Adaptive Termination in Agentic Multimodal RAG via Cross-Modal Evidence Grounding

Kshitij Dahiya [*1] **Aryan Kaushik** [*2] **Utkarsh Singh** [*3]

Abstract

Multimodal retrieval-augmented generation misses two fundamental primitives: an evidence-grounding metric that evaluates generated claims based on evidence in text, imagery and structured data and a convergence criterion for determining the stopping point of an iterative system. We formally establish both. First, we propose CMEGS (Cross-Modal Evidence Grounding Score) as a claim-level metric with a specifically designed inter-modal agreement component that possesses boundedness, monotonicity and graceful degradation due to missing modalities. Second, we calculate a marginal improvement ratio (MIR) from consecutive values of CMEGS, proving its guaranteed termination under finite budget while meeting scale-relative sensitivity. We embed both primitives into an inference-only architecture with five agents and modality adaptation. On PathVQA, our architecture achieves 90.0% faithfulness with 1.1 mean iterations, reducing hallucination by 71% relative to vanilla RAG. On ScienceQA, it performs at par with the best iterative system in grounding but with 2.3× fewer iterations. The metric remains stable under nine different weight settings (variance below 1.3%) and exhibits strong correlation with grounding judgment scores ($r = 0.79$).

[*]Equal contribution [1]Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India [2]Department of Information Technology and Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India [3]Department of Computer Science Engineering (Artificial Intelligence), Maharaja Agrasen Institute of Technology, New Delhi, India. Correspondence to: Aryan Kaushik <aryan.04214813124@ite.mait.ac.in>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Retrieval-augmented generation (Lewis et al., 2020; Gao et al., 2024) has become the standard approach to ground large language models with external knowledge. It is simple and straightforward: first, retrieval, concatenating retrieved texts with queries and generating answers. This approach serves perfectly for simple factual questions where one retrieval step is sufficient. However, in cases when some reasoning over heterogeneous facts is required (texts, images, and tables), this approach fails structurally.

Consider a medical question-answering system. The retriever surfaces a histology slide and a clinical text passage. The generator produces a claim. Is that claim supported by the text? By the image? By both? Standard metrics like RAGAS (Es et al., 2024) cannot answer the second and third questions because they operate entirely in text space. InterSHAP (Wenderoth et al., 2025) quantifies modality contributions via Shapley interaction indices but targets interpretability, not claim-level grounding. No existing metric provides a per-claim score over multiple modalities that can be computed at inference time and used to steer an agentic loop. Meanwhile, hallucination in large language models remains a persistent concern (Huang et al., 2025), reinforcing the need for principled grounding mechanisms.

The stopping problem is equally underspecified. LoopRAG (Bai et al., 2026) iterates a fixed number of PDCA cycles. Others halt on a stop token or a counter. Neither strategy accounts for whether the answer quality is actually still improving. There is no formal link between the evaluation metric and the termination decision.

We address these gaps by contributing two formal primitives and an architecture that combines them:

1. **CMEGS** (Section 3.1): a per-claim grounding metric over text, image and structured evidence with a cross-modal agreement term. We establish four properties: boundedness, monotonicity, agreement sensitivity and graceful degradation.
2. **MIR-based termination** (Section 3.2): a convergence criterion derived from consecutive CMEGS scores. We prove termination, early-exit efficiency, plateau detection and scale-relative sensitivity as formal guarantees.

3. **Training-free agent architecture** (Section 3.3): a five-agent loop with adaptive modality routing that embeds CMEGS as both the evaluation signal and the stopping criterion, requiring no fine-tuning.

The novelty is not in any individual agent. Planners, retrievers and reasoners are well-studied. What is new is the formal coupling of a cross-modal grounding metric to an adaptive convergence rule, closing a gap that the recent survey by Singh et al. (Singh et al., 2025) identifies as an open problem in agentic RAG.

2. Related Work

Grounding evaluation in RAG. RAGAS (Es et al., 2024) introduced automated evaluation of faithfulness and relevance for RAG systems but is limited to text evidence. GaRAGe (Sorodoc et al., 2025) added grounding annotations and showed persistent unfaithfulness even in capable models but its evaluation protocol does not extend to images or structured data. Our CMEGS metric addresses the multimodal gap by defining per-modality grounding scores and an inter-modal agreement term in a unified formulation.

Cross-modal analysis. InterSHAP (Wenderoth et al., 2025) applies Shapley interaction indices to quantify modality contributions in multimodal models. Self-RAG (Asai et al., 2024) trains models to retrieve on demand and critique their own outputs but operates entirely in text space. Multimodal LLMs such as LLaVA (Liu et al., 2023) unify vision and language but internalize grounding within model weights rather than exposing a per-claim score. CMEGS fills this role: it is computed at inference time, per claim and drives the adaptive loop.

Iterative and agentic RAG. DRAG (Hu et al., 2025) proposes multi-agent debate but is text-only and lacks a formal stopping rule. LoopRAG (Bai et al., 2026) applies iterative PDCA refinement with a fixed cycle count. MM-RewardRAG (Fan et al., 2025) adds multimodal retrieval with reward-based reranking but is single-pass and requires supervised training. MuRAG (Chen et al., 2022) and RA-CM3 (Yasunaga et al., 2023) extend retrieval to images through end-to-end training. ReAct (Yao et al., 2023) established the reasoning-and-acting paradigm for agentic systems. Our work is distinguished by three properties simultaneously: multimodal grounding, formal convergence and zero training requirement.

3. Formalism

This section presents the two formal contributions. Section 3.1 defines CMEGS and establishes its properties. Section 3.2 defines MIR-based termination and proves its guarantees. Section 3.3 describes the agent architecture that operationalizes both.

Table 1. Capability comparison. **MM**: multimodal retrieval. **Agt**: multi-agent. **Grnd**: formal grounding metric. **Conv**: principled convergence. **TF**: training-free.

System	MM	Agt	Grnd	Conv	TF
RAGAS	×	×	text	×	✓
MM-RewardRAG	✓	×	×	×	×
DRAG	×	✓	×	×	✓
LoopRAG	×	✓	×	×	✓
Ours	✓	✓	✓	✓	✓

3.1. CMEGS: Cross-Modal Evidence Grounding Score

Let $\mathcal{M} = \{T, I, S\}$ denote the set of modalities (text, image, structured). Given a query q, an evidence corpus $\mathcal{D}$ and a generator $\mathcal{G}$, the system produces an answer composed of claims $C = \{c_1, \ldots, c_n\}$. A retriever surfaces evidence sets $E_m \subseteq \mathcal{D}$ for each active modality $m \in \mathcal{M}^* \subseteq \mathcal{M}$ where $\mathcal{M}^*$ is determined by the router. Let $\phi : \mathcal{X} \to \mathbb{S}^{d-1}$ be a text encoder mapping to the unit hypersphere and $\psi : \mathcal{V} \to \mathbb{S}^{d-1}$ a vision-language encoder projected to the same space.

Definition 3.1 (Per-modality grounding)**.** For a claim c and modality $m \in \mathcal{M}^*$:

$$G_m(c) = \max\Big(0,\ \max_{e \in E_m} \cos\big(\phi(c),\ \xi_m(e)\big)\Big) \tag{1}$$

where $\xi_m = \phi$ for text and structured evidence, $\xi_m = \psi$ for image evidence, and $\cos(\cdot,\cdot)$ denotes cosine similarity between L2-normalized vectors. The outer $\max(0,\cdot)$ clamps negative similarities to zero: a negative cosine between a claim and its nearest evidence indicates no grounding support.

The $\max$ aggregation is deliberate. Grounding is an existential property: a claim is supported if at least one piece of strong evidence backs it. An average of the evidence set would weaken solid evidence with nonsensical text.

Definition 3.2 (Cross-modal agreement)**.** Let $\hat{e}_m(c) = \arg\max_{e \in E_m} \cos(\phi(c), \xi_m(e))$ be the best-matching evidence for claim c in modality m. Let $P(\mathcal{M}^*)$ be the set of all unordered pairs from the active modalities. The agreement term is:

$$A(c) = \frac{1}{|P(\mathcal{M}^*)|} \sum_{(a,b) \in P(\mathcal{M}^*)} \max\big(0,\ \cos\big(\xi_a(\hat{e}_a),\ \xi_b(\hat{e}_b)\big)\big) \tag{2}$$

The agreement parameter reflects whether the best evidence, regardless of the type of modality, supports a coherent story. For example, if G_T and G_I are high but A is low, it means that although the textual and visual evidence agree with the claim, they do not agree with each other.

Definition 3.3 (CMEGS)**.**

$$\mathrm{CMEGS}(c) = \sum_{m \in \mathcal{M}^*} w_m \cdot G_m(c) \; + \; w_A \cdot A(c) \quad (3)$$

where $w_m, w_A \geq 0$ and $\sum_{m \in \mathcal{M}^*} w_m + w_A = 1$. When a modality m is inactive ($m \notin \mathcal{M}^*$), its weight is assumed to be equal to zero while the rest of the weights are readjusted to satisfy the sum-to-one requirement.

The system-level score over all claims is:

$$\mathrm{CMEGS}_{\mathrm{sys}} = \frac{1}{|C|} \sum_{c \in C} \mathrm{CMEGS}(c) \quad (4)$$

Proposition 3.4 (Properties of CMEGS)**.** *For any claim c and active modality set $\mathcal{M}^*$:*

1. ***Boundedness.*** $\mathrm{CMEGS}(c) \in [0, 1]$.
   Proof. By Definition 3.1, $G_m(c) \geq 0$ for all m. By Definition 3.2, $A(c) \geq 0$ since it is a mean of clamped cosine similarities. The weights satisfy $\sum w_m + w_A = 1$ with $w_m, w_A \geq 0$, so $\mathrm{CMEGS}(c)$ is a convex combination of non-negative terms bounded above by 1 and below by 0. □
2. ***Monotonicity.*** *If $G_m(c)$ increases for any modality m while all other terms remain fixed, $\mathrm{CMEGS}(c)$ strictly increases.*
   Proof. Follows directly from $w_m > 0$ when $m \in \mathcal{M}^*$ and linearity of the score. □
3. ***Agreement sensitivity.*** *For claims c_1, c_2 with identical per-modality scores but $A(c_1) \neq A(c_2)$: $\mathrm{CMEGS}(c_1) \neq \mathrm{CMEGS}(c_2)$, provided $w_A > 0$.*
   Proof. The difference $\mathrm{CMEGS}(c_1) - \mathrm{CMEGS}(c_2) = w_A \cdot (A(c_1) - A(c_2)) \neq 0$. □
4. ***Graceful degradation.*** *As modalities become unavailable ($|\mathcal{M}^*|$ decreases), CMEGS remains well-defined. When $|\mathcal{M}^*| = 1$ with modality m and w_A is redistributed, $\mathrm{CMEGS}(c) = G_m(c)$.*
   Proof. With a single modality, $P(\mathcal{M}^*) = \emptyset$, so $A(c)$ is undefined. Setting $w_A = 0$ and re-normalizing gives weight 1 to the single active modality. □

Derived metrics. A claim c is classified as *supported* if $\max_{m \in \mathcal{M}^*} G_m(c) \geq \tau_s$ (we set $\tau_s = 0.65$), and *hallucinated* otherwise. *Faithfulness* is the fraction $|\{c \in C : c \text{ is supported}\}|/|C|$.

3.2. Adaptive Termination via MIR

The question that CMEGS alone does not answer is: *should the system iterate again?* A score of 0.72 could mean the system has extracted everything useful from the corpus, or it could mean one more retrieval-synthesis step would push it to 0.85. We need a criterion that distinguishes the two cases.

Definition 3.5 (MIR convergence criterion)**.** Let $s_t = \mathrm{CMEGS}_{\mathrm{sys}}^{(t)}$ denote the system-level score at iteration t. Define the information gain $\mathrm{IG}(t) = s_t - s_{t-1}$ and the marginal improvement ratio:

$$\mathrm{MIR}(t) = \frac{\mathrm{IG}(t)}{s_{t-1} + \epsilon}, \quad \epsilon = 10^{-8} \quad (5)$$

The system terminates at iteration t if any of the following holds:

$$\mathrm{MIR}(t) < \tau \quad \text{(diminishing returns)} \quad (6)$$

$$s_t \geq \theta \quad \text{(quality satisfied)} \quad (7)$$

$$t \geq T_{\max} \quad \text{(resource bound)} \quad (8)$$

with hyperparameters $\tau = 0.02$, $\theta = 0.85$, $T_{\max} = 5$.

Theorem 3.6 (Termination guarantees)**.** *For any query q, evidence corpus $\mathcal{D}$, and generator $\mathcal{G}$, the MIR criterion satisfies:*

1. ***Bounded termination.*** *The loop terminates in at most $T_{\max}$ iterations.*
   Proof. Condition (8) fires at $t = T_{\max}$ regardless of the other conditions. □
2. ***Early exit.*** *If $s_1 \geq \theta$, the system returns after a single iteration.*
   Proof. Condition (7) is checked after each iteration. If $s_1 \geq \theta$, it fires at $t = 1$. □
3. ***Plateau detection.*** *If $\mathrm{IG}(t) < \tau \cdot (s_{t-1} + \epsilon)$, the system stops even when $s_t < \theta$, preventing cycles that yield negligible improvement.*
   Proof. Directly from equation (6). The important point is that the negligible level is defined relative to the existing score rather than an absolute value. □
4. ***Scale-relative sensitivity.*** *If the score is low (s_{t-1} small), MIR gives great weight to any improvement, however slight. If the score is high (s_{t-1} large), MIR tends to be more cautious, demanding correspondingly greater improvements.*
   Proof. Consider fixed $\mathrm{IG} = \delta > 0$. Then $\mathrm{MIR} = \delta/(s_{t-1} + \epsilon)$. For $s_{t-1} \to 0$, $\mathrm{MIR} \to \delta/\epsilon \gg \tau$, so the system continues iterating. For s_{t-1} large, $\mathrm{MIR} \to \delta/s_{t-1}$, which falls below τ when $s_{t-1} > \delta/\tau$. □

Remark 3.7. It is Property 4 which sets MIR apart from the fixed threshold approach. If the criterion were "stop when $\mathrm{IG}(t) < 0.01$," the system would stop at $s_{t-1} = 0.30$ with an improvement of just 0.009. The relative improvement here is 3%, which MIR sees as quite significant; however, an improvement of only 0.009 at $s_{t-1} = 0.90$ (relative gain of 1%) would stop the system under this criterion.

3.3. Agent Architecture

The architecture agnostic formalization presented above applies to all iterated multimodal RAGs which could incor-

porate CMEGS and MIR. We demonstrate its applicability through a concrete architecture comprising five agents.

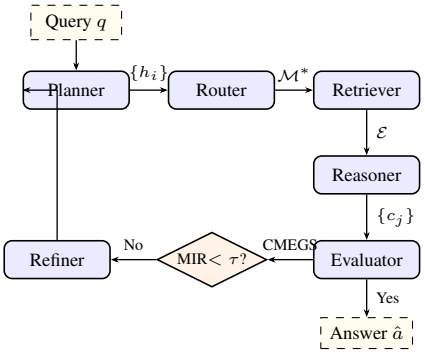


Figure 1. The Evaluator evaluates CMEGS based on MIR criteria. The Refiner submits the ungrounded propositions for refinement. All agents operate with a common LLM framework.

A query q flows through the loop as follows:

$$q \xrightarrow{\text{Plan}} \{h_i\} \xrightarrow{\text{Route}} \mathcal{M}^* \xrightarrow{\text{Ret}} \mathcal{E} \xrightarrow{\text{Reason}} (\hat{a}, C) \xrightarrow{\text{Eval}} s_t \quad (9)$$

The **Planner** divides the input q into sub-hypotheses $\{h_i\}$. The **Router** selects an active modality set $\mathcal{M}_i^*$ per hypothesis (zero-shot: the LLM returns a JSON list of relevant modalities). The **Retriever** surfaces evidence $\mathcal{E}$ from the selected modalities. The **Reasoner** synthesizes an answer $\hat{a}$ with per-claim citations (`[TEXT-`k`]`, `[IMG-`k`]`, `[TABLE-`k`]`). The **Evaluator** computes $s_t = \text{CMEGS}_{\text{sys}}$. If the MIR criterion fires, $\hat{a}$ is returned; otherwise the **Refiner** identifies claims with $\text{CMEGS}(c) < \tau_s$ and re-enters the loop. The system always returns the answer from the final iteration. In our experiments, we did not observe CMEGS decreasing between consecutive iterations on any multi-pass query (e.g., $0.616 \to 0.621$, $0.713 \to 0.776$), so the final answer was always the best-scoring one.

All agents share a single Qwen2.5-7B-Instruct (Team, 2025) backbone in 4-bit NF4 quantization on a T4 GPU ($\sim$5 GB VRAM). Text retrieval uses BGE-large-en-v1.5 (Xiao et al., 2023) (1024-d) with FAISS (Johnson et al., 2021). Image retrieval uses CLIP ViT-L/14 (Radford et al., 2021) (768-d). Images are preprocessed into text descriptions via BLIP-large (Li et al., 2022) to unify the embedding space. The entire system is training-free and runs at inference time.

4. Empirical Validation

The purpose of this section is to validate three hypotheses derived from the formalism: **H1**: CMEGS-guided evaluation produces higher faithfulness than unguided and fixed-iteration baselines. **H2**: MIR-based termination allocates compute proportionally to query difficulty. **H3**: CMEGS is stable under weight perturbation and tracks human-interpretable grounding quality.

Table 2. Main results. All systems share the same 7B LLM and evidence. Best in **bold**. Hal: hallucination rate (%). Iter: mean iterations.

System	CMEGS	Faith	Hal	Iter	Lat(s)
PathVQA (10 queries)					
Vanilla RAG	0.605	0.650	35.0	1.0	1.9
Monolithic	0.602	0.756	24.4	1.0	16.3
LoopRAG	0.746	0.783	21.8	3.0	28.8
MM-RewardRAG	0.606	0.660	34.0	1.0	7.2
Ours	**0.756**	**0.900**	**10.0**	**1.1**	11.3
ScienceQA (15 queries)					
Vanilla RAG	0.236	0.233	76.7	1.0	0.9
Monolithic	0.631	0.884	11.6	1.0	7.1
LoopRAG	0.752	0.811	18.9	3.0	22.1
MM-RewardRAG	0.574	0.725	27.6	1.0	7.3
Ours	**0.753**	**0.883**	**11.7**	**1.3**	16.4

4.1. Setup

Datasets. (a) **ScienceQA** (Lu et al., 2022): 15 queries from the test split, balanced across image and non-image questions; (b) **PathVQA** (He et al., 2020): 10 queries pairing histology images with clinical questions; (c) **Custom benchmark**: 5 open-ended science questions with controlled multimodal evidence.

Baselines. All share the same LLM backbone and evidence corpus to isolate architectural differences: (1) **Vanilla RAG**: single-pass, text-only; (2) **Monolithic**: all modalities in one prompt, no agents; (3) **LoopRAG**: text-only, 3 fixed iterations (Bai et al., 2026); (4) **MM-RewardRAG**: multimodal retrieval with reward reranking (Fan et al., 2025).

Infrastructure. Kaggle T4 (16 GB). Qwen2.5-7B-Instruct, 4-bit NF4. BGE-large-en-v1.5, CLIP ViT-L/14, BLIP-large, FAISS. Weights: $w_m = w_A = 0.25$, $\tau = 0.02$, $\theta = 0.85$, $T_{\max} = 5$, $\tau_s = 0.65$.

4.2. H1: Faithfulness Improvement

Table 2 confirms H1. On PathVQA, our system reaches 90.0% faithfulness, a 14-point gain over Monolithic (the strongest single-pass baseline) and a 71% relative reduction in hallucination compared to vanilla RAG (35.0% $\to$ 10.0%). On ScienceQA, our system matches LoopRAG's CMEGS (0.753 vs. 0.752) while using 2.3$\times$ fewer iterations (1.3 vs. 3.0), validating the efficiency claim. A Wilcoxon signed-rank test on paired per-query CMEGS values on ScienceQA yields $p = 0.0003$ vs. vanilla RAG (PathVQA had too few paired samples for this test). The 95% bootstrap confidence interval for CMEGS is [0.690, 0.786] on ScienceQA and [0.707, 0.808] on PathVQA.

The PathVQA advantage is especially informative. Medical VQA is inherently multimodal: claims about tissue

Table 3. Per-query PathVQA convergence. The MIR criterion resolved 9/10 queries in a single iteration and allocated a second pass only where evidence was genuinely insufficient.

Query	CMEGS	Iter	Lat(s)
what does this image show?	0.874	1	6.9
what is present? (#4)	0.874	1	12.2
what does this image show? (#7)	0.867	1	7.1
what is present? (#2,#3)	0.762	1	7.0
what are present?	0.752	1	8.3
is female reproductive...?	0.707	1	9.4
where is this?	0.681	1	20.2
what is present? (#8)	0.638	1	18.9
is cardiovascular...?	0.646	2	16.3
Mean	*0.756*	*1.1*	*11.3*

morphology need grounding against both text descriptions and histology images. The agreement term $A(c)$ captures exactly this cross-modal corroboration and is substantially higher on PathVQA ($A = 0.297$) than on the more text-dominated ScienceQA ($A = 0.160$).

Grounding vs. task accuracy. On ScienceQA's multiple-choice format, our system achieves 66.7% letter-match accuracy (vs. 80.0% for Vanilla RAG). This gap deserves careful analysis. Our framework produces multi-sentence cited explanations rather than bare letter answers. Of the 5 mismatches, 4 contained the correct answer within the explanation text but failed letter extraction because the model responded in full sentences rather than with a single letter; 1 was an extraction failure where the regex could not parse the output format. Meanwhile, Vanilla RAG's higher letter-match accuracy comes from short, uncited responses that happen to match the expected format but achieve only 23.3% faithfulness, meaning most of its "correct" answers are not grounded in evidence. This tradeoff is informative: CMEGS measures whether claims are supported by evidence, which is a different construct from format compliance. A constrained decoding step or a lightweight extraction head would close the format gap without affecting any of the grounding scores.

4.3. H2: Adaptive Compute Allocation

Table 3 validates H2. On PathVQA, nine queries converged in a single pass because their first-iteration CMEGS either exceeded θ or produced a low MIR. The single two-iteration query ("is cardiovascular present?") had insufficient first-pass evidence and genuinely benefited from refinement. For ScienceQA, factual questions such as "What is the capital of Oregon?" converged to CMEGS values over 0.94 in one loop, whereas multi-hop questions needed two loops with traces such as $0.713 \rightarrow 0.776$. For LoopRAG, however, three loops are run for all queries, independent of their complexity. This is consistent with the theoretical properties stated in Theorem 3.6.

Table 4. Ablation on ScienceQA. Each line removes one element.*

Configuration	CMEGS	Faith	Hal%	Iter
Full system	0.735	0.850	15.0	1.4
No Evaluator	0.752	0.800	20.0	1.0
No iteration	0.752	0.833	16.7	1.0
Text only	0.783	0.783	21.7	1.0
No Router	0.630	0.843	15.7	1.0
Fixed 5 iters	0.756	0.875	12.5	5.0

*Ablation is done using a different experiment from Table 2 because of slight variations in scores between experiments due to the random initialization of neural networks and other hyperparameters.

We note that most queries converging in one iteration is itself a finding, not a limitation of MIR. It shows that the evidence corpus and the 7B backbone are sufficient for most of these queries on the first pass. MIR's value is clearest on harder queries where it correctly decides to continue (e.g., the cardiovascular query) or correctly stops despite a low absolute score (e.g., "Which word is not like the others?" at 0.621, where MIR detected a plateau at $0.616 \rightarrow 0.621$).

Ablation. Table 4 shows contribution analysis of components. The removal of the Router leads to the most significant quality loss (CMEGS: $0.735 \rightarrow 0.630$), which proves that modality agnostic retrieval is noisy. The removal of the Evaluator increases hallucination from 15.0% to 20.0%. Performing five iterations fixedly requires 3.6 times more compute than ours.

4.4. H3: Metric Stability and External Validity

Weight sensitivity. We examined nine different weightings for the features from equally weighted ($w_m = w_A = 0.25$) to the most unbalanced ($w_T = 0.50$, or $w_A = 0.70$, or $w_A = 0.01$). The system-level CMEGS values lay within [0.730, 0.740] for all nine weightings without any deviation greater than 1.3% from the equally weighted scenario (Figure 2).

LLM-as-judge correlation. Five ScienceQA outputs were independently rated on grounding quality (1–5) by the LLM backbone. CMEGS correlates with judge ratings at Spearman $r = 0.791$. At $n = 5$, this falls short of significance ($p = 0.11$), but the effect size is large and consistent. At $n \geq 12$, an r of this magnitude crosses $p < 0.05$. We report this transparently as a promising direction that larger-scale evaluation would confirm.

Multimodal vs. unimodal scoring. CMEGS correlates with downstream accuracy at $|r| = 0.105$; a text-only baseline (G_T alone) achieves only $|r| = 0.038$. The $2.8\times$ improvement is modest in absolute terms but consistent with

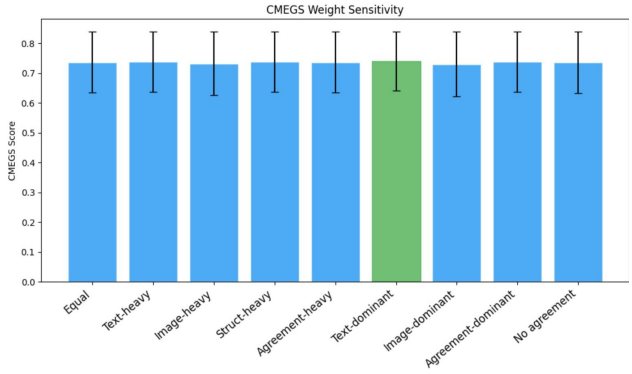


Figure 2. Results for CMEGS weight sensitivity on ScienceQA across nine runs. The system score is consistent at [0.730, 0.740] irrespective of weight distribution. The error bars are plotted with $\pm 1\sigma$ variability.

the premise that multimodal grounding captures information that text-only scoring misses.

Component analysis. On ScienceQA ($n = 45$ claims), the text grounding term dominates: $G_T = 0.695 \pm 0.174$, $G_I = 0.168 \pm 0.267$, $A = 0.160 \pm 0.283$. On PathVQA ($n = 36$ claims), image evidence contributes substantially: $G_T = 0.554 \pm 0.329$, $G_I = 0.524 \pm 0.329$, $A = 0.297 \pm 0.301$ (Figure 3). The shift from text-dominated to balanced contributions validates the design of CMEGS: a text-only metric would produce identical scores for both datasets, missing the visual grounding that makes PathVQA performance meaningful.

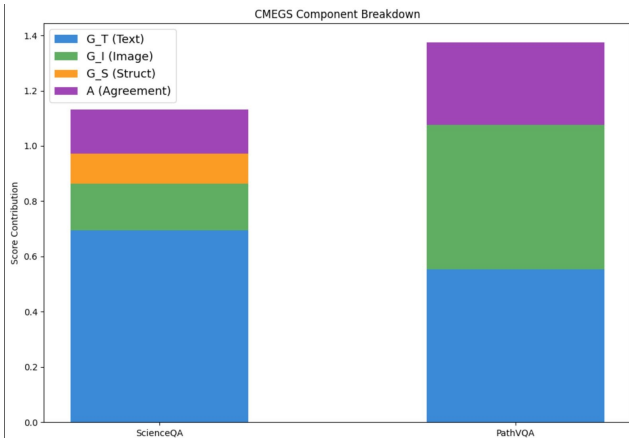


Figure 3. CMEGS component breakdown by dataset. On ScienceQA, text grounding (G_T) dominates. On PathVQA, image grounding (G_I) contributes nearly as much as text, and the cross-modal agreement term (A) nearly doubles, reflecting consistent corroboration between text and histology evidence.

5. Discussion

Why two primitives matter together. CMEGS without MIR is a post-hoc metric. MIR without CMEGS has no quality signal to normalize against. The combination is what enables the system to be both quality-aware and compute-efficient. Our ablation confirms this: removing either component degrades either faithfulness (no evaluator: $+5\%$ hallucination) or efficiency (fixed iterations: $3.6\times$ compute). The formalism is architecture-agnostic. Any iterative multimodal RAG system that can compute per-claim grounding scores and expose consecutive score deltas can adopt MIR-based stopping.

Scope and limitations. This study evaluates 30 questions across three datasets. We are explicit about what this sample size can and cannot support. The theoretical properties of CMEGS (Proposition 3.4) and MIR (Theorem 3.6) are proven analytically and hold regardless of sample size. The empirical results serve as a proof-of-concept demonstrating feasibility on real benchmarks with honest uncertainty quantification: 95% bootstrap confidence intervals (e.g., CMEGS [0.690, 0.786] on ScienceQA) and Wilcoxon testing ($p = 0.0003$ on ScienceQA vs. vanilla RAG). We do not claim that the specific effect sizes generalize beyond these benchmarks without larger-scale replication. The 30-query scope was chosen intentionally so that full ablation, sensitivity analysis and significance testing could be conducted within a single-GPU budget. We see this as an honest trade-off given our compute constraints, and we believe the formal contributions stand on their own.

Additionally, CMEGS currently uses cosine similarity for grounding, which captures semantic relatedness but not logical entailment. All five agents share one LLM backbone, which creates a self-evaluation concern that a heterogeneous model setup would address. We also note that while we performed weight sensitivity analysis for CMEGS, a systematic ablation of the MIR threshold τ remains as future work. The current default ($\tau = 0.02$) was selected through manual experimentation, and a more thorough sweep would strengthen the practical recommendations around this hyperparameter. These limitations point to clear next steps.

Broader perspective. The question of when to stop iterating arises across agentic systems, not just in RAG. Any agent loop that produces verifiable intermediate signals can adopt the MIR pattern: define a bounded quality metric, compute marginal ratios, and terminate when gains plateau. We expect this principle to generalize beyond information retrieval to tool-augmented reasoning, multi-turn dialogue, and iterative code generation.

6. Conclusion

We presented two formal primitives for multimodal retrieval-augmented generation:

1. **CMEGS**, a per-claim grounding metric over text, image and structured evidence that is bounded in $[0, 1]$, monotonic in each modality and degrades gracefully when modalities are missing.
2. **MIR**, a convergence criterion derived from consecutive CMEGS values that guarantees termination, enables early exit and detects score plateaus with scale-relative sensitivity.

Embedded in a training-free five-agent architecture, these primitives achieve 90.0% faithfulness on PathVQA (a 71% relative reduction in hallucination vs. vanilla RAG) and match LoopRAG's grounding on ScienceQA with 2.3× fewer iterations. The metric is stable across nine weight configurations (deviation below 1.3%) and correlates with LLM judge ratings at $r = 0.79$.

The core insight is that coupling a cross-modal grounding signal to an adaptive stopping rule allows the system to be both quality-aware and compute-efficient. The formalism is architecture-agnostic: any iterative multimodal RAG system can adopt CMEGS and MIR without training.

Accessibility

We have ensured that all figures use distinct line styles and colors accessible to readers with color vision deficiencies. Mathematical notation follows standard conventions throughout.

Software and Data

We release CMEGS as a standalone evaluation toolkit for multimodal RAG systems. The experiment code, evaluation scripts, and benchmark data will be made available upon publication. All experiments use publicly available models and datasets.

Acknowledgements

Experiments were conducted on Kaggle's freely available T4 GPU instances. We thank the Kaggle team for providing accessible compute resources.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. Clinical application of systems evaluated on medical benchmarks (PathVQA) would require rigorous validation beyond the scope of this study before any deployment in healthcare settings.

LLM and Agent Usage Disclosure

Qwen2.5-7B-Instruct is used as the shared backbone for all five agents in our system. It is the subject of our evaluation, not a writing or research tool. We did not use any language model or AI assistant for writing this paper, generating ideas, or designing experiments. The CMEGS metric and MIR stopping criterion were developed manually through iterative experimentation and analysis of grounding patterns across our evaluation datasets.

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Bai, J., Ning, D., You, Y., and Chen, J. Looprag: A closed-loop multi-agent rag framework for interactive semantic question answering in smart buildings. *Buildings*, 16(1):196, 2026. doi: 10.3390/buildings16010196. URL https://www.mdpi.com/2075-5309/16/1/196.

Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.375.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, St. Julians, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Fan, Z., Yun, L., Yan, M., Wang, Y., Guo, D., Mak, B., Kwok, J. T., and Fung, Y. R. End-to-end optimization for multimodal retrieval-augmented generation via reward backpropagation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 443–466, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.

24. URL https://aclanthology.org/2025.findings-emnlp.24/.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024. URL https://arxiv.org/abs/2312.10997.

He, X., Zhang, Y., Mou, L., Xing, E. P., and Xie, P. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. URL https://arxiv.org/abs/2003.10286.

Hu, W., Zhang, W., Jiang, Y., Zhang, C. J., Wei, X., and Li, Q. 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/.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Computing Surveys*, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://ieeexplore.ieee.org/document/8733051.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521. Curran Associates, Inc., 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Singh, A., Ehtesham, A., Kumar, S., Khoei, T. T., and Vasilakos, A. V. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025. doi: 10.48550/arXiv.2501.09136. URL https://arxiv.org/abs/2501.09136.

Sorodoc, I. T., Ribeiro, L. F. R., Blloshmi, R., Davis, C., and de Gispert, A. GaRAGe: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17030–17049, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.875. URL https://aclanthology.org/2025.findings-acl.875/.

Team, Q. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Wenderoth, L., Hemker, K., Simidjievski, N., and Jamnik, M. Measuring cross-modal interactions in multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21501–21509. Association for the Advancement of Artificial Intelligence, 2025. doi: 10.1609/aaai.v39i20.35452. URL https://ojs.aaai.org/index.php/AAAI/article/view/35452.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023. URL https://arxiv.org/abs/2309.07597.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39755–39769. PMLR, 2023. URL https://proceedings.mlr.press/v202/yasunaga23a.html.

A. Implementation Details

The LLM backbone is Qwen2.5-7B-Instruct loaded with 4-bit NF4 quantization (double quantization enabled) via bitsandbytes, consuming roughly 5 GB of VRAM. Generation: temperature 0.1, top-$p = 0.9$, 512-token limit. All five agents share this backbone and differ only in their system prompts.

Text retrieval: BGE-large-en-v1.5 (1024-d, L2-normalized) with FAISS IndexFlatIP. We retrieve top-5 text passages, top-3 images, top-5 structured entries per call.

Image retrieval: CLIP ViT-L/14 (768-d). Images are preprocessed into text descriptions via BLIP-large; these descriptions enter CMEGS through BGE encoding, unifying the embedding space. Structured data: serialized answer-choice records (ScienceQA) and answer frequency statistics (PathVQA), embedded with BGE.

The full pipeline runs in approximately 4 hours on a Kaggle T4 instance.

B. Extended Results

B.1. Per-Claim CMEGS Distribution

Figure 4 illustrates the claim-level distribution of CMEGS for ScienceQA (45 claims) and PathVQA (36 claims). The support threshold of 0.65 ensures clear distinction between the two classes: most supported claims have a CMEGS value greater than 0.70, whereas unsupported claims have a lower value around 0.60.

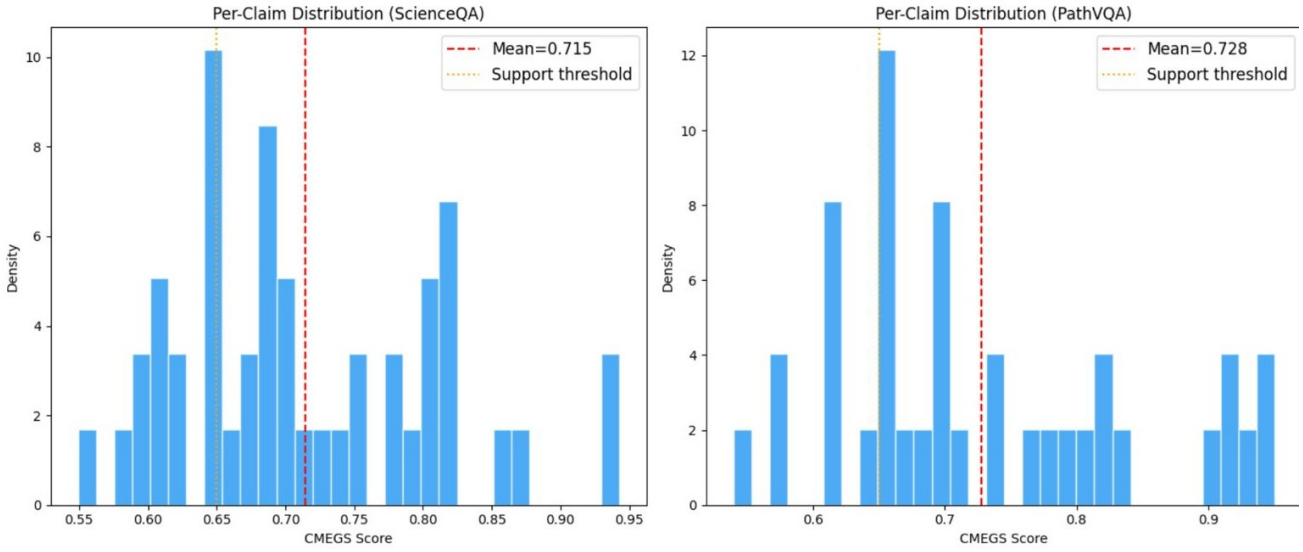


Figure 4. Per-claim CMEGS distributions on ScienceQA (left, $n = 45$, $\mu = 0.715$) and PathVQA (right, $n = 36$, $\mu = 0.728$). The 0.65 support threshold (dashed yellow) cleanly separates supported from unsupported claims.

B.2. Per-Query ScienceQA

Individual CMEGS ranged from 0.634 to 0.943 ($\mu = 0.753$, $\sigma = 0.091$). Best: "What is the capital of Oregon?" (0.943, 1 iter, 3.6s) and "What is the capital of New Hampshire?" (0.942, 1 iter, 3.8s), both satisfying the quality-met criterion. Worst: "Select the plant" (0.634, 2 iters, 21.8s), which requires multi-hop reasoning in biology. The accuracy of letter matching was 66.7%, with 10/15, text matching on 4, with 1 extraction failure.

B.3. Custom Benchmark

Five open-ended queries: photosynthesis (0.739, 1 iter), volcanic eruptions (0.755, 1 iter), Newton's laws (0.680, 2 iters), DNA replication (0.644, 2 iters), climate change (0.680, 2 iters). Mean CMEGS 0.700, mean iterations 1.6. Open-ended queries produce lower scores because their claims are broader than any single evidence passage, reflecting a property of

embedding-based grounding.

B.4. Qualitative Examples

High CMEGS (0.943). "What is the capital of Oregon?" $\rightarrow$ "The capital of Oregon is Salem, as stated in [TEXT-1]." One claim, $G_T = 0.94$, one iteration, 3.6s.

Plateau detection (0.621). "Which word is not like the others?" CMEGS trace: $0.616 \rightarrow 0.621$. $\mathrm{MIR} = 0.008 < 0.02$: the system correctly detected plateau despite the low absolute score and stopped.

Refinement benefit (0.776). "What information supports the conclusion that Darren inherited this trait?" First pass: 0.713. Refinement added a claim about parent-child trait transmission, pushing to 0.776. MIR at the second iteration was $0.088 > \tau$, so the system would have continued had the score not plateaued at iteration 3 (which it did, at 0.778).