\Delta W_{\ell,\text{pool}}^t = \bar{B}_\ell^t \bar{A}_\ell^t, \tag{3}$$

$$W_{\ell,\text{eff}}^{t+1} = W_\ell + \alpha \Delta W_{\ell,\text{pool}}^t. \tag{4}$$

Here s_i^t is the soft factor scale and α is the injection scale. Thus the canonical system composes messages in *factor space*: it scales both LoRA factors and averages A and B separately before multiplication. It is not the arithmetic mean of per-message dense updates $B_{\ell,i}^t A_{\ell,i}^t$; expanding the FactorMean product contains cross-message terms proportional to $s_i^t s_j^t B_{\ell,i}^t A_{\ell,j}^t$. All canonical main-table results use this operator. The injected factors are removed after each generation step, so the stored backbone weights are never overwritten, and all next-round agents share the same temporary update while retaining distinct document-conditioned prompts and local histories.

Text2LoRA is a frozen DyPRAG-style translator (Tan et al., 2025): a backbone-specific hypernetwork (Ha et al., 2017) that maps the final hidden-state representation of x_i^t to LoRA pairs (Hu et al., 2022). For each adapted module, $A_{\ell,i}^t \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B_{\ell,i}^t \in \mathbb{R}^{d_{\text{out}} \times r}$. In LAM-RAG, the translator input is not a raw document but an answer-and-rationale state produced under (q, d_i); the factors can be pooled, dropped, and injected transiently without changing stored backbone weights. LAM-RAG uses this frozen, backbone-matched translator only as a fixed text-to-parameter interface; our contribution is its use for inter-agent communication with inference-time flow control. Here Text2LoRA denotes the DyPRAG parameter translator; it is distinct from Text-to-LoRA (Charakorn et al., 2025), which generates task adapters from task descriptions. Appendix A records projector provenance, artifact policy, and evaluation details.

3.3 Reliability-Aware Flow Control

Flow control has two distinct roles in the implemented FactorMean operator. MAD pruning produces a binary mask $m_i^t \in \{0, 1\}$ and therefore the retained set $\mathcal{S}_t$; soft scaling produces a nonnegative factor scale s_i^t that multiplies both $A_{\ell,i}^t$ and $B_{\ell,i}^t$ before the factor-space averaging in Eqs. 1–2. Thus s_i^t should not be interpreted as a linear coefficient on the dense update $B_{\ell,i}^t A_{\ell,i}^t$. The controller is inference-time rather than learned: MAD pruning suppresses low-reliability outliers, while soft scaling modulates the contribution of retained factors. This provides an intermediate control regime between hard selection and uniform factor averaging.

Grounded reliability score. Let c_i^t be answer-scope generation confidence and u_i^t local evidence support. Confidence is computed under the same temporary weights that generated the answer, and only over the `Answer:` span:

$$c_i^t = \exp\left(\frac{1}{|I_i^t|} \sum_{k \in I_i^t} \ell_{i,k}^t \right), \tag{5}$$

$$\ell_{i,k}^t = \log p_{W_{\text{eff}}^t}(y_{i,k}^t \mid y_{i,<k}^t,\, q,\, d_i). \tag{6}$$

where I_i^t indexes answer tokens. Confidence alone cannot distinguish grounded answers from confident hallucinations, so we anchor it with substring support against the agent's local document:

$$u_i^t = \mathbb{I}\big[\mathrm{norm}(a_i^t) \text{ is a substring of } \mathrm{norm}(d_i)\big], \tag{7}$$

where $\mathrm{norm}(\cdot)$ lowercases and strips punctuation and articles. The grounded reliability score is

$$r_i^t = c_i^t\big((1-\beta) + \beta u_i^t\big) \tag{8}$$

with a fixed mixture coefficient β shared across all reported runs. We use lexical support as the default because it is cheap, auditable, and conservative. Table 14 in Appendix E compares BGE-based semantic and hybrid support under the same controller; semantic variants do not consistently improve final F1.

MAD pruning. Let $\widetilde{r}^t$ be the round median and MAD^t the median absolute deviation. We retain message i iff

$$r_i^t \geq \widetilde{r}^t - \lambda \cdot \mathrm{MAD}^t, \tag{9}$$

yielding $m_i^t = 1$ for retained agents and $m_i^t = 0$ otherwise. MAD pruning is example-adaptive: when agents mostly agree, the threshold is conservative; when one message is a low-reliability outlier, the same rule suppresses it without requiring a fixed pruning fraction. We request a minimum of two retained messages, clipped by the number of available agents (so a one-document example retains its single message), to avoid collapsing multi-answer cases.

Soft scaling. Among retained messages, we modulate injection strength with the soft scaling rule:

$$\hat{r}_i^t = \frac{r_i^t - \widetilde{r}^t}{\mathrm{MAD}^t + \varepsilon_{\mathrm{den}}}, \tag{10}$$

and then apply

$$s_i^t = s_{\min} + (s_{\max} - s_{\min}) \cdot \sigma\left(\frac{\hat{r}_i^t}{\tau}\right), \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function. If $\mathrm{MAD}^t \leq \varepsilon_{\mathrm{scale}}$, we set $s_i^t{=}1$ for all retained messages; otherwise we use $\varepsilon_{\mathrm{den}}$ in Eq. 10. Appendix A.1 lists the fixed configuration. Soft scaling complements MAD pruning by assigning retained messages graded factor scales. Because both LoRA factors are scaled before FactorMean, s_i^t is not a linear coefficient on a per-message dense update.

Inference loop. At test time, each round alternates between local answer generation, message translation, and flow control; if another round remains, the controlled message pool is transiently injected into the next round (Appendix Algorithm 1). The default uses $T{=}2$ rounds, i.e., one LoRA-message exchange. Final answers are produced by a textual aggregator over surviving agent outputs; if the retained answer set does not change between the two rounds, the aggregator reads the first-round outputs (Appendix A).

4 Experiments

4.1 Experiment Setup

Tasks and benchmarks. We evaluate RAMDocs (Wang et al., 2025b) for noisy/conflicting evidence, AmbigDocs (Lee et al., 2024) for same-name ambiguity, and an answer-set adaptation of FaithEval's inconsistent-context split (Ming et al., 2025), where the two context-supported answers form the target set. We use Qwen2.5-1.5B/7B-Instruct (Qwen Team, 2024) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024); a fixed TriviaQA split (Joshi et al., 2017) is diagnostic only. Appendix Table 4 gives dataset sizes and document/answer statistics.

Baselines. We evaluate all methods in a matched-evidence setting: retrieved documents are fixed, so the comparison targets answer reconciliation and communication rather than end-to-end retrieval quality. Baselines include direct answering (Vanilla), RAG with concatenated retrieved documents (Lewis et al., 2020), LLM-Debate (Du et al., 2024), MADAM-RAG (Wang et al., 2025b), DyPRAG-combine (Tan et al., 2025), Self-Refine (Madaan et al., 2023), Astute-RAG (Wang et al., 2025a), DRAG (Hu et al., 2025), MAIN-RAG (Chang et al., 2025), and VOTE-RAG (Xie & Sun, 2026). VOTE-RAG is included in Table 1; DRAG, MAIN-RAG, and LatentMAS (Zou et al., 2026) are reported in Appendix Table 5; Appendix C summarizes their retained answer-stage components.

Table 1: Main results under matched evidence (EM/F1, %). Bold/underline: best/second best. Multi-round baselines use their published depth (T=3); LAM-RAG uses T=2. Supplementary baselines: Table 5.

Backbone	Method	RAMDocs		AmbigDocs		FaithEval		Avg.	
		EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑
Qwen2.5 1.5B	Vanilla	0.2	4.10	0.0	7.61	0.0	8.82	0.1	6.84
	RAG NeurIPS'20	3.4	16.09	4.3	21.25	11.7	30.19	6.5	22.51
	Self-Refine NeurIPS'23	6.2	32.58	5.0	26.05	17.7	38.18	9.6	32.27
	LLM-Debate ICML'24	9.4	40.79	6.3	39.40	24.8	53.01	13.5	44.40
	DyPRAG-combine arXiv'25	2.0	13.56	1.4	15.90	5.4	15.90	2.9	15.12
	Astute-RAG ACL'25	2.4	15.89	0.4	12.29	2.7	27.48	1.8	18.55
	VOTE-RAG arXiv'26	8.4	37.14	5.2	34.79	17.4	42.80	10.3	38.25
	MADAM-RAG COLM'25	12.6	53.99	9.4	45.08	26.9	54.65	16.3	51.24
	LAM-RAG (Ours)	**14.0**	**59.87**	**10.3**	**49.84**	**33.1**	**59.60**	**19.1**	**56.44**
Qwen2.5 7B	Vanilla	1.8	9.32	0.0	10.27	0.0	20.21	0.6	13.27
	RAG NeurIPS'20	13.2	53.07	6.6	44.58	29.8	63.68	16.5	53.78
	Self-Refine NeurIPS'23	16.0	54.89	7.2	43.95	27.6	62.17	16.9	53.67
	LLM-Debate ICML'24	15.6	52.02	10.6	46.24	20.5	58.76	15.6	52.34
	DyPRAG-combine arXiv'25	14.8	53.45	6.0	41.24	30.2	62.79	17.0	52.49
	Astute-RAG ACL'25	5.4	28.49	0.6	17.35	4.5	41.57	3.5	29.14
	VOTE-RAG arXiv'26	14.0	51.05	7.0	40.28	27.5	59.30	16.2	50.21
	MADAM-RAG COLM'25	19.2	57.95	**12.5**	52.03	28.3	62.80	20.0	57.59
	LAM-RAG (Ours)	**22.6**	**63.88**	10.5	**53.84**	**36.3**	**65.58**	**23.1**	**61.10**
Llama-3.1 8B	Vanilla	2.2	13.60	0.0	12.77	0.0	22.92	0.7	16.43
	RAG NeurIPS'20	17.6	58.69	**12.8**	**52.53**	27.0	58.16	19.1	56.46
	Self-Refine NeurIPS'23	19.0	62.74	8.6	48.38	19.5	55.87	15.7	55.67
	LLM-Debate ICML'24	12.8	51.56	9.5	42.31	19.1	49.61	13.8	47.83
	DyPRAG-combine arXiv'25	12.8	45.81	8.6	37.17	23.3	52.74	14.9	45.24
	Astute-RAG ACL'25	6.6	43.67	1.9	29.13	6.6	45.55	5.0	39.45
	VOTE-RAG arXiv'26	14.0	56.12	8.4	35.30	30.9	50.42	17.8	47.28
	MADAM-RAG COLM'25	16.8	58.02	11.9	45.96	28.8	58.74	19.2	54.24
	LAM-RAG (Ours)	**20.8**	**66.19**	10.9	49.67	**33.7**	**62.10**	**21.8**	**59.32**

Implementation details. All methods use the same pre-retrieved documents, answer-set parser, FP32 greedy decoding, and comparable generation budgets. The primary metric is token-level F1 with one-to-one greedy matching; EM credits a question only when the whole predicted answer set matches the gold set, so it is low for every method. Text-debate baselines use their canonical three rounds; LAM-RAG uses two rounds (one LoRA-message exchange), fixed backbone-matched Text2LoRA translators, and one controller configuration. No projector or controller parameter is selected on the evaluation benchmarks. Section 4.3 compares both families at every depth. Appendices A, A.1, F.3, and F.4 provide provenance, configuration, bootstrap uncertainty, and a three-trajectory nucleus-sampling check; with greedy decoding, each configuration is a single deterministic run (seed 42).

4.2 Main Results

Among text-exchange and single-pass baselines, LAM-RAG attains the highest average F1 for all three backbones (Table 1), improving on the strongest one by +5.20, +3.51, and +2.86 on Qwen2.5-1.5B, Qwen2.5-7B, and Llama-3.1-8B. It also has the highest average EM on every backbone (19.1/23.1/21.8 vs. 16.3/20.0/19.2 for the next best method), and on Qwen2.5-7B RAMDocs it processes fewer tokens than either multi-round text baseline (Section 4.7), so the gain does not come from processing more tokens. The gains are most consistent on RAMDocs and FaithEval, whose documents contain unsupported or conflicting answer clusters. AmbigDocs marks the boundary: its alternatives are usually explicit, and with Llama-3.1-8B, one-pass RAG remains stronger. The gains also narrow as the backbone gets stronger. In the other eight dataset–backbone settings LAM-RAG has the highest mean F1 (95% bootstrap intervals in Appendix F.3). Among supplementary baselines (Table 5), LatentMAS (Zou et al., 2026), which chains agents through one shared KV cache, exceeds LAM-RAG by 4.34 average F1 on Llama-3.1-8B but trails it by 14.42 on Qwen2.5-7B and gives no usable output on Qwen2.5-1.5B (Appendix C). LAM-RAG also exceeds DyPRAG-combine in all nine cells, favoring parameterized answer states over parameterized raw documents.

Table 2: Where the gain comes from (F1, %). **(a)** Depth: T=1 has no exchange; *zero msg.* replaces every message with zeros (identical to LAM-RAG at T=1). **(b)** Components: Text Debate[‡] is LLM-Debate; Text + MAD[§] prunes transcripts with the same rule (three rounds).

(a) Communication depth, Qwen2.5-1.5B

	RAMDocs			AmbigDocs			FaithEval		
Method	1	2	3	1	2	3	1	2	3
LLM-Debate	58.10	46.91	40.79	48.20	40.13	39.40	55.98	52.92	53.01
MADAM-RAG	58.10	56.57	53.99	48.20	45.35	45.08	55.98	54.69	54.65
LAM-RAG	58.38	**59.87**	59.25	47.30	49.84	**50.11**	58.04	**59.60**	59.51
zero msg.	58.38	58.10	58.41	47.30	47.55	47.51	58.04	57.65	57.65

(b) Components, RAMDocs

Variant	1.5B	7B	8B
Text Debate[‡]	40.79	52.02	51.56
Text + MAD[§]	44.02	53.82	59.10
LoRA msg. only	58.14	63.59	65.37
+ MAD	58.85	63.55	64.99
+ scaling	58.42	62.53	65.65
LAM-RAG (full)	**59.87**	**63.88**	**66.19**

4.3 Why Does Transcript Exchange Hurt?

Communication depth. Without exchange (T=1), LAM-RAG and the text-debate systems nearly coincide: all aggregate the same kind of independent local answers, and LAM-RAG differs mainly in pruning which outputs reach the aggregator (58.38 vs. 58.10 on RAMDocs; Table 2a). The systems then diverge. Transcript exchange keeps F1 below the no-exchange level at every depth on all three benchmarks: from T=1 to T=3, MADAM-RAG loses 4.11/3.12/1.33 points and LLM-Debate 17.31/8.80/2.97. One round of LoRA-message exchange instead raises F1 by 1.49/2.54/1.56 points, and a second round leaves it essentially unchanged; the text-exchange decline recurs on Qwen2.5-7B (Appendix Table 6). The gap in Table 1 therefore has two parts: a larger part from avoiding the conformity cost of transcripts, and a modest gain from message content, +1.77/2.29/1.95 F1 over the zero-message pipeline at the same depth. Zero messages leave F1 flat through T=3 (58.38/47.30/58.04 at T=1 vs. 58.41/47.51/57.65 at T=3), so extra rounds without message content neither help nor hurt. Across backbones (Table 2b), pruning transcripts recovers little of the text-exchange loss.

Receiver behavior. We compare each agent's round-one and round-two answers using the systems' own containment rule. On RAMDocs with Qwen2.5-1.5B and the same agent set for every condition, 28.8% of MADAM-RAG agents and 66.0% of LLM-Debate agents switch to another agent's round-one answer, compared with 4.4% under LAM-RAG. Round-one gold answers still covered in round two fall to 83.0% and 62.2% under text exchange, versus 94.8% for LAM-RAG. AmbigDocs and FaithEval show the same pattern (peer adoption 19.2% and 10.7% for MADAM-RAG vs. 0.5% and 2.1% for LAM-RAG; Appendix Table 9). LoRA messages thus couple agents weakly. They still carry a question-conditioned signal, since zero, random, and cross-question messages each reduce RAMDocs F1 by 1.30–1.93 points (Table 7), but they do not make peer answers copyable, so each agent stays anchored to its own document. Set-valued reconciliation needs exactly this, and it explains why LoRA exchange does not degrade with depth.

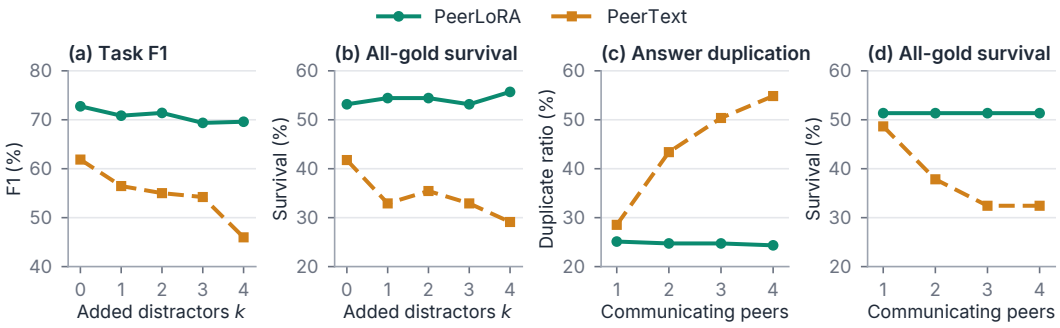

Figure 3: Source-matched carrier robustness (a,b) and the clean multi-peer diagnostic (c,d) on Qwen2.5-7B. (a) F1 and (b) all-gold answer survival under k added distractors; (c) answer duplication and (d) all-gold answer survival as the number of communicating peers increases.

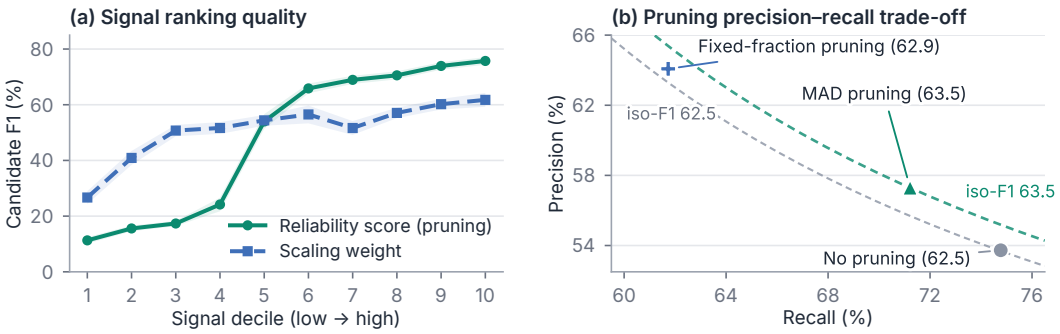

Figure 4: RAMDocs flow-control diagnostics. (a) Candidate F1 by signal decile, pooled over the three backbones (bands: ± 1 bootstrap s.d.). (b) Pruning precision–recall on Qwen2.5-1.5B (T=2); dashed: iso-F1 curves.

4.4 How Robust Is the Communication Carrier to Distractors?

The weak coupling observed above also makes the low-rank carrier more robust to added distractors than explicit peer text (Figure 3a,b). We add $k \in \{0, \ldots, 4\}$ irrelevant documents to 79 clean two- or three-answer RAMDocs questions and compare PeerLoRA with PeerText under matched peer sources, with pruning and scaling disabled, so the diagnostic isolates the carrier rather than the full system. From $k = 0$ to $k = 4$, PeerLoRA F1 changes from 72.75 to 69.60, while PeerText falls from 61.87 to 45.98. The paired gap therefore widens by +12.74 F1 (95% CI [6.21,19.46]), and the all-gold survival interaction is +15.19 points [5.06,25.32]; with early stopping removed, fixed-T=3 gap interactions stay positive (+9.73 on 7B, +11.93 on 1.5B; Appendix F.2).

Without distractors, a clean multi-peer diagnostic shows the conformity mechanism directly: from 1 to 4 communicating peers, PeerText agents increasingly duplicate each other while PeerLoRA stays flat, changing the duplicate/survival gaps by $-27.09/+16.22$ points on 7B and $-44.56/+27.03$ on 1.5B (Figure 3c,d).

4.5 What Does Flow Control Add to LoRA Messages?

Flow control refines the carrier: it adds 1.73/0.29/0.82 F1 over LoRA messages alone on 1.5B/7B/8B (Table 2b), its signals track candidate quality, and MAD pruning balances precision and recall (Figure 4). It does not improve fixed-round source-matched F1 (Appendix F.2), so we treat it as an

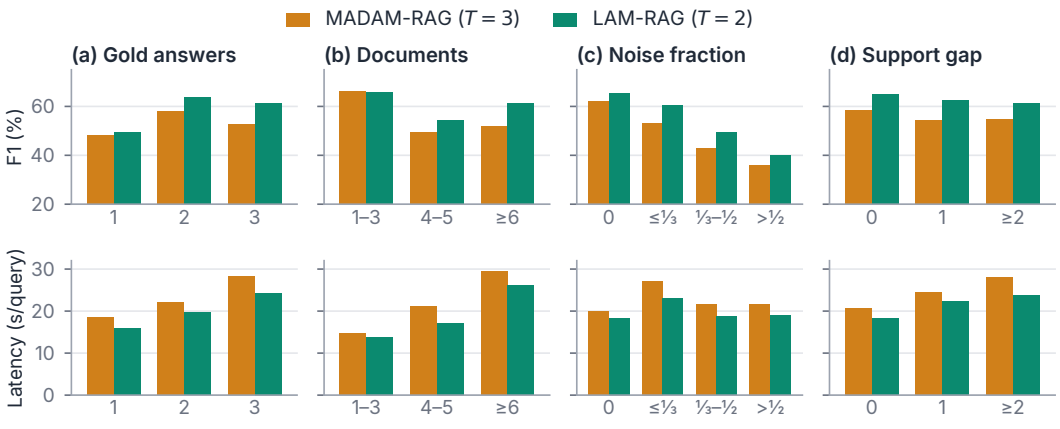

Figure 5: RAMDocs F1 (top) and latency (bottom) by question regime on Qwen2.5-1.5B: (a) gold answers, (b) retrieved documents, (c) noise fraction, (d) support gap among gold answers (multi-answer questions). Bin counts: Appendix Table 17.

influence regulator rather than the source of robustness. It also lowers the retained weight on lexically unsupported round-one states by 8.7–12.8 points across backbones (Table 13; Appendix E.2).

4.6 When Does LAM-RAG Help?

LAM-RAG's RAMDocs F1 advantage grows from +1.09 to +8.50 as the number of answers rises from one to three, and from -0.46 to +9.38 between 1–3 and ≥ 6 documents (Figure 5). Gains persist across noise and support-imbalance regimes (the highest-noise bin has only 16 examples), and under same-round batching LAM-RAG is faster in every bucket.

4.7 What Are the Cost and Interface Boundaries?

Figure 6 compares accuracy with the tokens the backbone processes, not full compute cost. On Qwen2.5-7B RAMDocs, more rounds of transcript exchange cost more and score less: from the shared no-exchange point (61.7 F1, 2.7k tokens), MADAM-RAG falls to 58.0 F1 at 9.9k tokens and LLM-Debate to 52.0 at 14.4k by T=3, whereas one round of LoRA-message exchange raises LAM-RAG from 62.3 to 63.9 F1 for about 2.2k additional tokens. Factorized injection keeps peak memory within 0.5 GB of text-only MADAM-RAG (29.44 vs. 28.93 GB on Qwen2.5-7B) with outputs identical to dense injection (Appendix Table 20).

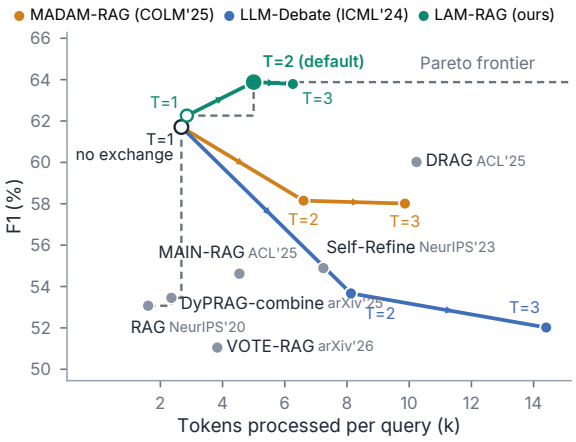

Figure 6: Accuracy–token trade-off on RAMDocs (Qwen2.5-7B). Lines trace T=1 → 2 → 3 from the shared no-exchange point; dashed: Pareto frontier. Tokens include translator encodings and judge calls; Astute-RAG (28.5 F1) is off-scale.

The interface boundary is equally clear: LoRA messages suit intermediate influence, not answer enumeration. Relative to the LoRA-message-only ablation, LoRA-only final aggregation lowers F1 by 49.0 points on average and adding transcripts back into agent communication lowers it by 7.6, each winning none of nine dataset–backbone settings (Appendix Table 21). LAM-RAG therefore keeps text at both ends of the pipeline.

5 Limitations

LAM-RAG targets set-valued QA; when alternatives are already explicit, as on AmbigDocs with Llama-3.1-8B, single-pass RAG can be stronger. The gain from message content is modest (+1.77–2.29 F1 over zero messages), and we do not show faithful transfer of individual peer facts. The translator is backbone-specific and needs weight access, which excludes API-only models, and the controller uses lexical support. We test backbones up to 8B with fixed evidence and mostly greedy decoding; larger models and end-to-end retrieval remain open.

6 Conclusion

Debate pulls agents toward consensus, the wrong objective when documents support different answers. Across three benchmarks and three backbones, transcript exchange lowers answer-set F1 below the no-exchange level, because agents adopt their peers' answers and supported alternatives disappear. LAM-RAG separates answer expression from inter-agent influence: local and final answers stay in text, while influence travels as transient LoRA messages that expose no copyable answers yet still carry useful information. This weak-coupling channel preserves answer diversity, attains the highest average F1 among text-exchange and single-pass methods, degrades less than peer text as distractors or peers are added, and costs fewer tokens than multi-round debate. For multi-agent systems more broadly, the communication carrier, not only the protocol, shapes what agents converge to.

LLM USAGE STATEMENT

In preparing this manuscript, large language models were used only for language polishing, sentence rephrasing, and improving the clarity of expression. They did not generate the paper's research ideas, core claims, or arguments. All reported numbers are taken from saved experimental outputs, and the authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix A specify the frozen translators, the evaluation settings, the full inference procedure (Algorithm 1), and the fixed configuration (Table 3). Appendix B gives dataset statistics and prompt templates, Appendix C documents every baseline adaptation, and Appendix F reports bootstrap intervals and a sampling check. All main results use greedy decoding with a fixed seed. An anonymized artifact with inference code, baseline wrappers, and reproduction scripts is linked in Appendix A and included as supplementary material.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Wenrui Bao, Huan Wang, Jian Wang, Zhangyang Wang, Kai Wang, and Yuzhang Shang. Good agentic friends do not just give verbal advice: They can update your weights. *arXiv preprint arXiv:2605.13839*, 2026. URL https://arxiv.org/abs/2605.13839.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2607–2622, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.131. URL https://aclanthology.org/2025.acl-long.131/.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-LoRA: Instant transformer adaption. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7485–7514, 2025. URL https://proceedings.mlr.press/v267/charakorn25a.html.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26981–26995, Suzhou, China, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1371. URL https://aclanthology.org/2025.emnlp-main.1371/.

Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

Masnun Nuha Chowdhury, Nusrat Jahan Beg, Umme Hunny Khan, Syed Rifat Raiyan, Md Kamrul Hasan, and Hasan Mahmud. Courtroom-style multi-agent debate with progressive RAG and role-switching for controversial claim verification. *arXiv preprint arXiv:2603.28488*, 2026. URL https://arxiv.org/abs/2603.28488.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *Proceedings of the 14th International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.03215.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Llama 3 contributors. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkpACe1lx.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15839–15853, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.770. URL https://aclanthology.org/2025.acl-long.770/.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TrloAXEJ2B.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Samir Yitzhak Gadre, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 16867–16878, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1466/.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Yoonsang Lee, Xi Ye, and Eunsol Choi. AmbigDocs: Reasoning across documents on different entities under the same name. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=mkYCfO822n.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *Proceedings of the 13th International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/48404cd9ce03946c6b7177691f3267a1-Abstract-Conference.html.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through embeddings. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06272.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Vignav Ramesh and Kenneth Li. Communicating activations between language model agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51094–51116, 2025. URL https://proceedings.mlr.press/v267/ramesh25a.html.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905, 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248. PMLR, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

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

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge fusion of large language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4fd5cfd2e31bebbccfa5ffa354c04bdc-Abstract-Conference.html.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O. Arik. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30553–30571, Vienna, Austria, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1476. URL https://aclanthology.org/2025.acl-long.1476/.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *Proceedings of the Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=z1MHB2m3V9.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6106–6131, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.331. URL https://aclanthology.org/2024.acl-long.331/.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

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 *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Zequn Xie and Zhengyang Sun. Mitigating hallucination on hallucination in RAG via ensemble voting. *arXiv preprint arXiv:2603.27253*, 2026. URL https://arxiv.org/abs/2603.27253.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024. URL https://arxiv.org/abs/2401.15884.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/299a08ee712d4752c890938da99a77c6-Abstract-Conference.html.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.20639.

Appendix

Appendix A gives implementation details, the full algorithm, and the default configuration; Appendix B gives dataset statistics and prompts; Appendix C documents baseline adaptations and supplementary baselines; Appendix D reports communication and message diagnostics; Appendix E reports flow-control diagnostics and case studies; and Appendix F covers robustness, uncertainty, and cost.

A Implementation Details

Translator. LAM-RAG uses a backbone-matched DyPRAG-style Text2LoRA translator (Tan et al., 2025) as a fixed text-to-parameter interface. Training follows DyPRAG's stage-2 procedure: augmented QA passages (HotpotQA, 2WikiMultiHopQA, PopQA, and ComplexWebQuestions, generated with Llama-3.2-1B-Instruct) first yield passage-level LoRA adapters on the target backbone, and a lightweight hypernetwork (a two-layer MLP per adapted module, bottleneck p=32, rank-2 factors) is then trained to predict these factors from the backbone's hidden states. Because its input and output spaces are tied to the backbone, we train separate translators for Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct, and Llama-3.1-8B-Instruct, and use the earliest saved checkpoint, following DyPRAG's observation that early checkpoints transfer better to unseen text. No translator checkpoint is trained, tuned, early-stopped, or selected on the evaluation benchmarks.

Message capacity. DyPRAG parameterizes retrieved passages; LAM-RAG instead parameterizes answer-and-rationale states generated under (q, d_i), which are shorter and already question-conditioned, and this input performs better on every backbone (Table 8). The translator still compresses each state into a final hidden representation, so a LoRA message should not be read as a lossless encoding of its text.

Evaluation details. The backbone and translator remain frozen during evaluation; the only test-time state changes are transient LoRA messages, removed after each round. All LAM-RAG runs use FP32 for both backbone and translator. The Qwen2.5-1.5B translator was trained on RTX 4090 GPUs; the other translators, all benchmark inference, and latency measurements used Ascend 910C NPUs. Generation is greedy (`do_sample=False`) with at most 128 new tokens per agent message and 512 for the final aggregator.

Agent granularity. LLM-Debate, MADAM-RAG, and LAM-RAG share the same one-document-per-agent assignment. It preserves identifiable sources for support estimation, pruning, and scaling, at the cost of one document-conditioned generation and translator pass per active agent per round. We do not claim that it is the optimal cost–quality granularity: cluster-level routing could reduce cost but would change source attribution and could merge conflicts before independent control.

Anonymous artifact. For review, we provide an anonymized software snapshot at `https://anonymous.4open.science/r/lam-rag-artifact-2026/` and as supplementary material. It includes inference code, prompt templates, answer-set parsing scripts, matched-evidence baseline wrappers, controller configurations, and reproduction scripts, and excludes author names, affiliations, commit history, and private paths. Translator checkpoints are included when the backbone and data licenses permit; otherwise we provide reconstruction scripts, checksums, and the expected directory layout.

Communication depth and final-source selection. Two rounds capture most of the depth gain: on Qwen2.5-1.5B, moving from T=1 to T=2 improves F1 on all three datasets, while T=3 is marginal (Table 6). With the two-round cap, if the retained answer set is already stable at round two, the final aggregator consumes the round-one retained outputs; otherwise it consumes the round-two outputs. Across the nine main-table settings, round-one outputs are aggregated in 24.0–66.0% of examples (24.0/25.9/49.4% on Qwen2.5-1.5B, 46.4/36.4/66.0% on Qwen2.5-7B, and 57.2/47.7/65.6% on Llama-3.1-8B for RAMDocs/AmbigDocs/FaithEval). In these examples the LoRA message affects the output only through the stability check, so the reported gains arise from the remaining examples.

A.1 Algorithm and Configuration

Algorithm 1 gives the full inference loop and Table 3 the fixed configuration used for all reported runs.

Algorithm 1: LAM-RAG inference procedure.

Input: q, $\mathcal{D}=\{d_i\}_{i=1}^{N}$, T, base weights W
Output: Answer set $\mathcal{A}$

Initialize one agent per d_i; set $W_{\text{eff}}^{1} = W$.
for $t = 1, \ldots, T$ **do**
  /* Generate LoRA messages */
  for each agent i **do**
    $x_i^t \leftarrow \text{LLM}_{W_{\text{eff}}^t}(q, d_i)$;
    $(A_i^t, B_i^t) \leftarrow \text{Text2LoRA}(x_i^t)$;
    $r_i^t \leftarrow \text{Reliability}(a_i^t, d_i)$;
  end
  /* Control message flow */
  $(\mathcal{S}_t, s^t) \leftarrow \text{FlowControl}(r^t)$;
  if the post-control answer set converges **then**
    break;
  end
  if $t < T$ **then**
    /* Inject controlled messages */
    $(\bar{A}^t, \bar{B}^t) \leftarrow \text{FactorMean}(\mathcal{S}_t, s^t)$;
    $W_{\text{eff}}^{t+1} \leftarrow W + \alpha \bar{B}^t \bar{A}^t$;
  end
end
  /* Aggregate surviving textual answers */
$\mathcal{A} \leftarrow \text{Aggregate}(\text{FinalSources}_t)$;
return $\mathcal{A}$; App. A

Table 3: Default LAM-RAG configuration.

Component	Setting
Translator	Frozen Text2LoRA
Bottleneck / rank	p=32, r=2
Injection scale	α=1
Rounds	T=2, stable final source
Confidence / support	Answer-only / substring
Reliability	β=0.5
MAD pruning	λ=3.0, min = 2
Soft-scaling temp.	τ=4.0
Soft-scaling range	$s \in [0, 1.5]$
Scaling off if	$\text{MAD}^t \leq 10^{-3}$
Denominator ε	10^{-4}
Final aggregation	Text, retained agents

Controller sensitivity. The flow-control configuration is not tuned to a narrow operating point: Figure 7 sweeps the MAD threshold, the reliability mixture, and the soft-scaling range under the two-round RAMDocs protocol.

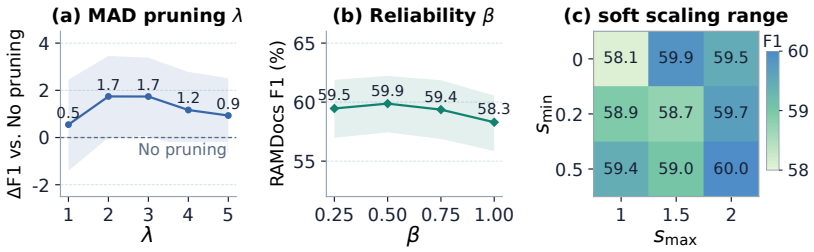

Figure 7: Controller sensitivity on RAMDocs with Qwen2.5-1.5B under the fixed two-round LAM-RAG protocol. Panel (a) reports ΔF1 over the no-pruning variant as the MAD threshold λ varies; panels (b–c) report F1 while varying the reliability mixture β and soft-scaling range. The default configuration lies in a stable region rather than at an isolated optimum.

B Datasets and Prompts

Table 4 summarizes the evaluation sets. RAMDocs is the conflict-heavy benchmark released with MADAM-RAG, AmbigDocs preserves same-name ambiguity, and FaithEval fixes two-document inconsistent-context cases.

Table 4: Dataset statistics for the main evaluation. Docs/query and Gold answers/query report averages, with ranges in parentheses. RAMDocs stresses noisy multi-document integration, AmbigDocs preserves same-name ambiguity, and FaithEval tests faithful answer-set construction under inconsistent contexts.

Dataset	#Queries	Docs/q	Gold/q
RAMDocs	500	5.53 (1–12)	2.20 (1–3)
AmbigDocs	1,000	2.83 (2–7)	2.83 (2–7)
FaithEval	1,000	2.00 (2–2)	2.00 (2–2)

B.1 Prompt Templates

Natural language is used for local answer proposals and final answer-set enumeration. Between rounds, LAM-RAG does not append agent transcripts to the visible prompt; each agent's answer-and-rationale state is translated into a transient LoRA message, and the controlled pool is injected only for the next generation step.

Local Agent i: evidence-conditioned proposal.

```
Role: You are an agent reading one document to answer a question.
Use the assigned document as visible evidence. Answer the question based
only on this document and any internal signal integrated into the model
state for this question.
Return exactly: Answer: {answer}. Explanation: {explanation}.

[Question] {question}
[Document i] {document_i}
```

Final Aggregator: candidate-constrained answer-set enumeration.

```
Role: You are an aggregator reading answers from multiple agents.
Enumerate all answer strings supported by the surviving agent responses.
If multiple correct answers exist, include all of them. If no supported
answer is available, output "unknown". Do not add unsupported answers.
Return exactly: All Correct Answers: [{answer_1}, {answer_2}, ...].
Explanation: {explanation}.

[Question] {question}
[Agent responses]
Agent 1: {answer and explanation}
Agent 2: {answer and explanation}
...
```

C Baselines

Protocol. All baselines consume the same provided document set, use the same answer-set parser, and are decoded greedily with comparable generation caps; output-format normalization makes every backbone produce the same answer-list schema without changing evidence, decoding, parser, or decision rule. Self-Refine follows the single-agent self-reflection baseline of MADAM-RAG (Wang et al., 2025b), adapted from Madaan et al. (2023): the model answers from all documents, critiques its answer, and refines it for two rounds.

Answer-stage adaptations. DRAG, VOTE-RAG, and MAIN-RAG originally include retrieval, filtering, or voting over evidence. Under matched evidence these upstream stages are controlled, and we evaluate the component that consumes a fixed evidence set: DRAG keeps proponent–opponent debate and moderation (omitting Query Debate), MAIN-RAG keeps predictor–judge filtering and prediction from the retained subset, and VOTE-RAG keeps independent generation and majority voting (omitting Retrieval Voting). These are therefore not end-to-end rankings of the original pipelines. MAIN-RAG's near-zero AmbigDocs EM reflects incomplete ambiguity-set recovery rather than failed runs: all 1,000 raw outputs are non-empty on every backbone, parsed answer sets are non-empty in 997/968/1,000 examples, and it retains 1.46–1.54 documents per query against 2.83 gold answers.

LatentMAS. LatentMAS (Zou et al., 2026), a general latent-communication method, is adapted to one agent per document: each agent reads its document with MADAM-RAG's round-one prompt, runs m latent steps, and appends to a KV cache shared along the agent chain, from which a final agent decodes the answer set (official code, commit `9a9e4d3`). As in the original paper, m and latent realignment are tuned, here per backbone on 200 held-out questions disjoint from the evaluation sets (m=10 for Qwen2.5-7B and m=80 for Llama-3.1-8B, both without realignment). On Qwen2.5-1.5B every setting degenerates; the official code does the same on its own GSM8K task (0/50 versus 26/50 for a single agent) on both NPU and GPU. LatentMAS trails LAM-RAG by 14.42 average F1 on Qwen2.5-7B and exceeds it by 4.34 on Llama-3.1-8B (Table 5).

Table 5: Supplementary baselines under the matched-evidence protocol (EM/F1, %). †Adapted to the fixed evidence (Appendix C). –: LatentMAS yields no usable output on Qwen2.5-1.5B.

Backbone	Method	RAMDocs		AmbigDocs		FaithEval		Avg.	
		EM	F1	EM	F1	EM	F1	EM	F1
Qwen2.5-1.5B	DRAG† ACL'25	5.6	29.21	1.5	25.73	10.9	40.73	6.0	31.89
	VOTE-RAG† arXiv'26	8.4	37.14	5.2	34.79	17.4	42.80	10.3	38.25
	MAIN-RAG† ACL'25	6.2	23.30	0.0	19.22	0.8	16.95	2.3	19.82
	LatentMAS† ICML'26	–	–	–	–	–	–	–	–
Qwen2.5-7B	DRAG† ACL'25	20.2	60.02	7.7	43.04	15.6	56.00	14.5	53.02
	VOTE-RAG† arXiv'26	14.0	51.05	7.0	40.28	27.5	59.30	16.2	50.21
	MAIN-RAG† ACL'25	14.8	54.62	0.0	36.89	1.0	47.24	5.3	46.25
	LatentMAS† ICML'26	12.6	48.52	2.0	34.99	9.9	56.53	8.2	46.68
Llama-3.1-8B	DRAG† ACL'25	16.8	62.76	9.4	45.93	23.4	60.13	16.5	56.27
	VOTE-RAG† arXiv'26	14.0	56.12	8.4	35.30	30.9	50.42	17.8	47.28
	MAIN-RAG† ACL'25	16.6	61.48	0.2	38.48	0.3	43.71	5.7	47.89
	LatentMAS† ICML'26	19.6	66.86	13.3	53.76	45.7	70.36	26.2	63.66

D Communication and Message Diagnostics

D.1 Communication Depth on Two Backbones

Table 6 repeats the depth comparison of Table 2a on Qwen2.5-7B. Transcript exchange again lowers F1 below the no-exchange level on all three benchmarks ($-3.56/-1.11/-1.43$ from T=1 to T=2 for MADAM-RAG). LoRA-message exchange raises F1 on RAMDocs and AmbigDocs ($+1.62/+0.79$) and is slightly below its no-exchange level on FaithEval (-0.38), where MAD pruning of round-one outputs already gives LAM-RAG a higher T=1 starting point.

Table 6: Communication depth on two backbones (F1, %). T=1 has no exchange; T=2/3 add one/two exchanges. T=3 MADAM-RAG and T=2 LAM-RAG entries are the main-table runs; – marks settings not run.

Backbone	Method	RAMDocs			AmbigDocs			FaithEval		
		1	2	3	1	2	3	1	2	3
Qwen2.5-1.5B	MADAM-RAG	58.10	56.57	53.99	48.20	45.35	45.08	55.98	54.69	54.65
	LAM-RAG	58.38	**59.87**	59.25	47.30	49.84	**50.11**	58.04	**59.60**	59.51
Qwen2.5-7B	MADAM-RAG	61.71	58.15	57.95	53.07	51.96	52.03	63.70	62.27	62.80
	LAM-RAG	62.26	**63.88**	63.79	53.05	**53.84**	–	**65.96**	65.58	–

D.2 Message Content and Translator Input

Table 7 keeps injection active while removing or mismatching the question-conditioned signal; every control lowers RAMDocs F1. Table 8 holds the loop, checkpoint, controller, and decoding budget

fixed and changes only the translator input from Answer+Explanation to Question+Document on the same 500 RAMDocs questions. Answer-state inputs are shorter and perform better on each backbone: the mean paired difference is +0.99 F1, with a question-cluster 95% CI of [+0.32, +1.69] over the three fixed backbones.

Table 7: LoRA-message controls on RAMDocs with Qwen2.5-1.5B. Controls keep injection active while removing or mismatching the question-conditioned signal. Intervals and p values are paired-bootstrap results versus LAM-RAG.

Variant	ΔF1	95% CI	p
LAM-RAG	$+0.00$	–	–
Zero msg.	-1.77	$[-3.38, -0.10]$	0.037
Random msg.	-1.93	$[-3.57, -0.39]$	0.019
Cross-question	-1.30	$[-2.23, -0.40]$	0.003

Table 8: Translator input on RAMDocs (F1): answer state versus question plus document (Q+Doc). The Qwen-7B value is from the paired diagnostic; the canonical main-table run reports 63.88.

Backbone	Answer	Q+Doc	ΔF1
Qwen-1.5B	59.87	58.63	+1.24
Qwen-7B	63.74	62.92	+0.82
Llama-8B	66.19	65.27	+0.92

D.3 Receiver Answer Changes

Table 9 compares round-one and round-two answers of the same agents on Qwen2.5-1.5B (LLM-Debate: the second round of its three-round run). An answer counts as changed when neither normalized string contains the other, the rule the systems use for convergence. *Peer adoption* is a change to an answer another agent gave in round one; *gold kept* is the share of round-one-covered gold answers still covered in round two; *distinct* is the ratio of distinct round-two to round-one answers. All conditions use the agents LAM-RAG retains after round-one pruning; using all agents for the text systems gives the same pattern (RAMDocs peer adoption 32.1% for MADAM-RAG and 67.6% for LLM-Debate). The raw change rate also counts paraphrases, so peer adoption is the more informative column on AmbigDocs.

Table 9: Receiver behavior from round one to round two on Qwen2.5-1.5B (%, except distinct ratio). Text exchange raises peer adoption and reduces answer diversity; LoRA messages do not.

Dataset	Condition	Changed	Peer adoption	Gold kept	Distinct	Agents
RAMDocs	LLM-Debate	69.99	65.99	62.19	0.605	2,499
	MADAM-RAG	34.89	28.81	83.04	0.696	2,499
	LAM-RAG	22.17	4.40	94.77	0.989	2,499
AmbigDocs	MADAM-RAG	35.33	19.19	77.23	0.796	2,720
	LAM-RAG	38.20	0.51	89.00	1.000	2,720
FaithEval	MADAM-RAG	22.80	10.70	83.53	0.872	2,000
	LAM-RAG	23.35	2.10	91.40	1.008	2,000

D.4 Factor-Space Composition

FactorMean averages factors before multiplying them, $\Delta W_{\text{FactorMean}} = \left(\frac{1}{M}\sum_i B_i\right)\left(\frac{1}{M}\sum_j A_j\right) = \frac{1}{M^2}\sum_{i,j} B_i A_j$, so it contains cross-message terms $B_i A_j$ ($i \neq j$); UpdateMean, $\frac{1}{M}\sum_i B_i A_i$, averages complete per-message updates. With pruning and scaling disabled, the choice changes 8/100 final answer sets on a Qwen2.5-1.5B prefix (Table 10), while a source-matched diagnostic on 79 multi-answer RAMDocs questions with Qwen2.5-7B favors neither operator (-0.59 F1 $[-1.48, 0.00]$ at k=0; $+0.51$ $[-0.26, 1.39]$ at k=4). We therefore document the canonical operator's semantics without claiming that it dominates update-space fusion.

Table 10: Controlled FactorMean versus UpdateMean diagnostic on a deterministic 100-question RAMDocs prefix with Qwen2.5-1.5B, $T=2$, and pruning/scaling disabled. All 100 examples in this prefix have single-answer gold sets. The interval is a paired question-level bootstrap.

Composition	F1	ΔF1 vs. FactorMean
FactorMean	**52.47**	0.00
UpdateMean	51.06	-1.40 $[-3.25, -0.09]$

E Flow-Control Diagnostics

E.1 Control-Signal Quality

Grounded reliability is a stronger candidate-level signal than raw confidence or lexical support alone (Table 11): top1 is the F1 of the highest-ranked candidate, pair is pairwise ranking accuracy, ROC/PR are threshold AUCs, and best-pr. is the fraction of decisions that remove all oracle-best candidates (lower is safer). Soft lexical support is a token-overlap reference; the online controller uses binary substring support. MAD pruning keeps higher-support, more gold-aligned messages than it removes (Table 12), and its precision–recall position lies between fixed-fraction pruning and no pruning (Figure 4).

Table 11: Ranking quality of candidate-level control signals (RAMDocs and TriviaQA average); Reliability is grounded reliability. Grounded reliability improves top-ranked candidate F1, pairwise ranking accuracy, ROC-AUC, and PR-AUC, while reducing the risk of pruning all oracle-best candidates.

Signal	top1	pair	ROC	PR	best-pr.
Confidence	.477	.566	.517	.454	.033
Soft support	.502	.609	.616	.524	.021
Reliability	**.566**	**.676**	**.788**	**.696**	**.013**

Table 12: MAD pruning diagnostics on RAMDocs. K/P denotes kept/pruned candidate means; lower best-pruned ratio is safer.

Backbone	Support	Cand. F1	Gold-exact	Best-pr.
Qwen-1.5B	.61/.18	.45/.19	.29/.04	0.92%
Qwen-7B	.55/.22	.47/.23	.29/.07	0.93%
Llama-8B	.87/.40	.61/.30	.44/.12	0.81%

E.2 Local Support, Calibration, and Failure Cases

An answer state is lexically unsupported when its normalized answer is absent from its own document. Over 500 questions and 2,766 non-empty round-one messages per backbone, pruning preferentially removes unsupported states and scaling lowers their retained weight (Table 13); the weight share is a controller statistic, not a causal decomposition of the update, and not a hallucination rate, since aliases and paraphrases can fail lexical support. Semantic support is not a consistent replacement (Table 14): in an offline audit of 7,598 Qwen2.5-1.5B candidates (threshold 0.70), semantic-only matches widen coverage (14.6–48.9% of candidates) but have much lower candidate-level F1 (22.4–37.6) than candidates supported by both signals (66.9–70.0), so the controller keeps substring support. Across the three RAMDocs round-one logs, 80 exact-gold candidates lack lexical support; 29 are pruned and 19 are not recovered.

E.3 Case Studies

Table 15 shows how the controller acts on one RAMDocs question: LoRA messages add supported candidates and one unsupported candidate, MAD pruning removes the unsupported one, and soft scaling preserves a further supported alternative. Table 16 gives two verified failures of local lexical support.

Table 13: Share of round-one states that are lexically unsupported on RAMDocs (%): over all, kept, and pruned states, and as a share of retained controller weight.

Backbone	All	Kept	Pruned	Weight
Qwen-1.5B	60.09	56.70	91.76	51.39
Qwen-7B	45.41	37.12	87.31	32.61
Llama-8B	23.36	12.74	73.80	11.10

Table 14: Semantic-support diagnostic on Qwen2.5-1.5B with the fixed $T=3$ controller used for this ablation. ΔF1 is relative to substring support on the same dataset; these values are not the main $T=2$ scores.

Dataset	Support	F1	ΔF1
RAMDocs	substring	59.25	0.00
	embedding	58.84	−0.42
	hybrid	60.42	+1.17
FaithEval	substring	59.51	0.00
	embedding	59.24	−0.27
	hybrid	59.42	−0.09

Table 15: Gibraltar Barracks case study on RAMDocs. The example illustrates how LoRA-message exchange adds supported answer candidates, while the full controller suppresses an unsupported candidate. Boldface marks retained supported answers, shaded non-bold text marks unsupported candidates, '+' marks additions, and strikethrough marks controller removals.

Question: When were the Gibraltar Barracks established?

Variant	F1	Answer-set evolution
Text Debate	50.0	**1878**
LoRA msg. only	57.1	**1878**; +**1797**; +1748
LoRA msg. + MAD	66.7	**1878**; **1797**; ~~1748~~
LoRA msg. + scaling	71.4	**1878**; **1797**; +**early 1970s**; 1748
LAM-RAG (full)	83.3	**1878**; **1797**; +**early 1970s**; ~~1748~~

Note: msg. = message.

Table 16: Failure cases: local lexical support neither captures all valid answers nor verifies query relevance.

Case	Evidence and intermediate state	Outcome
Peter Jacobs	The passage describes Hawaii Ironman results but omits "Triathlete." Confidence is 0.829; support 0 reduces reliability to 0.415 and weight to 0.406.	The correct candidate is later pruned and absent from the final set.
McCoy Tyner	An added passage about the album *Inception* lexically supports the unrelated candidate McCoy Tyner.	Output changes from Andy Williams to Andy Williams / McCoy Tyner; example F1 falls from 0.667 to 0.500.

F Robustness, Uncertainty, and Cost

F.1 Regime Breakdown

Table 17 gives the exact values behind Figure 5, with bin sizes.

F.2 Source-Matched Carrier Controls

The carrier comparison uses 79 RAMDocs questions with two or three gold answers and no original noise. Original documents and gold answers stay fixed while irrelevant documents from other questions are added under a frozen leakage filter. PeerText receives the previous peer Answer+Explanation states as text; PeerLoRA encodes the same peer states with the frozen translator and applies a receiver-specific peers-minus-self FactorMean update. Source matching controls the eligible sender identities, not identical message contents after trajectories diverge. Both conditions disable pruning and scaling and keep textual final aggregation; across all jointly executed rounds, 2,815 receiver-route checks found no self leakage or source-set mismatch. Uncertainty is a paired question-level bootstrap.

Table 17: Qwen2.5-1.5B regime trade-offs relative to MADAM-RAG under same-round batching (LAM-RAG T=2, MADAM-RAG T=3). $+\Delta$F1 is better; $-\Delta$Latency is faster.

Regime (n)	ΔF1	ΔLat.
RAMDocs: answer multiplicity		
1 answer (100)	$+1.09$	-2.59
2 answers (200)	$+5.67$	-2.52
3 answers (200)	$+8.50$	-3.88
RAMDocs: document count		
1–3 docs (100)	-0.46	-0.93
4–5 docs (162)	$+4.66$	-3.87
≥ 6 docs (238)	$+9.38$	-3.44
RAMDocs: support imbalance		
Balanced (95)	$+6.43$	-2.58
Mild gap $= 1$ (130)	$+8.07$	-2.03
Strong gap ≥ 2 (175)	$+6.70$	-4.41
RAMDocs: noise ratio		
0 (163)	$+3.52$	-1.54
$(0, 1/3]$ (249)	$+7.33$	-4.16
$(1/3, 1/2]$ (72)	$+6.61$	-2.86
$> 1/2$ (16)	$+4.21$	-2.82
AmbigDocs: answer multiplicity		
2 answers (504)	$+2.42$	$+4.15$
3 answers (274)	$+6.58$	-1.33
≥ 4 answers (222)	$+7.86$	-2.62

Forcing exactly three rounds gives F1 gap interactions of +9.73 [2.89, 17.19] on Qwen2.5-7B and +11.93 [3.26, 20.29] on Qwen2.5-1.5B; the all-gold survival interactions are +7.59 [-2.53, 17.72] and +10.13 [3.80, 17.72], so the 7B fixed-round survival interval includes zero. In the 37-question multi-peer test without distractors, the F1 interaction is backbone-dependent (-0.07 on 7B, +11.82 on 1.5B), so we use it as mechanism evidence for diversity preservation, not as an F1 scaling law. Enabling the controller in the fixed-round protocol adds no task-level robustness (-3.16 [-6.95, 0.55] for k=4 versus k=0), although the reliability signal separates original from distractor agents (AUC 0.839); this motivates treating flow control as an influence regulator rather than a second robustness mechanism.

F.3 Bootstrap Confidence Intervals

Table 18 reports 95% intervals for LAM-RAG and the strongest other method in each setting. LAM-RAG has the highest mean F1 in eight of nine settings; the exception is Llama-3.1-8B on AmbigDocs, where RAG is stronger. The intervals are method-level rather than paired, so overlap does not by itself imply a non-significant difference.

Table 18: Method-level 95% bootstrap intervals for mean F1 (10,000 example resamples per method). The comparison method is the strongest non-LAM method of Table 1 in each setting.

Backbone	Dataset	LAM-RAG	Strongest other method	Other F1
Qwen2.5-1.5B	RAMDocs	**59.87** [57.41, 62.27]	MADAM-RAG	53.99 [51.23, 56.78]
	AmbigDocs	**49.84** [48.05, 51.66]	MADAM-RAG	45.08 [43.20, 46.97]
	FaithEval	**59.60** [57.31, 61.86]	MADAM-RAG	54.65 [52.35, 56.91]
Qwen2.5-7B	RAMDocs	**63.88** [61.31, 66.47]	MADAM-RAG	57.95 [55.14, 60.71]
	AmbigDocs	**53.84** [52.10, 55.63]	MADAM-RAG	52.03 [50.18, 53.97]
	FaithEval	**65.58** [63.45, 67.69]	RAG	63.68 [61.63, 65.71]
Llama-3.1-8B	RAMDocs	**66.19** [63.80, 68.51]	Self-Refine	62.74 [60.24, 65.21]
	AmbigDocs	49.67 [47.79, 51.58]	RAG	**52.53** [50.69, 54.39]
	FaithEval	**62.10** [59.75, 64.38]	MADAM-RAG	58.74 [56.48, 61.01]

F.4 Stochastic Decoding Check

To test whether the main pattern depends on greedy decoding, we rerun Qwen2.5-1.5B with nucleus sampling over three reproducible trajectories (Table 19). LAM-RAG keeps the highest mean F1 on RAMDocs and FaithEval, and its RAMDocs advantage is recall-oriented (60.94 versus 52.88 for MADAM-RAG). This is a single sampling configuration, not a decoding sweep.

Table 19: Nucleus-sampling check on Qwen2.5-1.5B (`temperature=0.7`, `top_p=0.9`; mean ± s.d. over three trajectories, %). LAM-RAG uses its default $T=2$; baselines use their canonical depth.

	RAMDocs		FaithEval	
Method	F1	Recall	F1	Recall
RAG	21.20±1.38	28.11±0.69	30.74±1.07	40.40±0.71
DyPRAG-combine	17.65±0.24	22.87±0.83	25.47±0.56	34.62±0.85
LLM-Debate	39.09±1.11	36.63±1.18	50.06±0.82	53.99±0.47
MADAM-RAG	50.05±1.70	52.88±1.78	50.94±0.63	55.02±0.72
LAM-RAG	**52.47**±0.86	**60.94**±0.67	**53.97**±0.25	**57.23**±0.50

F.5 Peak Memory and Interface Boundary

A direct implementation materializes the dense update BA for every adapted projection, roughly doubling peak memory on Qwen2.5-7B. The canonical implementation applies each message as a factorized product $(hA^\top)B^\top$; on a 100-question RAMDocs prefix it produces identical round-one, round-two, and final outputs on both Qwen backbones (maximum next-token logit difference 6.8×10^{-5}) and stays within 0.5 GB of text-only MADAM-RAG (Table 20). LoRA messages are designed for intermediate influence, not answer enumeration: LoRA-only final aggregation and restoring transcripts in agent communication both lower F1 (Table 21).

Table 20: Peak memory on RAMDocs (GB, single query, $T=2$) for text-only MADAM-RAG and LAM-RAG with dense or factorized injection. Factorized injection gives identical outputs to dense injection and brings LAM-RAG to within 0.5 GB of text-only MADAM-RAG.

Backbone	MADAM	Dense	Factorized	Δ vs. dense
Qwen-1.5B	5.97	10.43	**6.19**	−40.6%
Qwen-7B	28.93	50.59	**29.44**	−41.8%

Table 21: Interface diagnostics separate intermediate communication from final answer enumeration. ΔF1 is relative to the LoRA-message-only ablation over nine dataset–backbone settings.

Variant	Mean ΔF1	Wins
Final LoRA only	−49.0	0/9
Final Text+LoRA	−2.0	1/9
Agent Text+LoRA	−7.6	0/9