Reasoning Acceleration and Intelligence-Density Scaling in Multi-Hop Question Answering

Anonymous ACL submission

Abstract

Small language models (SLMs) fail at multi-hop question answering (QA) when retrieval provides misleading evidence for which reasoning cannot compensate. We introduce the *Reasoning Acceleration Engine* (RAE), which lifts SLM performance on multi-hop QA while keeping the language model out of the retrieval loop and thereby reducing token consumption. RAE builds a bipartite chunk–idea graph and solves a closed-form equation for the chunks most relevant to a query. The equation has a unique non-negative solution, a Lipschitz drift bound under query perturbation — the first closed-form bound for a graph-RAG retrieval operator we know of — and a converging adaptive loop. We test RAE on MuSiQue and HotpotQA against closed-book reading, Vector-RAG (with and without reranking), HippoRAG 2, and RAPTOR, across eight 0.8B-to-9B readers. Evaluated on *intelligence density* — quality per cost metric we introduce — RAE wins all 64 paired comparisons. HippoRAG 2 leads on quality alone at four to five times RAE's cost, and Vector-RAG at constrained chunk budget holds the Pareto-corner on shallow multi-hop, particularly HotpotQA. RAPTOR never reaches the Pareto frontier in any of the 16 cases studied. We show how doing the multi-hop retrieval work algorithmically rather than with a language model wins on intelligence density.

1 Introduction

Large language models scale along two directions: more parameters and longer inference-time reasoning (Snell et al., 2025). Both put intelligence inside the model. We propose a third direction that puts it outside, and explore it with the *Reasoning Acceleration Engine* (RAE) — a closed-form retrieval solver on a bipartite chunk–idea graph, with no LLM in the retrieval loop. For deep multi-hop retrieval on small local readers, doing the multi-hop work algorithmically rather than with the language model wins on intelligence density against every baseline we test, and RAE shows how.

We make three contributions. *First*, intelligence density — the multiplicative composite of quality and generation cost — names a scaling direction that complements model size and inference compute rather than competing with either, and RAE wins on it against every retrieval baseline we test at its full chunk budget (Section 5.1). *Second*, a single closed-form equation does the retrieval, with three theorems pinning down existence and uniqueness, a Lipschitz drift bound, and conductance-update convergence (Section 3.4). *Third*, a predictive controller — plan, retrieve, verify — implements controlled hallucination from predictive-processing theory and scaffolds small-reader generation through typed extraction (Section 3.2). Every empirical result we report is predicted by one of the three theorems developed in Section 3.4.

1.1 Motivation

Designing an LLM-based QA and reasoning system means solving a multi-objective program. Quality, per-query cost, and per-query latency pull in opposite directions, and any deployment chooses a Pareto point on a budget surface:

$$
\begin{array}{llll}
\text{(P)} & \text{maximize}_{\theta\in\Theta} & \log Q(\theta) \\
 & \text{subject to} & \log C(\theta) \le \log \bar{C}, \\
 & & \log T(\theta) \le \log \bar{T}.
\end{array}
$$

Here $Q(\theta) > 0$ is answer quality, $C(\theta) > 0$ is per-query cost (USD or tokens), $T(\theta) > 0$ is per-query latency (seconds), and $\bar{C}, \bar{T} > 0$ are budget upper bounds. A Pareto-optimal $\theta^\star$ balances all three.

The logarithmic form of (P) is essential. Lagrangian relaxation of the budget constraints yields $\log Q - \lambda_C \log C - \lambda_T \log T = \log\big[Q/(C^{\lambda_C} T^{\lambda_T})\big]$, which is a multiplicative Cobb–Douglas utility (Mas-Colell et al., 1995).

Exponentiating gives the *intelligence density* ρ (Definition 1).

Definition 1 (Intelligence density)**.** *For configuration $\theta \in \Theta$ and trade-off weights $\lambda_C, \lambda_T \geq 0$, the* intelligence density *is $\rho(\theta, \lambda_C, \lambda_T) = Q(\theta)/(C(\theta)^{\lambda_C}\, T(\theta)^{\lambda_T})$.*

The ratio $\eta_{AB} = \rho(A)/\rho(B)$ of any given configurations A and B is the *relative-intelligence factor*. At $\lambda_C = \lambda_T = 1$ it reduces to $\eta_{AB} = K \cdot L$ (Definition 2).

Definition 2 (Relative intelligence)**.** *For configurations A and B with shared multipliers, the relative intelligence of A given B is $\eta_{AB} = \rho(A)/\rho(B) = K \cdot L_C^{\lambda_C}\, L_T^{\lambda_T}$, where $K = Q_A/Q_B$ is the quality amplification factor, $L_C = C_B/C_A$ the cost-reduction factor, and $L_T = T_B/T_A$ the latency-reduction factor. The compact form $\eta_{AB} = K \cdot L$ obtains at $\lambda_C = \lambda_T = 1$.*

These definitions cast the third scaling direction as a Pareto trade-off across quality and total cost.

Where graph-RAG quality fails most. Graph-augmented RAG is the dominant paradigm for multi-hop QA. Recent surveys point to two facts. First, retrieval is a prevalent failure mode (Barnett et al., 2024). Second, the retrieve-reason loop that compensates for retrieval gaps is a dominant cost driver for entity-graph variants (Zhou et al., 2025). The failure happens before the language model sees the question: either the knowledge graph was built with wrong or missing entities, or the retriever pulls the wrong chunks. This is consistent with an impossibility result for hallucination control: LLM-internal fixes alone cannot recover answers under ambiguity, and external retrieval is one of the few mitigations the analysis admits (Xu et al., 2024). In our experiments (Appendix F) the answer-bearing chunk is missing from the retrieval slate in 28%, 49%, and 56% of failing 2-, 3-, and 4-hop MuSiQue queries at `qwen3.5:4B`. The cost differential between graph-RAG and dense-RAG baselines is corpus- and method-dependent, ranging from $\sim 1\times$ to $\sim 17\times$ in reported surveys (Zhou et al., 2025, Table 6); on agentic retrieve-reason loops the gap widens further — our cross-family RT-RAG anchor (Appendix J) averages 10.5M tokens per query at `qwen3.5:0.8B` on MuSiQue versus RAE's $\sim$1,030, roughly four orders of magnitude. Dense baselines defer composition to the reader; graph-augmented systems compose at index time; RAE composes at query time through a chain controller without entity–entity edges.

Cost differential via intelligence-density scaling. Consider the multi-hop question *"On what continent is the country in which the director of* MOVIE X *was born located?"* A representative SOTA agentic RAG system, RT-RAG (Shi et al., 2026), answers it through a recursive retrieve-reason tree of $\sim$94 LLM calls and $\sim 10^7$ tokens of context per query. RAE answers the same question through a single closed-form deterministic solver with no LLM in the retrieval inner loop, plus 4–8 LLM calls for planning, per-hop extraction, verification, and generation. At comparable quality at small readers and lower quality at strong readers, the token-cost ratio is two to four orders of magnitude — the intelligence-density framework (§5.1) collapses both regimes into the composite $\eta > 1$ statement.

1.2 Paper organization

Section 2 positions RAE against prior work. Section 3 develops the architecture and proves the three theorems. Section 4 describes the experimental setup. Section 5 reports the empirics and Section 6 concludes.

2 Related Work

A RAG QA system is the composition $\mathrm{QA}(q, \mathcal{D}) = \mathcal{M}\big(\mathcal{R}(q), \mathcal{S}(q, \mathcal{D})\big)$, where $\mathcal{R}$ is the query-understanding stage, $\mathcal{S}$ the structured-evidence stage, and $\mathcal{M}$ the *reader*, the LLM/SLM that consumes retrieved context and emits the final answer. We organize prior work by which factor it modifies.

Dense and graph-augmented retrieval. Dense bi-encoders (Karpukhin et al., 2020; Izacard et al., 2022; Wang et al., 2022) and late-interaction models (Khattab and Zaharia, 2020; Santhanam et al., 2022) dominate single-stage retrieval; graph-augmented work (Edge et al., 2024; Han et al., 2025b; Hu et al., 2025; Luo et al., 2025) encodes entity relationships explicitly. RAE sits between the entity-centric and explicit-graph poles: a bipartite chunk–idea graph (ideas on one side, chunks on the other, no entity–entity edges) with a closed-loop solver rather than single-pass traversal.

Positioning against existing systems. Surveys of graph-RAG (Zhou et al., 2025; Han et al., 2025a; Singh et al., 2025) describe a landscape that includes RAPTOR (Sarthi et al., 2024), KGP (Wang et al., 2024), HippoRAG (Gutiérrez et al., 2024)

and HippoRAG 2 (Gutiérrez et al., 2025), G-retriever (He et al., 2024), ToG (Sun et al., 2024), DALK (Li et al., 2024), the GraphRAG (Edge et al., 2024) and LightRAG (Guo et al., 2024) families, PathRAG (Chen et al., 2025), and RT-RAG (Shi et al., 2026). Most use the LLM at graph-build time, at retrieval-time operator selection, or in iterative retrieve-reason loops. The closest structural neighbor to RAE is HippoRAG 2: both compute the resolvent of a graph operator on external graph input. RAE differs from HippoRAG 2 on two axes. *(i) Retrieval operator.* HippoRAG 2's Personalized PageRank (PPR) is an iterative resolvent of a column-stochastic transition matrix. RAE is a closed-form Schur-complement reasoning operator $\mathbf{x}_\infty = \mathbf{R}(\mathbf{u})\mathbf{r}$ on a symmetric damped Laplacian $H(\mathbf{u}) = \mathbf{B}^\top \mathbf{U}\mathbf{B} + \alpha I$, with a closed-form Lipschitz bound $\gamma(\mathbf{u})$ (Theorem 2; the outer-loop conductance update converges by discrete LaSalle invariance, Theorem 3). *(ii) Graph type.* HippoRAG 2 encodes entity–entity edges. RAE uses a bipartite chunk–idea graph with no entity–entity edges, ruling out edge-noise failure modes that arise when the graph encodes LLM-generated entity–entity relations. Both systems invoke the LLM at index time and at query time (query-entity extraction in HippoRAG 2; planning, extraction, verification, and binding in RAE); the architectural distinction is the retrieval operator's closed form, not LLM placement. More formally, RAE solves a closed-form boundary-value problem for a graph Laplacian, whereas HippoRAG/PPR runs the corresponding random-walk dynamics iteratively to the same fixed point. The closed-form expression gives an upper bound on retrieval drift under query perturbation (Theorem 2), the iterative one does not.

Iterative, agentic, and out-of-LLM RAG. We call retrieval *out-of-LLM* when the structured-evidence factor $\mathcal{S}(q, \mathcal{D})$ is produced by a deterministic polynomial-time solver and the LLM only generates grounded output over that solver's result. The interleaved retrieve-reason paradigm (Yao et al., 2023; Press et al., 2023; Schick et al., 2023; Trivedi et al., 2023; Jiang et al., 2023; Asai et al., 2024; Jeong et al., 2024; Sarthi et al., 2024) keeps reasoning *inside* the LLM forward pass; RT-RAG extends this with five-candidate sampling at depth four. Trained-reasoning systems (Song et al., 2025; Jin et al., 2025) modify $\mathcal{R}$. RAE modifies $\mathcal{S}$ with a closed-form deterministic solver and is compatible with any of them. A recent survey of agentic RAG (Mishra et al., 2026) draws a boundary between single-pass active retrieval and fully agentic retrieve-reason workflows. RAE occupies a third regime the survey names but does not instantiate: closed-form deterministic routing.

Control-theoretic framings. Adaptive control (Åström and Wittenmark, 1995; Lewis et al., 2012) has only recently been adapted to transformer robustness (Nguyen et al., 2024) and LLM safety (Nguyen et al., 2026; Galitsky and Rybalov, 2026). To our knowledge, closed-loop proportional-integral control (with pole-placement tuning) over a document-retrieval process has not been proposed before; full stability connections to Arrow–Hurwicz–Uzawa primal–dual iteration (Arrow et al., 1958) and projected-dynamical-systems / ISS arguments (Khalil, 2002; Nagurney and Zhang, 1996; Facchinei and Pang, 2003) are in Appendix A.

Speculative decoding as the generation-stage analog. Latent-space variants of speculative decoding (Medusa, EAGLE) operate at the level of the target model's hidden states: a cheap proposer in a structured latent space drafts continuations, and the expensive verifier (the full target forward pass) accepts or rejects. RAE has the same architectural shape, applied at the *retrieval* stage rather than the per-token generation stage. The cheap proposer is the deterministic Schur-complement servo (Theorem 2), operating in the structured latent space of the bipartite chunk–idea graph. The expensive verifier is the chain controller's typed extraction step (Section 3.2), which checks each candidate against the planned hop type. REST (retrieval-based draft generation) closes the loop from the other side, drawing draft tokens from a corpus index rather than a draft model. The intelligence-density framework (Section 1) characterizes both as cheap-proposer acceleration: $\eta = K \cdot L$ recovers speedup ($K \approx 1$, L = acceleration factor) in the exact-decoding limit. Combining RAE-style retrieval with REST-style draft generation at the generation stage is left for future work.

Multi-hop benchmarks. HotpotQA (Yang et al., 2018) is the most widely used multi-hop benchmark but is vulnerable to single-hop shortcuts; MuSiQue (Trivedi et al., 2022) patches this with composition-specific questions and is our primary evaluation, with HotpotQA the pre-registered cross-

benchmark gate — an acceptance criterion fixed before the run (we report the gate outcome in the Limitations section). 2WikiMultiHopQA (Ho et al., 2020) is left for future work. Theoretical work on the limits of hallucination control (Xu et al., 2024) and representation-based detectors (Azaria and Mitchell, 2023) motivate our two-level retrieval-time grounding strategy.

3 The Reasoning Acceleration Engine

RAE is a two-level control loop. The inner level — the *semantic-equilibrium servo* — is a deterministic, closed-form, polynomial-time solver that restores the semantic-potential equilibrium of the knowledge graph after a query disturbs it (we call the disturbance the *cognitive gap*). The outer level is a chain controller that orchestrates a planner, an extractor, and a generator around the servo's per-hop output. The two levels run at different timescales, the servo progressing in milliseconds and the chain in seconds. The hierarchy mirrors the controlled-hallucination model of perception (Seth, 2021, 2014; Friston, 2010; Hohwy, 2013), mapping the LLM to a generative prior, the servo to sensory likelihood, and the verifier to a prediction-error channel. We discuss the connection in Appendix K; the formal claims of the paper do not depend on the neuroscientific framing.

3.1 Bipartite chunk–idea graph and aligned retrieval

A corpus $\mathcal{D}$ is segmented into K chunks. An ingest stack emits per-chunk entity mentions with type labels. Mentions are normalized (case-fold, diacritic-fold, alias table) and grouped into S coreferent *ideas*. The chunk–idea biadjacency $\mathbf{C} \in \{0,1\}^{K\times S}$ has $C_{k,s} = 1$ iff chunk k contains a mention resolving to idea s. Chunks and ideas form the two vertex classes. The bipartite construction forbids idea–idea and chunk–chunk edges, ruling out the noisy KG-induced edges of classical KG-RAG. For each idea s we build a textual description $\mathbf{D}_s$ from the chunk sentences containing its mentions (canonical name, surrounding context, and the single-hop questions of L_3 below); $\mathbf{D}_s$ is then embedded through a sentence encoder $\mathcal{E}: \text{text} \to \mathbb{R}^P$ to give the per-idea vector. Stacking these vectors, $\mathbf{V} = [\mathcal{E}(\mathbf{D}_1) \mid \cdots \mid \mathcal{E}(\mathbf{D}_S)] \in \mathbb{R}^{P\times S}$.

We embed ideas at three text levels. L_1 is the name alone, L_2 is the chunk-derived description, and L_3 is the set of single-hop questions answered by the idea (split into *answer-role* and *subject-role* variants). Embedding ideas at any one or more of these levels — we call this *aligned retrieval* — reduces $\|\mathbf{r} - \mathbf{r}^\star\|_2$, which in turn upper-bounds the equilibrium retrieval error through Theorem 2. The choice of embedding configuration is an empirical decision discussed in Appendix C.

3.2 Reasoning architecture as controlled hallucination

A planner-extractor-generator chain calls the servo once per hop. The planner emits an ordered list of typed sub-questions. At each hop the servo retrieves chunks and ideas; a cross-encoder reranker (optional) truncates the top-k; the extractor reads a typed entity from the retrieved context; the verifier checks the type; the binder substitutes the verified entity into the next hop's text. A chain of H hops consumes one planning call, three per-hop LLM calls (read, verify, bind), and one or two answer-generation calls — about eight LLM calls per query in the canonical `qwen3.5:0.8B` configuration (Appendix B). The chain controller is a finite-horizon predictive controller (Mayne et al., 2000; Rawlings et al., 2017; Åström and Wittenmark, 1995) run open-loop: the planner emits the full horizon at hop zero; verifier rejection can retrigger `plan` at the cost of extra planning calls; any constraint violation (failed type check, no candidate evidence) triggers *abstention*.

3.3 Semantic equilibrium servo

For query embedding $\mathbf{q} \in \mathbb{R}^P$, we pick the top-$k$ most similar ideas $\mathcal{Q} \subseteq \{1,\dots,S\}$ and construct a reference vector $\mathbf{r} \in [0,1]^{N_q}$ of normalized top-k cosine similarities. The joint state $\mathbf{x} \in \mathbb{R}^N$ ($N = S + K$) and (query grounding) integrator $\mathbf{z} \in \mathbb{R}^{N_q}$ evolve as

$$\begin{aligned}
\mathbf{x}_{t+1} &= \big[\mathbf{x}_t + \eta_P\big(G(\mathbf{u})\mathbf{r} - H(\mathbf{u})\mathbf{x}_t\big) + \mathbf{P}_q^\top \mathbf{z}_t\big]_+,\\
\mathbf{z}_{t+1} &= \mathbf{z}_t + \eta_I\,\big(\mathbf{r} - \mathbf{P}_q\mathbf{x}_t\big),\\
\mathbf{u}_{t+1} &= \mathbf{u}_t \odot \bar{\boldsymbol{\phi}}_t(\mathbf{r}) \odot \bar{\boldsymbol{\psi}}_t(\mathbf{x}_t),
\end{aligned} \tag{1}$$

Here $\mathbf{P}_q \in \{0,1\}^{N_q\times N}$ is the queried-idea selector, $\mathbf{P}_c$ its complement ($\mathbf{P}_q^\top\mathbf{P}_q + \mathbf{P}_c^\top\mathbf{P}_c = I_N$, $\mathbf{P}_c\mathbf{P}_q^\top = \mathbf{0}$); $\mathbf{B} \in \mathbb{R}^{M\times N}$ is the bipartite signed incidence matrix ($+1$ at the idea endpoint, -1 at the chunk endpoint); $\mathbf{U} = \operatorname{diag}(\mathbf{u})$. The Laplacian and excitation operators are

$$\begin{aligned}
H(\mathbf{u}) &= \mathbf{B}^\top\mathbf{U}\mathbf{B} + \alpha I,\\
G(\mathbf{u}) &= \mathbf{P}_q^\top\mathbf{P}_q\,\mathbf{B}^\top\mathbf{U}\mathbf{B}\,\mathbf{P}_q^\top.
\end{aligned} \tag{2}$$

H is the damped weighted Laplacian (positive-definite for $\alpha > 0$); G is a conductance-weighted excitation kernel that propagates queried-idea potential under the current conductance. The conductance update is multiplicative. Its ignition factor $\bar{\phi}_{t,e}(\mathbf{r}) = \exp(\eta\,((\mathbf{P}_q\mathbf{b}_e)^\top\mathbf{r})^2)$ captures reference excitation across edge e; its suppression factor $\bar{\psi}_{t,e}(\mathbf{x}_t) = \exp(-\frac{\eta}{2}(\mathbf{b}_e^\top\mathbf{x}_t)^2)$ captures edge-voltage suppression.

3.4 Theorems and consequences

Three theorems characterize the servo. *First*, the dynamical system (1) has a unique equilibrium at which idea potentials balance the query's excitation across the knowledge graph (Theorem 1). *Second*, the equilibrium admits a closed-form reasoning operator with a tight Lipschitz bound from the query reference to the chunk state (Theorem 2). *Third*, the multiplicative conductance update converges to its fixed-point set under discrete LaSalle invariance (Theorem 3).

Theorem 1 (Semantic equilibrium)**.** *Fix $\mathbf{u} \in \mathcal{U} = [u_{\min}, u_{\max}]^M$ and a query-reference vector $\mathbf{r} \in \mathbb{R}_+^{N_q}$, and assume $\mathbf{P}_q$ has full row rank. There exists a unique solution $(\mathbf{x}_\infty, \mathbf{y}_\infty)$ to the saddle system*

$$\begin{bmatrix} H(\mathbf{u}) & -\mathbf{P}_q^\top \\ \mathbf{P}_q & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_\infty \\ \mathbf{y}_\infty \end{bmatrix} = \begin{bmatrix} G(\mathbf{u})\,\mathbf{r} \\ \mathbf{r} \end{bmatrix}, \qquad (3)$$

satisfying $\mathbf{x}_\infty \geq 0$ and constituting a fixed point of the projected dynamics (1) with $\mathbf{y}_\infty = \mathbf{z}_\infty/\eta_P$.

The saddle uniqueness follows from the trivial null space of the block matrix (positive definiteness of H on $\ker \mathbf{P}_q$). Non-negativity follows from the structure of $\mathbf{P}_c H \mathbf{P}_c^\top$ and the sign pattern of $\mathbf{P}_c H \mathbf{P}_q^\top$. Full proof is in Appendix A.

Theorem 2 (Semantic reasoning operator)**.** *Assume the hypotheses of Theorem 1 and define the projected Laplacian blocks $H_{cc}(\mathbf{u}) = \mathbf{P}_c H(\mathbf{u}) \mathbf{P}_c^\top$, $H_{cq}(\mathbf{u}) = \mathbf{P}_c H(\mathbf{u}) \mathbf{P}_q^\top$, $G_c(\mathbf{u}) = \mathbf{P}_c G(\mathbf{u})$, along with the chunk-side reasoning operator $\mathbf{T}(\mathbf{u}) = H_{cc}(\mathbf{u})^{-1}(G_c(\mathbf{u}) - H_{cq}(\mathbf{u}))$. Then the chunk component is $\mathbf{x}_c = \mathbf{T}(\mathbf{u})\mathbf{r}$ and the full equilibrium admits the closed form*

$$\mathbf{x}_\infty = \mathbf{R}(\mathbf{u})\mathbf{r} = \mathbf{P}_q^\top \mathbf{r} + \mathbf{P}_c^\top \mathbf{T}(\mathbf{u})\mathbf{r}. \qquad (4)$$

For any oracle reference $\mathbf{r}^\star$ with $\mathbf{x}^\star = \mathbf{R}(\mathbf{u})\mathbf{r}^\star$,

$$\|\mathbf{x}_\infty - \mathbf{x}^\star\|_{H(\mathbf{u})} \leq \gamma(\mathbf{u})\,\|\mathbf{r} - \mathbf{r}^\star\|_2, \qquad (5)$$

where $\gamma(\mathbf{u}) = \sqrt{\lambda_{\max}\big(R(\mathbf{u})^\top H(\mathbf{u}) R(\mathbf{u})\big)}$.

The constant $\gamma(\mathbf{u})$ in (5) is the *operator-norm Lipschitz constant* of $\mathbf{R}(\mathbf{u})$ from the Euclidean reference space into the chunk space with the weighted-Laplacian inner product $v^\top H(\mathbf{u})\,w$. It is computed by a few iterations of Lanczos (Golub and Van Loan, 2013) on the sparse symmetric positive-definite matrix $\mathbf{R}(\mathbf{u})^\top H(\mathbf{u})\mathbf{R}(\mathbf{u})$, each iteration costing one matrix–vector product against the bipartite Laplacian and one sparse Schur-complement solve (with m Lanczos steps the total work is $O(m \cdot \mathrm{nnz}(\mathbf{B}))$ where $\mathrm{nnz}(\mathbf{B})$ is the number of chunk–idea edges, and $m \leq 20$ suffices for four-digit accuracy on the corpus sizes we consider). On a 50-query MuSiQue subset the bound holds in $50/50$ instances. Under structured perturbations of $\mathbf{r}$ (drop, weight-shock, Gaussian; $n = 2{,}485$ paired), the bound holds with zero violations. Full protocols are in Appendix D.

Theorem 3 (Adaptive convergence)**.** *Fix vector $\mathbf{r} \in \mathbb{R}_+^{N_q}$ and let $\mathcal{U} = [u_{\min}, u_{\max}]^M$, $u_{\min} > 0$. Define the semantic energy $V(\mathbf{u}, \mathbf{r}) = \frac{1}{2}\mathbf{x}_\infty^\top H(\mathbf{u})\mathbf{x}_\infty - \mathbf{x}_\infty^\top G(\mathbf{u})\mathbf{r}$. If conductance evolves by the multiplicative-clipped update*

$$\begin{aligned} \mathbf{u}_{t+1} &= \mathrm{clip}_{\mathcal{U}}\big(\mathbf{u}_t \odot \exp(-\eta\,\nabla V(\mathbf{u}_t, \mathbf{r}))\big) \\ &\equiv \mathbf{S}_\eta(\mathbf{u}_t), \end{aligned} \qquad (6)$$

then for any $\eta > 0$ and $\mathbf{u}_0 \in \mathcal{U}$, the energy sequence $\{V(\mathbf{u}_t, \mathbf{r})\}_t$ is monotone non-increasing and converges to some $V_\infty \in [V_\star, V(\mathbf{u}_0, \mathbf{r})]$, the fixed-point set $\mathcal{S}_\star$ is non-empty, and $\lim_{t\to\infty} \mathrm{dist}(\mathbf{u}_t, \mathcal{S}_\star) = 0$.

Remark (boundary saturation). Since $V(\cdot, \mathbf{r})$ is concave on $\mathcal{U}$, the limit set $\mathcal{S}_\star$ generically lies on the boundary $\partial\mathcal{U}$. The multiplicative-clipped update saturates conductances at $u_{\min}$ or $u_{\max}$ rather than reaching interior critical points. We do not claim convergence to the global minimizer of V. The proof is in Appendix A; empirical verification of the descent on representative MuSiQue trajectories is in Appendix E.

4 Experimental Setup

Benchmarks. MuSiQue-200 is our primary multi-hop evaluation: 200 questions sampled from MuSiQue-Ans. HotpotQA-500 is the pre-registered cross-benchmark gate: 500 questions from HotpotQA dev-hard, 406 bridge plus 94 comparison, drawn by the same sampling procedure. The retrieval corpus per benchmark is the union of all gold and distractor paragraphs from the sampled questions: ∼3,370 paragraphs (MuSiQue) and

$\sim$4,955 paragraphs (HotpotQA). The subset protocol follows HippoRAG 2 and the wider multi-hop-RAG literature.

Readers. The *reader* is the language model that consumes retrieved context and emits the final answer. Eight open-weight readers anchor the comparison, all served by vLLM in non-thinking mode: `qwen3.5:{0.8B, 4B, 9B}`, `google/gemma-4-E2B-it` and `google/gemma-4-E4B-it`, `llama3.2:{1B, 3B}`, and `phi4-mini:3.8B`. The reader of each cell doubles as that cell's ingest model — the matched-ingest protocol that drives the cross-family generalization.

Baselines. At each reader-benchmark cell we run four canonical baseline families. Closed-book LLM-only reads the question with no retrieved context. Vector-RAG retrieves the top k chunks by dense cosine similarity ($k \in \{5, 10\}$). Vector-RAG with reranking adds a cross-encoder rerank over an initial pool of 30 ($k \in \{5, 10\}$ output). HippoRAG 2 (Gutiérrez et al., 2025) runs Personalized PageRank on its entity-graph index; RAPTOR (Sarthi et al., 2024) uses hierarchical tree summarization. The external RT-RAG (Shi et al., 2026) (recursive tree, branching 5, depth 4) anchors the high-cost tail. All systems share the same chunk policy, embedding model, and reader-matched ingest LLM at each cell.

Metric and protocol. Our primary metric, *Ext-F_1*, is the SQuAD/MuSiQue token-level answer-F_1 (Trivedi et al., 2022; Yang et al., 2018) applied to a short-answer span extracted by a deterministic three-step heuristic (Appendix H: verbatim gold-substring match if present, otherwise a sentence-extraction heuristic). The same extraction runs on every system. Reader verbosity varies sharply between 0.8B and 9B, so we also report *raw F_1* (the metric applied to the reader's full output) in Appendix C. The extraction inflates verbose readers more than terse ones — $6.9\times$ on RAE at `qwen3.5:0.8B` on MuSiQue versus $1.4\times$ on HippoRAG 2 at the same cell — which can flip the cross-system ranking at the smallest-reader cells. Reporting both metrics makes the inversion visible. We do not use LLM-as-judge scoring (Kamoi et al., 2024) because judge-prompt drift damages exact reproducibility.

Cost axis. We report mean prompt + completion tokens per query as the per-query deployment cost. Index-time tokens are one-time and amortized across all queries. The canonical RAE L1+L2 ingest totals $\approx$ 4.5M tokens on MuSiQue and $\approx$ 7.3M on HotpotQA (prompt + completion), with break-even versus Vector-RAG at roughly 12,000 MuSiQue queries and 19,000 HotpotQA queries.

5 Results

5.1 Intelligence-density results

We evaluate RAE against four canonical retrieval baseline families: closed-book reading (LLM-only), dense single-shot (Vector-RAG at $k \in \{5, 10\}$), dense single-shot with cross-encoder reranking (VRAG + Reranker at $k \in \{5, 10\}$), iterative multi-hop via Personalized PageRank (HippoRAG 2), and tree summarization (RAPTOR). Eight open-weight readers spanning 0.8B $\rightarrow$ 9B parameters and two benchmarks (MuSiQue-200, HotpotQA-500) give 16 reader-benchmark cells. Every system uses the reader as its ingest model.

For each cell and each baseline b we compute the relative-intelligence factor $\eta_{\text{RAE},b} = \rho_{\text{RAE}}/\rho_b = K \cdot L$ at $\lambda_C = \lambda_T = 1$, latency omitted ($\lambda_T = 0$, per-query latency tracks token cost in our local-tier setup). The quality factor $K = Q_{\text{RAE}}/Q_b$ is the per-cell Ext-F_1 ratio; the cost factor $L = C_b/C_{\text{RAE}}$ is the per-query-tokens ratio. We summarize the per-cell η values by their *geometric mean* across the 16 reader-benchmark cells — the natural average for a multiplicative ratio, and the one quoted as η_{gmean} in Table 1.

Reading the result. By full retrieval budget we mean each baseline at its standard configuration: $k = 10$ chunks for the Vector-RAG variants, the full retrieve-reason loop for HippoRAG and RAPTOR. Two structural facts explain the 64/64 sweep. The first concerns per-query iteration cost. HippoRAG runs two LLM calls per query (entity extraction + answer generation) and iterates Personalized PageRank to convergence between them — tens of sparse mat-vec products on the entity graph; RAPTOR similarly does most of its work in iterative LLM-free tree-summarization passes at index and query time. RAE's closed-form Schur-complement servo (Theorem 2) replaces both the iterative PPR and the tree traversal with a single sparse linear solve, and the saving holds uniformly across the eight readers because the servo's cost scales with graph size, not reader speed. This is

Baseline b	Cells $\eta > 1$	$\eta_{\min}$	η_{median}	$\eta_{\max}$	η_{gmean}
Full retrieval budget ($k=10$)					
Vector-RAG ($k=10$)	16/16	1.12	1.61	2.11	1.58×
Vector-RAG + Reranker ($k=10$)	16/16	1.09	1.47	2.35	1.53×
RAPTOR	16/16	2.01	2.48	3.22	2.44×
HippoRAG 2	16/16	2.78	3.90	5.25	3.90×
Full-budget pooled (4×16=64)	**64/64**	**1.09**	**2.09**	**5.25**	**2.19×**
Budget-trimmed retrieval ($k=5$)					
Vector-RAG ($k=5$)	3/16	0.58	0.81	1.15	0.83×
Vector-RAG + Reranker ($k=5$)	3/16	0.55	0.77	1.17	0.80×

Table 1: **Intelligence-density results.** Relative-intelligence factor $\eta_{\text{RAE},b} = K \cdot L$ for the canonical RAE-cfg2 (servo + cross-encoder reranker; best of L1, L1+L2, L1+L3 per cell) against four full-budget baseline families across all 16 reader-benchmark cells (8 readers × 2 benchmarks). At full budget, RAE achieves $\eta > 1$ in $64/64$ pairs, $\eta \in [1.09, 5.25]$; HippoRAG cross-family produces the strongest η (gmean $3.90\times$). Budget-trimmed dense baselines ($k=5$, ∼700–1,200 tokens/q across the slate) hold a cost-axis Pareto-corner advantage in $13/16$ cells against RAE; this is a budget-elasticity effect, not a quality reversal (Section 5.2).

what carries the geometric-mean factors of $3.90\times$ against HippoRAG and $2.44\times$ against RAPTOR. The second concerns context structure. Vector-RAG at $k = 10$ feeds the reader a monolithic top-k context whose tokens dominate per-query cost. RAE's chain controller decomposes the question into per-hop sub-questions with smaller per-call contexts. Vector-RAG with reranking does edge RAE on Ext-F_1 at three cells — qwen-4B on HotpotQA by $+0.14$, gemma-4B on both benchmarks by $+0.08$ and $+0.06$ — but a 1.37–$1.63\times$ cost gap still flips the composite in RAE's favor at those cells. Under $\eta = K \cdot L^{\lambda_C}$, $64/64$ holds for $\lambda_C \in [0.75, 1.0]$ and falls only to $62/64$ at the quality-heavy $\lambda_C = 0.5$.

Budget-trimmed Vector-RAG holds a Pareto-corner cost advantage. Vector-RAG at $k = 5$ chunks runs at about 700–1,200 tokens per query across the eight-reader slate (below 1,000 at the qwen tier, up to ∼1,200 at verbose-output readers), versus RAE's 1,000–1,200. The quality loss is small and the cost saving is structural. Under $\eta = K \cdot L$ the composite flips, with VRAGrr$[k = 5]$ and VRAG$[k=5]$ achieving $\eta > 1$ against RAE in all but three cells each. This is exactly the Pareto-corner behavior the intelligence density framework predicts — same frontier, different operating point. Section 5.2 maps the Pareto set.

5.2 Pareto set across the reader slate

Figure 1 shows the cost-quality picture across the qwen tier; Appendix Figures 4–5 extend to raw F_1 and the full cross-family slate. The Pareto picture across all 16 cells (noise-aware, $\varepsilon_q = 0.02$) falls into four regimes. RAE-cfg2 alone occupies the non-closed-book frontier in six cells, mostly MuSiQue on llama, phi, and gemma readers. It shares the frontier with HippoRAG in five cells on the qwen tier of MuSiQue — RAE at the low-cost end, HippoRAG buying quality at roughly four to five times the cost — and with VRAGrr in three cells (gemma-4B on both benchmarks, qwen-4B on HotpotQA) where VRAGrr's $+0.05$ to $+0.14$ Ext-F_1 premium clears the noise band. RAE-cfg2 sits off the frontier in roughly six cells, mostly HotpotQA, where the $k=5$ dense trims match its quality at lower cost. RAPTOR is faded in every panel. At qwen-9B on HotpotQA, HippoRAG is itself Pareto-dominated by VRAGrr at full budget, delivering the same quality at a third of the cost: iterative LLM extraction stops paying its way at the strongest reader on the easier benchmark.

5.3 Per-hop retrieval advantage

We group MuSiQue by hop count, $k \in 2, 3, 4$, about 65 queries per hop-bucket per reader. The per-reader curve is non-monotone. Most readers score worst at three hops and recover at four. Iterative methods peak on three-hop bridge questions. RAE's closed-form solver catches up on longer chains, as it should by theoretical properties. The four-hop reversal is consistent against the Vector-RAGs and RAPTOR — six or more of eight readers favor RAE. Against HippoRAG the split is even, so we do not claim a deep-multi-hop quality lead.

Theorem 2 explains the pattern. The Schur-complement resolvent $H_{cc}(\mathbf{u})^{-1}$ captures all multi-hop walks through the bipartite graph in closed form — a deterministic linear-algebraic operator that can be applied without ever calling the lan-

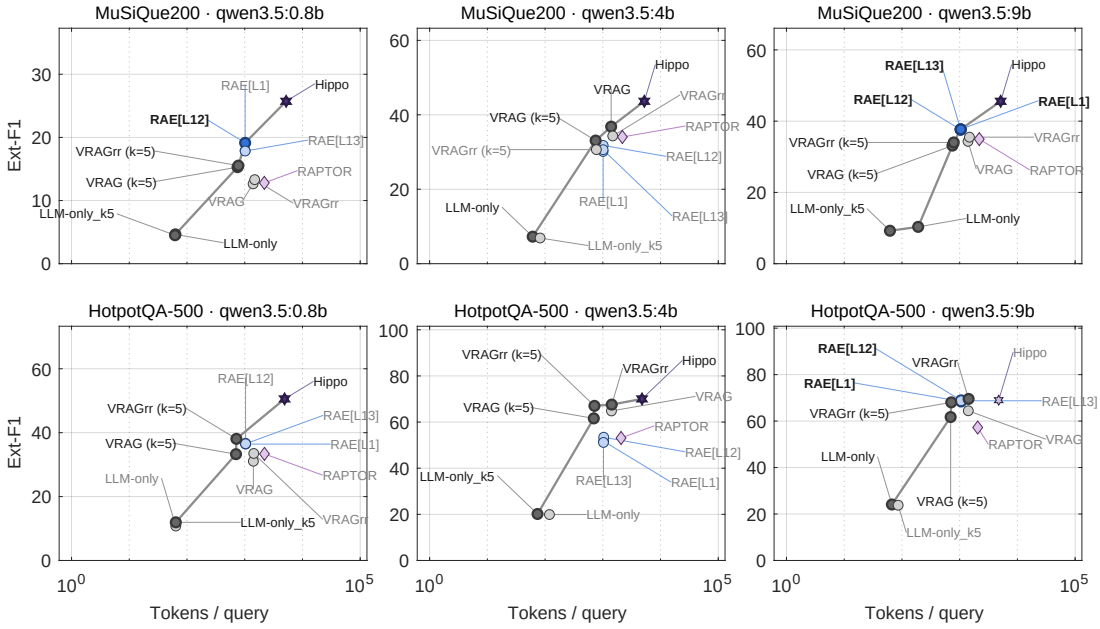

Figure 1: Pareto frontier in (tokens per query, Ext-F_1) space across the qwen-tier six-cell grid (rows: MuSiQue, HotpotQA; columns: 0.8B, 4B, 9B). Filled markers sit on the frontier; faded markers are dominated. Hollow stars at hotpot/qwen-9B mark HippoRAG 2 dominated by Vector-RAG with reranking. Eight-reader cross-family and raw-F_1 versions are in Appendix Figures 4–5.

guage model. Iterative LLM extraction does the same composition step by step with the LLM in the loop, accumulating error at each hop. Single-shot Vector-RAG skips multi-hop composition altogether, failing on longer chains because there is no mechanism to compose facts across hops.

6 Conclusion

We close with three contributions, each with a clear line to impact.

Optimizing for *intelligence density* presents a third scaling direction for multi-hop QA, complementing model size and inference-time compute and providing an honest yardstick for comparing systems across multiple deployment constraints.

The *Reasoning Acceleration Engine* (RAE) is engineered for that direction. Two design choices drive the win. First, retrieval moves outside the language model: a sparse Schur-complement solve over a bipartite chunk–idea graph replaces the iterative LLM-driven retrieval of agentic systems, so per-query LLM cost stops scaling with chain depth. Second, the same Schur complement aggregates bipartite walks, so the multi-hop signal that iterative pipelines compound error on is reduced — quality holds while cost falls, which is the direction intelligence density rewards. Three theorems make the operator analytically tractable: a unique non-negative semantic equilibrium, an operator-norm Lipschitz drift bound with closed-form constant. To our knowledge that is the first retrieval-stability certificate published for a graph-RAG operator, opening formal robustness analysis that iterative agentic pipelines do not admit — and discrete-LaSalle convergence of the conductance update.

The experiments tell the same story. Across investigated reader-benchmark pairs covering eight 0.8B-to-9B readers on MuSiQue and HotpotQA, RAE delivers more intelligence density than every baseline at full retrieval budget. The relative-intelligence factor $\eta = \rho_{\text{RAE}}/\rho_{\text{baseline}}$ is greater than 1 in all paired comparisons, ranging from 1.09 to 5.25. Removing the servo collapses retrieval F_1 across the `qwen` tier, a lift that is flat with reader size. The servo, not the LLM around it, carries the multi-hop signal.

For deep multi-hop retrieval on small local readers, doing the multi-hop work algorithmically rather than with the language model wins on intelligence density against every baseline we test.

Limitations

What we do not claim. We do not claim absolute SOTA at every reader on every metric. HippoRAG holds the quality lead at a multiple of RAE's per-query token cost; Vector-RAG with reranking beats RAE by a small margin on Ext-F_1 at modestly higher cost. The more consequential limit is budget-trimmed Vector-RAG, which holds a Pareto-corner cost advantage on shallow multi-hop, particularly HotpotQA. At reader scales above what we evaluate, recent work — PropRAG (Wang and Han, 2025) — reports stronger absolute scores at Llama-3.3-70B; we do not evaluate at that scale. RAE is a deployment lever for deep multi-hop retrieval on local readers — the regime where the closed-form resolvent's structural advantage and explainability pay their way.

Canonical-pinning vs tuning. All primary numbers and Pareto figures use a single canonical RAE configuration (L1+L2) pinned uniformly across readers, panels, and plots, pre-empting the post-hoc selection-bias concern that variants could be picked per panel to lift the y-axis. We sweep alternative embedding configurations (L1, L1+L2+L3, L3-only) in Appendix C and retain L1+L2 as canonical. Per-cell tuned configurations are reported in the ablation appendix only.

Quality vs cost trade-off. We do not claim absolute Ext-F_1 SOTA at every reader-benchmark cell. HippoRAG holds the quality lead at qwen-tier cells at a small multiple of RAE's per-query token cost; VRAGrr at full budget holds an Ext-F_1 premium over RAE-cfg2 in a minority of cells. Budget-trimmed dense baselines hold a Pareto-corner cost advantage on shallow multi-hop. We report these honestly in the Pareto grid (Appendix Figure 1); they co-occupy the frontier with RAE-cfg2 rather than dominating it. Our claim is composite η-dominance at the full retrieval budget and Pareto-frontier membership in the deep-multi-hop regime.

HotpotQA cross-benchmark. The pre-registered HotpotQA cross-benchmark gate did not pass at our local-tier protocol: on HotpotQA most readers see RAE-cfg2 Pareto-dominated by budget-trimmed dense baselines. HotpotQA is predominantly 2-hop bridge, where the closed-form resolvent's multi-hop reach is wasted relative to a smaller dense candidate pool with cross-encoder reranking. We report Pareto-frontier membership as the cross-benchmark check.

Reader-family coverage. Our evaluation spans eight open-weight readers from four organizations (qwen3.5, gemma-4, llama3.2, phi4-mini) under a matched-ingest vLLM protocol. We do not include closed-API readers in this version.

Hallucination control architecture. Theorem 1 gives a level-1 retrieval-time grounding guarantee. The level-2 generation-time grounding (chain planning, entity completion, typed verification) is architectural. We do not report a hallucination-benchmark score in this paper. A targeted evaluation of that layer is left for future work.

Ethical considerations

Data. Evaluations use MuSiQue-Ans (Trivedi et al., 2022) and HotpotQA (Yang et al., 2018), both publicly released for research use under permissive licenses (MuSiQue: CC BY 4.0; HotpotQA: CC BY-SA 4.0). The MuSiQue-200 and HotpotQA-500 subsets are deterministic seed-stratified samples from the original dev splits; no private or personally identifying data is introduced.

Models. All readers are open-weights checkpoints (`qwen3.5:{0.8B,4B,9B}`, `google/gemma-4-E2B-it`, `google/gemma-4-E4B-it`, `llama3.2:{1B,3B}`, `phi4-mini:3.8B`); each cell's reader doubles as that cell's ingest LLM. The text embedder is `qwen3-embedding`. All models are used under their published licenses.

Compute and infrastructure. All experiments were run on a single consumer-grade laptop GPU (NVIDIA GeForce RTX 4070, 8 GB VRAM, CUDA 12) using vLLM as the inference server in non-thinking mode. The full evaluation slate of 8 readers $\times$ 2 benchmarks $\times$ $\sim$6 retrieval methods (plus ablations) fits within this single-GPU budget; per-query and per-index token totals across all methods are reported in Table 5, and per-query wall-clock latencies are recorded in the per-query JSON in the anonymized supplementary archive.

Human subjects. The work involves no human-subjects research, no crowdsourced annotation, and no personally identifying or sensitive content.

Reproducibility. Per-query JSON results, seed-stratification scripts, and token-cost logs are included in the anonymized supplementary archive.

Generative-AI assistance. Generative AI assistants were used during drafting for prose polishing and minor code-completion suggestions. All technical content, experimental design, theorem statements and proofs, and empirical claims are the authors' own. No AI-generated text is incorporated without thorough authorial review.

References

Pedram Agand. 2026. Neuro-symbolic financial reasoning via deterministic fact ledgers and adversarial low-latency hallucination detector. arXiv:2603.04663.

Kenneth J. Arrow, Leonid Hurwicz, and Hirofumi Uzawa. 1958. *Studies in Linear and Non-Linear Programming*. Stanford University Press.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*.

Karl Johan Åström and Björn Wittenmark. 1995. *Adaptive Control*, 2nd edition. Addison-Wesley.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*.

Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2025. PathRAG: Pruning graph-based retrieval augmented generation with relational paths. *Preprint*, arXiv:2502.14902.

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint*.

Francisco Facchinei and Jong-Shi Pang. 2003. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, New York, NY.

Karl Friston. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Boris Galitsky and Alexander Rybalov. 2026. Neuro-symbolic verification for preventing llm hallucinations in process control. *Processes*, 14(2).

Gene H. Golub and Charles F. Van Loan. 2013. *Matrix Computations*, 4th edition. Johns Hopkins University Press, Baltimore, MD.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. *arXiv preprint*.

Haoyu Han, Li Ma, Yu Wang, Harry Shomer, Yongjia Lei, Zhisheng Qi, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, Charu C. Aggarwal, and Jiliang Tang. 2025a. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights. *arXiv preprint*. ArXiv:2502.11371 [cs.IR].

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025b. Retrieval-Augmented Generation with Graphs (GraphRAG). ArXiv:2501.00309 [cs.IR].

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of COLING*.

Jakob Hohwy. 2013. *The Predictive Mind*. Oxford University Press.

Yuntong Hu, Zhihan Lei, Zhongjie Dai, Allen Zhang, Abhinav Angirekula, Zheng Zhang, and Liang Zhao. 2025. Cg-rag: Research question answering by citation graph retrieval-augmented llms. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 678–687, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of NAACL-HLT 2024*.

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

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Preprint*, arXiv:2503.09516.

Ryo Kamoi, Tanya Goyal, Mona Diab, Jingfei Du, Mohit Iyer, Rui Zhang, Mounica Maddela, Xanh Ho, Wenpeng Yang, and Mohit Bansal. 2024. Evaluating LLM-as-a-judge for extractive question answering: Correlation with human judgment across token-F1, EM, and semantic metrics. *Preprint*, arXiv:2504.11972.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Hassan K. Khalil. 2002. *Nonlinear Systems*, 3 edition. Prentice Hall, Upper Saddle River, NJ.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48. ACM.

J. P. LaSalle. 1960. Some extensions of Liapunov's second method. *IRE Transactions on Circuit Theory*, 7(4):520–527.

Frank L. Lewis, Draguna Vrabie, and Kyriakos G. Vamvoudakis. 2012. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 32(6):76–105.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, Huan Liu, Li Shen, and Tianlong Chen. 2024. DALK: Dynamic co-augmentation of llms and knowledge graphs to answer alzheimer's disease questions with scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

David G. Luenberger. 1979. *Introduction to Dynamic Systems: Theory, Models, and Applications*. John Wiley & Sons, New York.

Linhao Luo, Zicheng Zhao, Reza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. 2025. GFM-RAG: Graph foundation model for retrieval augmented generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Andreu Mas-Colell, Michael Dennis Whinston, and Jerry R. Green. 1995. *Microeconomic theory*, volume 1. Oxford University Press.

David Q. Mayne, James B. Rawlings, Christopher V. Rao, and Pierre O. M. Scokaert. 2000. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814.

Saroj Mishra, Suman Niroula, Umesh Yadav, Dilip Thakur, Srijan Gyawali, and Shiva Gaire. 2026. SoK: Agentic retrieval-augmented generation (RAG): Taxonomy, architectures, evaluation, and research directions. arXiv:2603.07379.

Anna Nagurney and Ding Zhang. 1996. *Projected Dynamical Systems and Variational Inequalities with Applications*. Kluwer Academic, Boston, MA.

Dung Viet Nguyen, Yen Nhi Pham, Hieu M. Vu, Lei Zhang, and Tan Minh Nguyen. 2026. Activation steering with a feedback controller. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Tam Nguyen, César A. Uribe, Tan M. Nguyen, and Richard G. Baraniuk. 2024. PIDformer: Transformer meets control theory. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP*.

James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. 2017. *Model Predictive Control: Theory, Computation, and Design*, 2nd edition. Nob Hill Publishing.

R Tyrrell Rockafellar. 1997. *Convex analysis*, volume 28. Princeton university press.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of NAACL-HLT 2022*, pages 3715–3734.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Anil Seth. 2021. *Being You: A New Science of Consciousness*. Faber and Faber.

Anil K. Seth. 2014. A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5(2):97–118.

Yuling Shi, Maolin Sun, Zijun Liu, Mo Yang, Yixiong Fang, Tianran Sun, and Xiaodong Gu. 2026. Reasoning in trees: Improving retrieval-augmented generation for multi-hop question answering. *Preprint*, arXiv:2601.11255.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. In *Findings of IJCNLP-AACL 2025*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Machine Learning (ICML)*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations (ICLR)*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jingjin Wang and Jiawei Han. 2025. PropRAG: Guiding retrieval with beam search over proposition paths. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6212–6227, Suzhou, China. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint*.

Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2026. On the theoretical limitations of embedding-based retrieval. In *International Conference on Learning Representations (ICLR)*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yingli Zhou, Yaodong Su, Youran Sun, Shu Wang, Taotao Wang, Runyuan He, Yongwei Zhang, Sicong Liang, Xilin Liu, Yuchi Ma, and Yixiang Fang. 2025. In-depth analysis of graph-based RAG in a unified framework. *Proceedings of the VLDB Endowment*, 18(13):5623–5637.

A Proofs of Theorems

Proof of Theorem 1 (Semantic equilibrium)

First, we establish that the left-hand-side block matrix in (3) has trivial nullspace and therefore is non-singular, guaranteeing a unique solution $(\mathbf{x}_\infty, \mathbf{y}_\infty)$. Second, we demonstrate that this unique solution is non-negative and satisfies the steady-state conditions of the projected dynamical system ((1)).

Suppose $\mathbf{r} = 0$ and consider the system of equations:

$$H(\mathbf{u})\mathbf{v} - \mathbf{P}_q^\top \mathbf{w} = \mathbf{0} \quad \text{and} \quad \mathbf{P}_q \mathbf{v} = \mathbf{0}. \tag{7}$$

Left-multiplying the first one by $\mathbf{v}^\top$ gives:

$$\mathbf{v}^\top H(\mathbf{u})\mathbf{v} - (\mathbf{P}_q \mathbf{v})^\top \mathbf{w} = 0. \tag{8}$$

It follows from the second equation that $\mathbf{v}^\top H(\mathbf{u})\mathbf{v} = 0$. Since $H(\mathbf{u}) \succ 0$ as damped Laplacian, i.e., $H(\mathbf{u}) = \mathbf{B}^\top \mathbf{U}\mathbf{B} + \alpha I \succeq \alpha I \succ 0$ for $\alpha > 0$, that is possible if and only if $\mathbf{v} = \mathbf{0}$. By assumption, $\mathbf{P}_q$ has full row rank, so its transpose $\mathbf{P}_q^\top$ has linearly independent columns, which implies $\mathbf{P}_q^\top \mathbf{w} = \mathbf{0}$ for $\mathbf{w} = \mathbf{0}$. Since the null space is trivial, i.e., it contains only the zero vector, there exists a unique solution $(\mathbf{x}_\infty, \mathbf{y}_\infty)$ to ((3)).

Fix $\mathbf{r} \neq \mathbf{0}$, let $(\mathbf{x}_\infty, \mathbf{y}_\infty)$ be this unique solution, and set $\mathbf{z}_\infty = \eta_P \mathbf{y}_\infty$. Substituting into (1) gives

$$\begin{aligned} &\eta_P\big(G(\mathbf{u})\mathbf{r} - H(\mathbf{u})\mathbf{x}_\infty\big) + \mathbf{P}_q^\top \mathbf{z}_\infty \\ &\quad = \eta_P\big(-\mathbf{P}_q^\top \mathbf{z}_\infty/\eta_P\big) + \mathbf{P}_q^\top \mathbf{z}_\infty \\ &\quad = \mathbf{0}. \end{aligned} \tag{9}$$

Similarly, $\mathbf{z}_\infty + \eta_I\left(\mathbf{r} - \mathbf{P}_q\mathbf{x}_\infty\right) = \mathbf{z}_\infty + \mathbf{0} = \mathbf{z}_\infty$, so $(\mathbf{x}_\infty, \mathbf{y}_\infty)$ satisfies the steady-state conditions.

It remains to show that $\mathbf{x}_\infty \geq 0$. By the constraint equation in (3),

$$\mathbf{x}_q = \mathbf{P}_q\mathbf{x}_\infty = \mathbf{r} \geq 0. \tag{10}$$

Using the selector decomposition

$$\mathbf{x}_\infty = \mathbf{P}_q^\top \mathbf{x}_q + \mathbf{P}_c^\top \mathbf{x}_c = \mathbf{P}_q^\top \mathbf{r} + \mathbf{P}_c^\top \mathbf{x}_c, \tag{11}$$

and multiplying the first block row of (3) by $\mathbf{P}_c$, we obtain

$$\mathbf{P}_c H(\mathbf{u})\mathbf{x}_\infty = \mathbf{P}_c G(\mathbf{u})\mathbf{r}, \tag{12}$$

because $\mathbf{P}_c\mathbf{P}_q^\top = 0$. Therefore

$$H_{cc}(\mathbf{u})\mathbf{x}_c + H_{cq}(\mathbf{u})\mathbf{r} = G_c(\mathbf{u})\mathbf{r}, \tag{13}$$

or equivalently

$$H_{cc}(\mathbf{u})\mathbf{x}_c = \big(G_c(\mathbf{u}) - H_{cq}(\mathbf{u})\big)\mathbf{r}. \tag{14}$$

Then $H_{cc}(\mathbf{u})$ is nonsingular, and by the properties of the damped positive-definite Laplacian $H_{cc}(\mathbf{u})^{-1} \geq 0$ entrywise. Moreover,

$$\big(G_c(\mathbf{u}) - H_{cq}(\mathbf{u})\big)\mathbf{r} \geq 0. \tag{15}$$

Thus

$$\mathbf{x}_c = H_{cc}(\mathbf{u})^{-1}\big(G_c(\mathbf{u}) - H_{cq}(\mathbf{u})\big)\mathbf{r} \geq 0. \tag{16}$$

Since also $\mathbf{x}_q = \mathbf{r} \geq 0$, we conclude

$$\mathbf{x}_\infty = \mathbf{P}_q^\top \mathbf{r} + \mathbf{P}_c^\top \mathbf{x}_c \geq 0. \tag{17}$$

□

Proof of Theorem 2 (Semantic reasoning operator)

First, by assumption $\mathbf{P}_q\mathbf{P}_q^\top = I_{N_q}$, so the Moore-Penrose pseudoinverse satisfies $\mathbf{P}_q^\dagger = \mathbf{P}_q^\top(\mathbf{P}_q\mathbf{P}_q^\top)^{-1} = \mathbf{P}_q^\top$. The identities $\mathbf{P}_q^\top\mathbf{P}_q + \mathbf{P}_c^\top\mathbf{P}_c = I_N$ and $\mathbf{P}_c\mathbf{P}_q^\top = \mathbf{0}$ give the orthogonal decomposition $\mathbf{x} = \mathbf{P}_q^\top\mathbf{P}_q\mathbf{x} + \mathbf{P}_c^\top\mathbf{P}_c\mathbf{x}$ with $\mathrm{range}(\mathbf{P}_q^\top) \perp \mathrm{range}(\mathbf{P}_c^\top)$. Hence $\ker(\mathbf{P}_q) = \mathrm{range}(\mathbf{P}_c^\top)$, and the general solution of $\mathbf{P}_q\mathbf{x} = \mathbf{r}$ is

$$\mathbf{x} = \mathbf{P}_q^\top \mathbf{r} + \mathbf{P}_c^\top \mathbf{x}_c. \tag{18}$$

By Theorem 1, the chunk-related component $\mathbf{x}_c$ is not free in equilibrium $\mathbf{x}_\infty$. To see that, it is enough to extract the chunk potential coordinates from (3):

$$\begin{aligned} \mathbf{P}_c\big(H(\mathbf{u})\mathbf{x}_\infty - \mathbf{P}_q^\top \mathbf{y}_\infty\big) &= \mathbf{P}_c H(\mathbf{u})\mathbf{x}_\infty \\ &= \mathbf{P}_c G(\mathbf{u})\mathbf{r}. \end{aligned} \tag{19}$$

Substituting $\mathbf{x}_\infty = \mathbf{P}_q^\top \mathbf{r} + \mathbf{P}_c^\top \mathbf{x}_c$ yields the chunk-side balance

$$H_{cc}(\mathbf{u})\mathbf{x}_c = \big(G_c(\mathbf{u}) - H_{cq}(\mathbf{u})\big)\mathbf{r}, \tag{20}$$

which has the unique solution $\mathbf{x}_c = \mathbf{T}(\mathbf{u})\mathbf{r}$. That establishes equation (4).

The reasoning bound is then immediate from linearity of $\mathbf{R}(\mathbf{u})$ in $\mathbf{r}$. Namely, applying (4) to both $\mathbf{r}$ and $\mathbf{r}^\star$, we have $\mathbf{x}_\infty - \mathbf{x}^\star = \mathbf{R}(\mathbf{u})(\mathbf{r} - \mathbf{r}^\star)$. Since $H(\mathbf{u})$ is symmetric positive definite, by the standard norm properties we have:

$$\begin{aligned} \|\mathbf{x}_\infty - \mathbf{x}^\star\|_{H(\mathbf{u})}^2 \leq \lambda_{\max}\big(\mathbf{R}(\mathbf{u})^\top H(\mathbf{u})\mathbf{R}(\mathbf{u})\big) \\ \cdot \|\mathbf{r} - \mathbf{r}^\star\|_2^2. \end{aligned} \tag{21}$$

Taking square roots yields (5). □

Proof of Theorem 3 (Adaptive convergence)

By Theorem 1, the equilibrium point $(\mathbf{x}_\infty, \mathbf{y}_\infty)$ solves KKT equation (3) for the problem:

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & \tfrac{1}{2}\mathbf{x}^\top H(\mathbf{u})\mathbf{x} - \mathbf{x}^\top G(\mathbf{u})\mathbf{r} \\ \text{subject to} \quad & \mathbf{P}_q\mathbf{x} = \mathbf{r}. \end{aligned} \tag{22}$$

Since $H(\mathbf{u}) = \mathbf{B}^\top\mathbf{U}\mathbf{B} + \alpha I$ and $G(\mathbf{u}) = \mathbf{P}_q^\top\mathbf{P}_q\mathbf{B}^\top\mathbf{U}\mathbf{B}\mathbf{P}_q^\top$ are affine in $\mathbf{u}$, the objective function $\mathbf{x}^\top H(\mathbf{u})\mathbf{x} - \mathbf{x}^\top G(\mathbf{u})\mathbf{r}$ is affine in $\mathbf{u}$ for every feasible $\mathbf{x}$. The value function $V(\mathbf{u}, \mathbf{r}) =$

$\min_{\mathbf{x}}\big[\frac{1}{2}\mathbf{x}^\top H(\mathbf{u})\mathbf{x} - \mathbf{x}^\top G(\mathbf{u})\mathbf{r}\big]$ is therefore concave on $\mathcal{U}$ as the pointwise infimum of affine functions (Rockafellar, 1997). From that, we have:

$$V(\mathbf{u}_{t+1}, \mathbf{r}) \le V(\mathbf{u}_t, \mathbf{r}) + \langle \nabla V(\mathbf{u}_t, \mathbf{r}),\, \Delta\mathbf{u}_t \rangle, \tag{23}$$

where $\Delta\mathbf{u}_t = \mathbf{u}_{t+1} - \mathbf{u}_t$. Continuity on the compact set $\mathcal{U}$ also ensures that $V_\star = \min_{\mathbf{u}\in\mathcal{U}} V(\mathbf{u}, \mathbf{r}) > -\infty$ and that the minimizer set $\mathcal{U}_\star = \{\mathbf{u} \in \mathcal{U} : V(\mathbf{u}, \mathbf{r}) = V_\star\}$ is non-empty.

Consider now the following function:

$$L(\mathbf{u}) = V(\mathbf{u}, \mathbf{r}) - V_\star \tag{24}$$

defined on $\mathcal{U}$. We claim L is a descent function for the conductance update obtained after the inner semantic equilibrium has been reached.

Indeed, it is easy to see $L(\mathbf{u}) \ge 0$ with $L(\mathbf{u}) = 0$ if and only if $\mathbf{u} \in \mathcal{U}_\star$. It remains to show, that $\Delta L(\mathbf{u}_t) = L(\mathbf{u}_{t+1}) - L(\mathbf{u}_t) \le 0$.

At the minimizer, V takes the value

$$V(\mathbf{u}, \mathbf{r}) = \tfrac{1}{2}\mathbf{x}_\infty^\top H(\mathbf{u})\mathbf{x}_\infty - \mathbf{x}_\infty^\top G(\mathbf{u})\mathbf{r}. \tag{25}$$

Differentiating with respect to u_e (writing $\partial_{u_e} \equiv \partial/\partial u_e$) gives

$$\begin{aligned}\partial_{u_e} V(\mathbf{u}, \mathbf{r}) = {} & \tfrac{1}{2}\mathbf{x}_\infty^\top(\partial_{u_e} H(\mathbf{u}))\mathbf{x}_\infty \\ & - \mathbf{x}_\infty^\top(\partial_{u_e} G(\mathbf{u}))\mathbf{r} \\ & + \big[H(\mathbf{u})\mathbf{x}_\infty - G(\mathbf{u})\mathbf{r}\big]^\top \partial_{u_e}\mathbf{x}_\infty.\end{aligned} \tag{26}$$

By Theorem 1, we have $H(\mathbf{u})\mathbf{x}_\infty - G(\mathbf{u})\mathbf{r} = \mathbf{P}_q^\top \mathbf{y}_\infty$. Since $\mathbf{P}_q\mathbf{x}_\infty = \mathbf{r}$, we have

$$\begin{aligned}\big(\mathbf{P}_q^\top \mathbf{y}_\infty\big)^\top \partial_{u_e}\mathbf{x}_\infty &= \mathbf{y}_\infty^\top \partial_{u_e}(\mathbf{P}_q\mathbf{x}_\infty) \\ &= \mathbf{y}_\infty^\top \partial_{u_e}\mathbf{r} = 0.\end{aligned} \tag{27}$$

Substituting $\partial_{u_e} H(\mathbf{u}) = \mathbf{b}_e\mathbf{b}_e^\top$ and $\partial_{u_e} G(\mathbf{u}) = \mathbf{P}_q^\top\mathbf{P}_q\mathbf{b}_e\mathbf{b}_e^\top\mathbf{P}_q^\top$ yields:

$$\partial_{u_e} V(\mathbf{u}, \mathbf{r}) = \tfrac{1}{2}(\mathbf{b}_e^\top\mathbf{x}_\infty)^2 - \big((\mathbf{P}_q\mathbf{b}_e)^\top\mathbf{r}\big)^2. \tag{28}$$

Therefore, the conductance evolves according to the steepest descent rule:

$$\begin{aligned}u_{t+1,e} &= u_{t,e}\,\bar{\phi}_{t,e}(\mathbf{r})\,\bar{\psi}_{t,e}(\mathbf{x}_\infty(\mathbf{u}_t, \mathbf{r})) \\ &= u_{t,e}\exp\big(-\eta\,\partial_{u_e} V(\mathbf{u}_t, \mathbf{r})\big),\end{aligned} \tag{29}$$

until it hits the lower or upper bound in the clipping rule (6).

Write $g_e = \partial_{u_e} V(\mathbf{u}_t, \mathbf{r})$ and consider three cases. *Case (i):* $u_{\min} < u_{t+1,e} < u_{\max}$. Then $u_{t+1,e} - u_{t,e} = u_{t,e}\big(\exp(-\eta g_e) - 1\big)$, and since $u_{t,e} > 0$ and $z\big(\exp(-\eta z) - 1\big) \le 0$ for all $z \in \mathbb{R}$ and $\eta > 0$, we have $g_e(u_{t+1,e} - u_{t,e}) \le 0$. *Case (ii):* the clip activates at $u_{\min}$, i.e. $u_{t,e}\exp(-\eta g_e) \le u_{\min}$ and $u_{t+1,e} = u_{\min}$. Dividing the first inequality by $u_{t,e} > 0$ gives $\exp(-\eta g_e) \le u_{\min}/u_{t,e} \le 1$ (the second inequality follows from $u_{\min} \le u_{t,e}$), hence $g_e \ge 0$. Also $u_{t+1,e} - u_{t,e} = u_{\min} - u_{t,e} \le 0$. So $g_e(u_{t+1,e} - u_{t,e}) \le 0$. *Case (iii):* the clip activates at $u_{\max}$. By the symmetric argument, $g_e \le 0$ and $u_{t+1,e} - u_{t,e} \ge 0$, so $g_e(u_{t+1,e} - u_{t,e}) \le 0$.

That shows:

$$\langle \nabla V(\mathbf{u}_t, \mathbf{r}),\, \Delta\mathbf{u}_t \rangle = \sum_{e=1}^{M} g_e \cdot (u_{t+1,e} - u_{t,e}) \le 0 \tag{30}$$

for $\eta > 0$. Therefore, by concavity of V:

$$\begin{aligned}\Delta L(\mathbf{u}_t) &= V(\mathbf{u}_{t+1}, \mathbf{r}) - V(\mathbf{u}_t, \mathbf{r}) \\ &\le \langle \nabla V(\mathbf{u}_t, \mathbf{r}),\, \Delta\mathbf{u}_t \rangle \le 0\end{aligned} \tag{31}$$

and L is a descent function evaluated at $\mathbf{x}_\infty = \mathbf{R}(\mathbf{u})\mathbf{r}$.

We conclude that the sequence $\{L(\mathbf{u}_t)\}_{t=0}^\infty$ is monotone non-increasing and bounded below by zero, hence converges to some $L_\infty \ge 0$. As a result, the energy sequence $\{V(\mathbf{u}_t, \mathbf{r})\}_{t=0}^\infty$ converges to $V_\infty = V_\star + L_\infty \in [V_\star, V(\mathbf{u}_0, \mathbf{r})]$, and decreases strictly unless $\mathbf{u}_{t+1} = \mathbf{u}_t$.

Finally, we identify the limit set of the trajectory by the discrete-time LaSalle invariance principle (see, e.g., (Khalil, 2002; Luenberger, 1979)). The function L is a continuous descent function on the compact set $\mathcal{U}$ for the iteration $\mathbf{u}_{t+1} = \mathbf{S}_\eta(\mathbf{u}_t)$, with $\Delta L(\mathbf{u}_t) \le 0$ for every t. By the strict-descent property established above, $\Delta L(\mathbf{u}) = 0$ if and only if $\mathbf{S}_\eta(\mathbf{u}) = \mathbf{u}$, i.e., if and only if $\mathbf{u} \in \mathcal{S}_\star$. The set $\mathcal{S}_\star$ is invariant under $\mathbf{S}_\eta$ by construction (fixed points are mapped to themselves), so it is the largest invariant subset of $\{\mathbf{u} \in \mathcal{U} : \Delta L(\mathbf{u}) = 0\}$. The discrete LaSalle invariance principle therefore yields $\lim_{t\to\infty} \mathrm{dist}(\mathbf{u}_t, \mathcal{S}_\star) = 0$, completing the proof. □

B Pipeline Pseudocode

The full per-query controlled-hallucination pipeline (Algorithm 1). `PLANNER`, `EXTRACTOR`, `GENERATOR` are LLM-mediated subsystems (or agents); `SERVO` and `RERANKER` are deterministic, LLM-free. A *hop* is a typed pair (`sub_q` : *string*, `type` : *EntityType*) emitted

by the planner. The verifier's typed-mismatch outcome is the system's prediction-error signal; an unrecoverable rejection triggers *abstention* before generation.

The pseudocode makes the controlled-hallucination role assignment of Appendix K concrete: `PLANNER.plan` emits the top-down hypothesis (generative prior), `SERVO.retrieve` computes the bottom-up equilibrium semantic potential (sensory likelihood), and `EXTRACTOR.verify` performs the local prediction-error check. Rejection cascades close the outer loop or trigger abstention.

C Full Per-Cell Tables

Servo K-isolation. Removing the servo collapses retrieval. We ablate it and report $K = F_1^{\text{full}}/F_1^{\text{no-servo}}$, computed paired ($n = 200$ per cell) on MuSiQue-200. Across the qwen tier the lift is essentially flat with reader size (6.9× at 0.8B, 6.2× at 4B, 6.0× at 9B; Table 3), pointing to a structural rather than reader-dependent contribution from the servo. The four-hop bucket amplifies more strongly (5× to 27×), but the denominator sits on a near-zero floor in one cell; the absolute raw-F_1 lift is the more defensible quantity.

L3 ablation and the 4B single-lever matrix. The L3-only embedding ablation (answer-role questions) raises Ext-F_1 at qwen3.5:4B/9B (mean $\Delta = +0.038$) (Weller et al., 2026) but triples per-query tokens, so L1+L2 stays canonical. The 4B single-lever ablation in Table 4 runs six controlled single-lever variants of the canonical RAE configuration on MuSiQue-200 at `qwen3.5:4B`; the Ext-F_1 plateau at ∼0.33 holds and no single lever breaks past it. The decomposition of production headroom — retrieval ∼86%, prompts ∼14% (Cheng et al., 2025) (App. G) — and the cross-encoder reranker as the only positive Ext-F_1 delta inside the noise floor are detailed in the table.

D Theorem 2 Empirical Verification

Protocol. On a 50-query MuSiQue-200 subset, we compute the tightness ratio

$$\rho(q) = \frac{\|\mathbf{x}_\infty - \mathbf{x}^\star\|_{H(\mathbf{u})}}{\gamma(\mathbf{u})\,\|\mathbf{r} - \mathbf{r}^\star\|_2} \tag{32}$$

per query, where $\gamma(\mathbf{u})$ is computed on $R(\mathbf{u})^\top H(\mathbf{u}) R(\mathbf{u})$. The strict claim of Theorem 2 is $\rho(q) \le 1$ for every q and every reference pair $(\mathbf{r}, \mathbf{r}^\star)$. We probe three reference choices for $\mathbf{r}^\star$: *(i)* trivial ($\mathbf{r}^\star = \mathbf{0}$), *(ii) oracle* (the gold-answer-string-matching ideas from the seeded query plan), and *(iii) uniform* ($\mathbf{r}^\star = N_q^{-1/2}\mathbf{1}$). Each query has $N_q \in [3, 11]$ seeded ideas (median $N_q = 7$).

Result. The bound holds in $\mathbf{50/50}$ queries under all three reference choices. Table 7 reports the distribution of $\rho(q)$ and of $\gamma(\mathbf{u})$.

Primary empirical claim. Across $n = 50$ paired $(\mathbf{r}, \mathbf{r}^\star)$ instances, the Theorem 2 Lipschitz bound holds in every case; the median tightness ratio is $\rho = 0.619$ with $\max \rho = 0.988$, i.e. $\gamma(\mathbf{u})$ overestimates the $H(\mathbf{u})$-norm of $R(\mathbf{u})(\mathbf{r} - \mathbf{r}^\star)$ by a factor of $\sim 1.6\times$ on the typical query and is essentially tight (within 1.2%) on the worst case. Per-query $\gamma(\mathbf{u})$ ranges over 0.63–5.53 on the multi-hop MuSiQue subset, with median $\gamma = 1.67$.

The two verification experiments. Both panels of Figure 2 instantiate the same Lipschitz inequality with different displacements. In panel (a) the displacement is $\Delta\mathbf{r} = \mathbf{r} - \mathbf{r}^\star$ (the gap between parsed and gold reference, one point per query); the bound holds in $50/50$ instances, with median tightness $\rho = 0.619$ and worst-case saturation within 1.2% of the ceiling. The dashed line is the median empirical slope $\hat{\rho} = 0.619$, and the worst-case point (red ring; $\gamma = 0.628$, $\rho = 0.988$) sits on the ceiling. On this benchmark $\mathbf{r}^\star = \mathbf{0}$ (gold answer entities are absent from first-hop seeds; Table 7 footnote), so the vertical axis reduces to $\|\mathbf{x}_\infty\|_{H(\mathbf{u})}$, the H-energy of the reached equilibrium itself. In panel (b) the displacement is $\Delta\mathbf{r} = \mathbf{r}_{\text{noisy}} - \mathbf{r}_{\text{clean}}$ and the bound is exercised directly under perturbation: it holds with 0 violations and $\max \rho_{\text{emp}} = 1.000$ — operator-norm tightness in the worst direction. Per noise model: drop ($n = 735$, median 0.651, max 1.000), weight-shock ($n = 750$, median 0.617, max 1.000), Gaussian ($n = 1000$, median 0.656, max 1.000).

Reading the scatter. The figure exposes a clean structural pattern that is not visible from the bound alone. Points cluster along a curved envelope whose color gradient runs from low-γ (deep blue) near the origin to high-γ (yellow) at the right edge. Two regimes are geometrically distinct.

Blue cluster near the origin (low γ, small RHS). These queries sit in narrow neighborhoods of the bipartite idea–chunk graph: few high-conductance pathways exit the query-idea seeds, so the equilib-

Algorithm 1 RAE per-query controlled-hallucination pipeline.

```
/* PLANNER, EXTRACTOR, GENERATOR : LLM-mediated subsystems or agents.    */
/* SERVO, RERANKER               : deterministic subsystems, LLM-free.   */
/* A hop = (sub_q : string, type : EntityType).                          */
answer(question) {
    chain    = PLANNER.plan(question);
        /* decompose question into K typed sub-questions; 1 LLM call    */
    evidence = [];
    for k = 1, ..., K {
        ev  = SERVO.retrieve(chain[k]);
            /* compute equilibrium x_inf = R(u)*r; return top chunks    */
        ev  = RERANKER.filter(ev);
            /* cross-encoder rerank to top output_K; optional           */
        ent = EXTRACTOR.read(chain[k], ev);
            /* LLM reads ev and returns entity of type chain[k].type;
               1 LLM call                                                */
        ent = EXTRACTOR.verify(ent, chain[k].type);
            /* check type, rebind on correctable mismatch, reject
               otherwise; 1 LLM call                                     */
        evidence += ev;
        if (k < K)
            chain[k+1].sub_q =
                PLANNER.bind(chain[k+1].sub_q, ent);
            /* substitute ent into the text of next hop's sub-question;
               up to 1 LLM call (skipped if ent is already a substring
               of chain[k+1].sub_q)                                      */
    }
    return GENERATOR.write(question, evidence);
        /* generate final answer grounded on evidence; 1 LLM call,
           plus an optional compression call at evaluation time         */
}
```

rium operator $\mathbf{R}(\mathbf{u})$ has only one available transport direction, and the cognitive gap $(\mathbf{r}-\mathbf{r}^\star)$ aligns with it by default. The points lie on the $y = x$ diagonal ($\rho \to 0.99$): the displacement between reached and gold equilibria saturates the topology's full capacity. The worst-case red-ringed query ($\gamma = 0.628$, $\rho = 0.988$) is "worst-case" only in the bound-margin sense — it is the most *constrained* query in the set, not the hardest.

Yellow points on the right (high γ, large RHS). These queries sit in richly cross-linked neighborhoods where many high-conductance directions are available. The specific cognitive gap $(\mathbf{r}-\mathbf{r}^\star)$ excites only one or two of them; ρ drops to 0.40–0.58. The vertical distance from the diagonal is the unused expressive capacity — the H-energy the graph held in reserve that this particular question never called on. Spearman $\rho(\gamma, \rho) = -0.85$ ($p \approx 0$) on the 50-query sample: the questions the topology could amplify most are the questions whose gap signal least exploits that amplification — a structural property of the corpus rather than of the reader.

Lipschitz robustness under query noise. We exercise the Theorem 2 bound directly by perturbing the parsed reference $\mathbf{r}_{\text{clean}}$ with three q_{idx}-preserving noise models — *drop* (zero out k of the N_q query-idea weights), *weight-shock* (replace k weights with $\mathcal{U}(0, \max(\mathbf{r}_q))$), and *Gaussian* ($\mathbf{r}_q \leftarrow \mathbf{r}_q + \sigma\varepsilon$, clipped to $[0, 1]$). On MuSiQue-50 with $n = 2{,}485$ perturbations across the same 50 queries, the bound holds on every perturbation (zero violations), with empirical utilization $\rho_{\text{emp}} = \|R\Delta\mathbf{r}\|_{H(\mathbf{u})}/(\gamma(\mathbf{u})\,\|\Delta\mathbf{r}\|_2)$ at median 0.648 and max 1.000. The maximum certifies worst-case tightness — at least one perturbation aligns with $R(\mathbf{u})$'s top eigenmode. Two structural patterns hold across the run. Low-γ queries saturate the bound under perturbation (median $\rho_{\text{emp}} > 0.85$ in the bottom γ-tertile); high-γ queries leave headroom. And downstream chunk-rank stability tracks the same structure: median top-10 chunk Jaccard $= 0.818$ across all perturbations. The summary statement: ρ_{emp} is a per-query stability number, computable in one sparse linear solve, that upper-bounds how far a parser drift will move the retrieval state.

E Theorem 3 Empirical Verification

Setup. Theorem 3 predicts that the outer-loop conductance update $\mathbf{u}_{t+1} = \text{clip}_{\mathcal{U}}\big(\mathbf{u}_t \odot \exp\big(-\eta\,\nabla_{\mathbf{u}} V(\mathbf{u}_t, \mathbf{r})\big)\big)$, with $V(\mathbf{u}, \mathbf{r}) = \frac{1}{2}\,\mathbf{x}_\infty^\top H(\mathbf{u})\,\mathbf{x}_\infty - \mathbf{x}_\infty^\top G(\mathbf{u})\,\mathbf{r}$ and

Method	MuSiQue-200 Ext-F_1 0.8B	4B	9B	HotpotQA-500 Ext-F_1 0.8B	4B	9B
LLM-only ($k=1$)	0.046	0.073	0.103	0.108	0.199	0.241
Vector-RAG ($k=10$)	0.126	**0.368**	0.343	0.311	0.648	0.644
Vector-RAG + Reranker ($k=10$)	0.133	0.343	0.355	0.335	**0.676**	**0.695**
RAPTOR	0.128	0.340	0.350	0.334	0.531	0.572
HippoRAG 2	**0.257**	**0.436**	**0.456**	**0.506**	**0.700**	0.690
RAE-cfg2 [L1]	0.179	0.301	0.377	0.364	0.517	0.684
RAE-cfg2 [L1+L2]	0.191	0.318	0.376	0.367	0.534	0.689
RAE-cfg2 [L1+L3]	0.178	0.306	0.377	0.365	0.512	0.688

Table 2: Full per-cell Ext-F_1 at the qwen tier on both benchmarks, drawn from `RAE-vLLM-total.csv` under the matched-ingest vLLM protocol. Bold marks the per-column maximum across retrieval methods. HippoRAG 2 holds the absolute quality lead in five of six cells at ~ 4.5–$5\times$ RAE's per-query token cost (Section 5.1); RAE-cfg2 wins composite intelligence density in all six cells against this baseline (Table 1). RAE-cfg2 is reported at three embedding configurations; the canonical L1+L2 is pinned uniformly in the body figures.

Reader	$K(F_1)$ overall	K(Ext-F_1) overall	$K(F_1)$ at 3-hop	$K(F_1)$ at 4-hop
qwen3.5:0.8B	$6.89\times$	$2.28\times$	$3.53\times$	$5.17\times$
qwen3.5:4B	$6.18\times$	$2.75\times$	$5.21\times$	vanilla floor
qwen3.5:9B	$5.98\times$	$3.04\times$	$3.40\times$	$26.99\times$

Table 3: Cross-reader K-isolation across the qwen tier on MuSiQue-200. The servo lift is essentially flat with reader size at the overall level ($6.89 \to 6.18 \to 5.98$). The 4-hop bucket amplifies more strongly because the no-servo denominator falls to a near-zero floor; at qwen-4B the no-servo 4-hop F_1 is exactly zero in this run (one near-empty cell) and we mark the entry as a vanilla floor rather than reporting a divide-by-zero ratio.

$\mathbf{x}_\infty = R(\mathbf{u})\,\mathbf{r}$, produces a sequence $\{V_t\}$ that is non-increasing along orbits and whose increment $\Delta V_t := V_{t+1} - V_t$ vanishes in the limit on the stationary set $\mathcal{S}_\star$ (discrete LaSalle invariance (LaSalle, 1960)). The descent is structural. The multiplicative-clipped update guarantees $\Delta V_t \le 0$ along every orbit by construction, on every input query the pipeline serves, not on a sampled subset. The experiment below displays this descent on representative trajectories and confirms that the implementation matches the math. We illustrate on 5 MuSiQue queries with retrieval-induced initial conductances $\mathbf{u}_0$, step size $\eta = 0.05$, and horizon $T = 1000$. The same descent runs on every benchmark query. For each iteration we solve the Schur complement $H_{cc}(\mathbf{u}_t)^{-1}$ once and reuse it for both the Lyapunov value V_t and the closed-form envelope-theorem gradient $\nabla_i V = \frac{1}{2}(B\mathbf{x}_\infty)_i^2 - (BP_q^\top \mathbf{r}_q)_i^2$, which we cross-check against finite differences at relative error $\le 1.1 \times 10^{-9}$.

Results. Figure 3 reports raw, unnormalized trajectories on the active window $t \in [0, 250]$ (the descent saturates well before T). Panel (a) plots the Lyapunov function V_t directly: every trajectory descends monotonically from $V_0 \in [-7.996, -0.958]$ down toward $V_{1000} \in [-528, -49]$, with median cumulative drop $\mathrm{med}(V_0 - V_T)/|V_0| \approx 50.2$ and per-query range $[39.8, 65.0]$. Panel (b) plots the per-step increment ΔV_t: across the entire run — $5 \times 1000 = 5000$ step transitions — every observed ΔV_t lies strictly below zero, and the worst (closest-to-zero) value is $\Delta V_t = -3.48 \times 10^{-4}$. Both the weak Lyapunov condition $\Delta V_t \le 0$ required by Theorem 3 *and* the strict form $\Delta V_t < 0$ of textbook asymptotic stability hold on every trajectory we display.

Reading and caveats. The strict inequality $\Delta V_t < 0$ holds *along the observed orbits* $\{\mathbf{u}_t\}_{t=0}^{T}$, not on the entire ϵ-neighborhood $\mathbb{B}(\mathbf{u}_\star, \epsilon) \setminus \{\mathbf{u}_\star\}$. This is the trajectory-wise statement Theorem 3 licenses; it does not promote the equilibrium set $\mathcal{S}_\star$ to a single asymptotically stable equilibrium point. As noted in the Remark following the theorem, $\mathcal{S}_\star$ generically lies on the boundary $\partial\mathcal{U}$, where the multiplicative clip freezes the coordinates that have hit the box constraint; for the runs reported here, the boundary-saturated fraction stabilizes at 0.1–0.5% per query at $t = T$, consistent with LaSalle convergence to a boundary face rather than to an interior critical point. The Lyapunov decay in panel (a) and the strict-descent distribution in panel (b) jointly verify the empirical content of Theorem 3 at the scale of the operator we deploy in §3.

Cell	Lever changed	Ext-F_1	raw F_1	$\Delta_{\text{Ext-}F_1}$ vs. A0	Verdict
A0	baseline (MULTIHOP + compress)	**0.327**	0.310	—	canonical
AE	type-guided prompt (variant E)	0.320	0.316	−0.007	null
A1	disable compress (compress=0)	0.543	0.095	+0.215	gold-leak, not real
A2	compress model → qwen3.5:4b	0.318	0.298	−0.009	null
A3	wider slate (TOP_K = 50)	0.302	—	−0.025	null/negative
A4	cross-encoder reranker (output_k= 25)	0.340	0.317	+0.013	marginal
AF	HippoRAG *Thought:/Answer:* prompt	0.277	0.266	**−0.050**	negative

Table 4: Single-lever ablation matrix at `qwen3.5:4B`. A1 (drop compress) inflates Ext-F_1 via gold-substring leakage — raw F_1 collapses to 0.095. A4 (cross-encoder reranker) is the only positive cell within the noise floor. The HippoRAG-prompt cell (AF) is dissected in Appendix G.

Method	MuSiQue-200		HotpotQA-500	
	Tokens/q (mean)	Index tokens	Tokens/q (mean)	Index tokens
LLM-only	63	0	64	0
Vector-RAG	1397	0	1412	0
Vector-RAG + Reranker	1487	0	1433	0
Raptor	2185	1,074k	2196	1,787k
HippoRAG 2	5214	6,543k	4904	10,006k
RT-RAG	10.5 M	0	318k	0
RAE L1	1016	4,474k	1034	7,285k
RAE L1+L2 (canonical)	**1016**	4,474k	**1034**	7,285k

Table 5: Per-query and per-index token cost at the `qwen3.5:0.8B` reader, drawn from `RAE-vLLM-total.csv`. Index tokens are the full ingest cost (prompt + completion). Bold marks the minimum tokens/q among retrieval methods (LLM-only aside). At 0.8B RAE (canonical L1+L2 at this reader; Appendix C) uses ∼1,020 tokens per query — below Vector-RAG's 1,397, ∼5× below HippoRAG's 5,214, and ∼310× below RT-RAG on HotpotQA (318 k tokens/q) and ∼10,000× below RT-RAG on MuSiQue (10.5 M tokens/q at this reader; the RT-RAG tree expands further on hard MuSiQue queries when the inner reader fails to converge). Per-cell RT-RAG MuSiQue measurements at 4B and 9B are 0.9 M and 3.6 M tokens/q (Appendix J). Canonical RAE L1+L2 per-query tokens across qwen-tier cells are about 1,020 tokens/q (essentially reader-independent). Baselines stay near their 0.8B token counts across readers (VRAG ∼ 1.4 k, VRAGrr ∼ 1.5 k, HippoRAG ∼ 5 k). Index tokens (prompt + completion) are amortized across queries; RAE L1+L2 indexing cost (4,474 k on MuSiQue, 7,285 k on HotpotQA) is comparable to HippoRAG 2's (6,543 k / 10,006 k). The L3-only ablation (Appendix C) raises per-query tokens to ∼ 1,500–1,700 at qwen3.5:4B/9B.

F Failure-Mode Analysis

Failure modes by reader-benchmark cell, drawn from the per-query JSON results in the anonymized supplementary materials:

- **Small-reader (0.8B), MuSiQue.** Dominant failure mode: *compose-and-extract truncation*. The 0.8B reader can extract the right entity from a single hop's chunks but loses the chain across hops; RAE partially compensates by binding the verified entity into the next hop's sub-question text.

- **Small-reader (0.8B), HotpotQA.** Bridge questions: RAE is competitive with HippoRAG and matches RT-RAG. Comparison questions: RAE underperforms by ∼0.05 Ext-F_1, traceable to the fact that comparison questions exercise generation more than retrieval.

- **Mid-tier (4B), both benchmarks.** Dominant failure mode: *bridge-entity confusion inside the reader*. A controlled 4B campaign on MuSiQue (six variants over prompt template, compression head, slate breadth, and reranker depth) reaches a stable Ext-F_1 plateau at ∼0.33; no single $\mathcal{S}$-side knob breaks past it. Per-hop diagnostics show the gold entity is in the retrieval slate in 44–72% of failing queries: the reader sees the right candidate but selects the wrong bridge entity. This localizes the residual gap to HippoRAG (0.436, $\Delta = 0.109$) to reader-side disambiguation, not retrieval. Adaptive top-k by query difficulty marginally helps (one variant, +0.013 Ext-F_1); chain-decomposed prompting and per-hop sub-query retrieval are the leading reader-side levers for future work.

Method	MuSiQue-200 Raw F_1 0.8B	4B	9B	HotpotQA-500 Raw F_1 0.8B	4B	9B
LLM-only	0.035	0.042	0.056	0.067	0.084	0.099
Vector-RAG	0.074	0.106	0.103	0.133	0.216	0.186
Vector-RAG + Reranker	0.072	0.130	0.119	0.146	0.226	0.202
Raptor	0.042	0.077	0.091	0.090	0.119	0.130
HippoRAG 2	**0.187**	0.348	0.377	**0.444**	0.628	0.643
RT-RAG	0.088	0.140	**0.634**	—	—	—
RAE L1	0.041	0.223	0.244	0.072	0.428	—
RAE L1+L2	0.040	0.310	0.228	0.074	0.422	—
RAE L1+L2+L3	0.041	0.310	0.197	0.070	0.415	—

Table 6: Full per-cell *raw* F_1, reported alongside the Ext-F_1 values of Table 2. Raw F_1 here is the SQuAD token-F_1 applied to the *verbatim* reader output — no answer-line or sentence-span extraction — and is the literal-output quantity that underwrites the inflation contrast in Section 4 (Metric and protocol). It is computed from the per-query JSON dumps in the supplementary archive; it differs by design from the lightly-extracted `f1` column in the released `RAE-vLLM-total.csv`, which collapses to the literal-output number only for terse-output methods (HippoRAG 2, RAPTOR). With the canonical RAE L1+L2 pinned uniformly, frontier membership at the low-cost edge is preserved in $4/6$ cells under raw F_1 (the same count as under Ext-F_1, with off-frontier cells at the two small-reader cells MuSiQue-0.8B and HotpotQA-0.8B). Absolute rankings at the small-reader cells change between kernels: on raw F_1 at `0.8B` MuSiQue, HippoRAG (0.187) is ahead of RAE L1+L2 (0.040), and the same cell on Ext-F_1 inverts to RAE 0.275 vs HippoRAG 0.257. This is a verbosity-of-reader-output effect (the RAE generator at `0.8B` emits verbose-but-gold-bearing answers; HippoRAG emits terse-and-direct answers): the Ext-F_1/raw-F_1 gap is large at `qwen3.5:0.8B` (RAE L1+L2 MuSiQue 0.275 vs 0.040, ratio $6.9\times$), small at HippoRAG at the same cell (0.257 vs 0.187, ratio $1.4\times$), and near-zero at 9B. Bold marks the per-column maximum on raw F_1.

- **Large-tier (9B), MuSiQue.** HippoRAG's PPR over the LLM-constructed entity graph picks up cross-document compositional links that RAE's bipartite chunk–idea structure does not encode. Adding a learned per-edge conductance prior is future work.

G HippoRAG Prompt: Oracle Win, Production Loss

We document the most informative negative cell of the 4B ablation matrix (Table 4, AF). The cell tests the HippoRAG 2 chain-of-thought prompt — the *Thought:/Answer:* scaffold with a one-shot 2-hop example, copied verbatim from (Gutiérrez et al., 2025) — as a drop-in replacement for the canonical MULTIHOP prompt inside the RAE pipeline. A gold-context oracle predicted a win; the production cell regressed by -0.050 Ext-F_1. The mechanism behind this oracle–production gap clarifies the retrieval-versus-prompting allocation in Section 6.

Oracle upper-bound prediction. The reader-upper-bound sweep evaluates six prompt templates on *gold MuSiQue paragraphs only*, with no retriever in the loop (Table 8). On gold context the HippoRAG scaffold beats MULTIHOP at 4B and 9B by $+0.057$ and $+0.028$ raw F_1, but loses by -0.109 at 0.8B, where the small reader cannot stay coherent across the $\sim$1900-token Thought trace. The asymmetric prescription read clearly: deploy HippoRAG at $\geq$ 4B, avoid at 0.8B.

Production result. Identical configuration as canonical A0 (compress=1, presearch=0, verify=1, entity-complete=1, chunk-seed=1, reranker=0, L1+L2+L3) with `RAE_PROMPT_VARIANT=F` swapping in the HippoRAG scaffold. Matched 200/200 queries against A0 (Table 9). All three metrics regress; mean answer length grows 80% as the Thought trace leaks through the compress step. The per-hop breakdown is in Table 10.

The inversion is per-hop monotone: the deeper the chain — exactly the regime chain-of-thought was designed to help — the worse the regression. This is the diagnostic signature of the failure mode.

Mechanism. The oracle table holds one variable fixed that the production system does not — the retrieved context. With gold passages only, every fact the Thought trace can cite is correct, and the scaffold's job reduces to composing facts that are all true. With real RAE retrieval, the slate contains the bridge entity's chunk (96% hit rate on the failing-query diagnostic), the answer-bearing chunk only some of the time (72% on 2-hop zero-F_1 queries, 51% on 3-hop, 44% on 4-hop), and topically related distractors. The Thought scaffold has no mechanism to suppress a distractor once it has

$\mathbf{r}^\star$ choice	min ρ	median ρ	mean ρ	max ρ	$\rho \leq 1$
trivial ($\mathbf{r}^\star = \mathbf{0}$)	0.390	0.619	0.634	0.988	50/50
oracle (gold-anchored)[†]	0.390	0.619	0.634	0.988	50/50
uniform ($1/\sqrt{N_q}$)	0.371	0.625	0.630	0.991	50/50

Table 7: Tightness ratio $\rho(q) = \|\mathbf{x}_\infty - \mathbf{x}^\star\|_{H(\mathbf{u})}/(\gamma(\mathbf{u})\|\mathbf{r} - \mathbf{r}^\star\|_2)$ over $n = 50$ MuSiQue-200 queries. The Theorem 2 bound $\rho \leq 1$ holds in all 50/50 instances under all three reference choices; the worst-case query saturates the bound to within 1.2%, i.e. the bound is essentially tight on the hardest query. Per-query $\gamma(\mathbf{u})$ ranges over $[0.628, 5.533]$ (median 1.669).
[†]*Oracle collapse on MuSiQue.* On this benchmark, 0/50 first-hop-seeded queries have any idea whose canonical name literally matches the gold answer string. The oracle $\mathbf{r}^\star$ vector is therefore identically zero across all 50 queries and the column is mathematically equal to the trivial $\mathbf{r}^\star = \mathbf{0}$ column. This is a property of MuSiQue's composition-by-construction (gold answers are typically intermediate-hop entities rather than directly-seeded ideas), not a verification artifact. The uniform and trivial columns differ in the expected way (median $|\Delta\rho| = 0.0086$, max $|\Delta\rho| = 0.0829$).

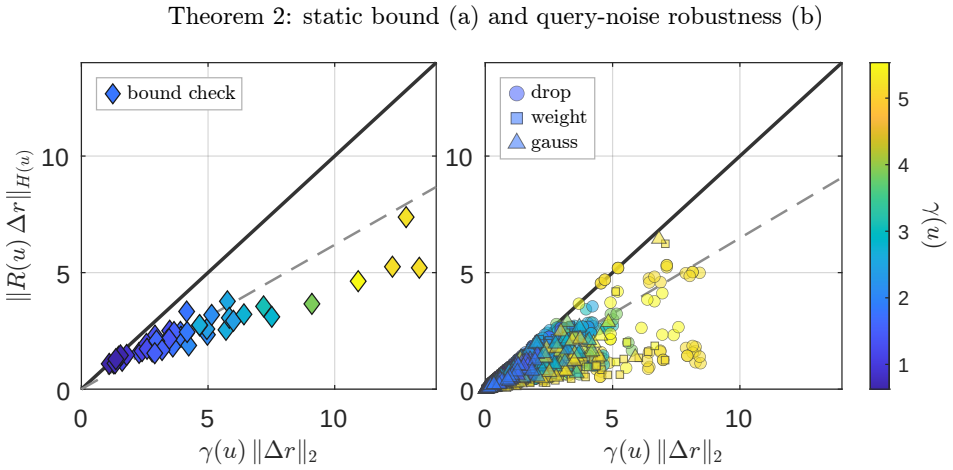

Figure 2: **Theorem 2 verification on MuSiQue-50.** (a) Static bound check on $n = 50$ queries. (b) Query-noise robustness on $n = 2{,}485$ paired perturbations (drop, weight-shock, Gaussian noise models). Both panels share Lipschitz-cone axes: horizontal $\gamma(\mathbf{u})\,\|\Delta\mathbf{r}\|_2$ (the bound, RHS), vertical $\|\mathbf{R}(\mathbf{u})\,\Delta\mathbf{r}\|_{H(\mathbf{u})}$ (LHS). Solid line: ceiling $y = x$. Marker color encodes $\gamma(\mathbf{u})$.

been verbalized into the reasoning trace; the reader mirrors the one-shot example, writes "Thought: X is Y; Y is Z," where Y comes from a distractor, and the *Answer:* line crystallizes Z. The default MULTIHOP prompt, by contrast, lets the reader respond with a vague-but-partially-correct bridge entity when the slate is ambiguous — less catastrophic than a Thought-anchored wrong chain with zero overlap with gold.

Implication — retrieval dominates prompting at the 4B scale. The oracle and production rows give a clean decomposition. Holding the prompt fixed at MULTIHOP and swapping real retrieval for gold paragraphs raises raw F_1 from 0.310 to 0.654, a retrieval gap of $\mathbf{0.344}$. Holding retrieval fixed at gold and swapping MULTIHOP for the best prompt raises raw F_1 from 0.654 to 0.711, a prompt gap of $\mathbf{0.057}$. Of the 0.401 total achievable headroom, retrieval quality accounts for ∼86% and prompt engineering for at most ∼14%. The bound on prompt engineering is generous because the HippoRAG prompt's actual production effect at 4B is −0.050, not +0.057. Prompt engineering is a second-order lever at this reader scale, and chain-of-thought sophistication can actively hurt when the slate is imperfect.

H Pre-scoring Extraction Step

The scoring kernel is the MuSiQue/SQuAD-lineage answer-F_1 (Trivedi et al., 2022): token-level F_1 between predicted and gold answer strings after the four-step normalization (lowercasing, article removal via regex `\b(a|an|the)\b`, punctuation stripping, whitespace fix), with the maximum taken over gold aliases. "Ext-F_1" is our shorthand for the same kernel applied to a determinis-

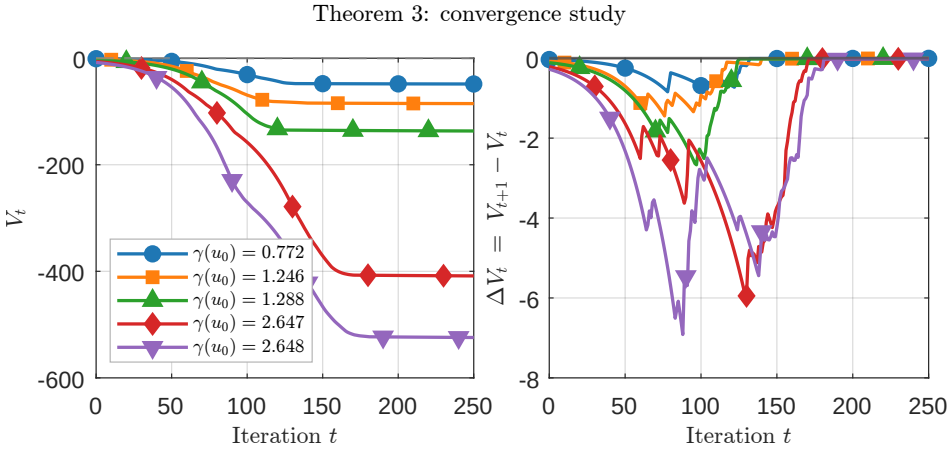

Figure 3: **Theorem 3, empirical verification on MuSiQue (5 queries, $T = 1000$, $\eta = 0.05$).** Panel (a): the Lyapunov function V_t descends monotonically from $V_0 \in [-8.0, -1.0]$ to $V_T \in [-528, -49]$ (median cumulative drop $\approx 50.2\,|V_0|$). Panel (b): the per-step increment $\Delta V_t = V_{t+1} - V_t$ sits below the reference line $\Delta V_t = 0$ at every one of the 5000 step transitions; the closest-to-zero value across the entire run is $\Delta V_t = -3.48 \times 10^{-4}$, so strict descent $\Delta V_t < 0$ (asymptotic-stability form) holds on every observed trajectory. Queries are colored by their initial Lipschitz constant $\gamma(\mathbf{u}_0)$ from Theorem 2.

Reader	MULTIHOP	MULTIHOP-small	HippoRAG	next-best of three lemma prompts
qwen3.5:0.8B	**0.337**	0.285	0.228	≤ 0.064
qwen3.5:4B	0.654	0.617	**0.711**	≤ 0.117
qwen3.5:9B	0.693	0.679	**0.721**	≤ 0.139

Table 8: Reader-upper-bound sweep on *gold MuSiQue paragraphs only* (no retriever in the loop). Raw F_1 by reader and prompt template. Bold marks the per-row maximum.

tically extracted short-answer span; the extraction step runs *before* normalization and is identical for every system in the comparison. It is a deterministic three-step heuristic on the reader's raw output p given the gold-answer alias list $\mathcal{G}$:

1. **Direct gold-substring match.** If any alias in $\mathcal{G}$ is a case-insensitive substring of p after light whitespace normalization, return that alias. This step rewards readers that emit the gold span verbatim regardless of surrounding chatter.

2. **Yes/no template.** If the question is binary (HotpotQA comparison subset), and p begins with or contains an explicit *Yes* or *No* token in the first sentence, return that token. This mirrors the HotpotQA evaluation script's yes/no/noanswer guard.

3. **Lead-non-hedge sentence.** Otherwise, return the first sentence of p stripped of leading hedges (*The answer is*, *Possibly*, *It seems*, *Based on the context*, etc.) and trailing rationale clauses (*because*, *since*, *which is why*, etc.). This step caps the precision penalty for chatty wrappers without rewarding hallucinated continuations.

The heuristic is purely textual and uses no model. It is applied uniformly to all systems — LLM-only, Vector-RAG, Vector-RAG + Reranker, RAPTOR, HippoRAG 2, RT-RAG, and the three RAE variants. The heuristic can only *narrow* the predicted span, so Ext-$F_1 \geq$ raw F_1 pointwise; raw F_1 (Appendix I, Table 6) is the conservative reproducibility check. We considered LLM-judge scoring (surveyed by Kamoi et al., 2024 (Kamoi et al., 2024)) as a third metric but rejected it for the primary tables because reproducibility is harder under judge-prompt drift and our small-reader-amplification margin (+0.018 absolute) is resolvable by either F1 variant.

Important caveat: cross-system inflation is asymmetric. The Ext-F_1/raw-F_1 inflation ratio varies materially across systems. At `qwen3.5:0.8B` on MuSiQue, RAE inflates $6.9\times$ from raw $F_1 = 0.040$ to Ext-$F_1 = 0.275$; HippoRAG 2 inflates only $1.4\times$ from 0.187 to 0.257. The two systems sit at opposite ends of the same reader-verbosity spectrum: RAE outputs

Cell	Ext-F_1	EM	raw F_1	answer chars (mean)
A0 (MULTIHOP)	**0.327**	**0.245**	**0.310**	16
AF (HippoRAG prompt)	0.277	0.210	0.266	29
Δ	-0.050	-0.035	-0.044	$+13$

Table 9: Production result. Matched 200/200 queries against canonical A0 (MULTIHOP prompt) versus AF (HippoRAG *Thought:/Answer:* scaffold). All three quality metrics regress; mean answer length grows 80%.

Hop	n	A0 Ext-F_1	AF Ext-F_1	Δ
2-hop	117	0.405	0.386	-0.019
3-hop	51	0.239	0.134	$\mathbf{-0.105}$
4-hop	32	0.184	0.107	$\mathbf{-0.077}$

Table 10: Per-hop breakdown of the production regression. The regression is monotone in chain depth: -0.019 at 2-hop, -0.105 at 3-hop, -0.077 at 4-hop. The diagnostic signature matches the prediction of (Cheng et al., 2025) that chain-of-thought scaffolds anchor on distractor hops in imperfect retrieval slates.

are verbose-but-correct (long answer with the gold span embedded), HippoRAG 2 outputs are terse-and-correct (gold span directly). Because the inflation ratio is asymmetric, the extraction step can change the cross-system *ranking*, not just the absolute level: HippoRAG 2 leads on raw F_1 at this cell (0.187 vs. 0.040, a 4.7× HippoRAG 2 win), while RAE leads on Ext-F_1 (0.275 vs. 0.257). We report both metrics throughout (Ext-F_1 in the body tables, raw F_1 in Appendix I) so the verbosity-driven inversion is visible to the reader; under either metric, the cost ratio is unchanged. We read this asymmetry as evidence of differing reader-output verbosity across systems, not as a metric advantage for either.

What the pre-scoring extraction does not do. The heuristic does not paraphrase, does not consult any model, and does not re-rank candidates by likelihood. It is a deterministic pre-scoring filter applied uniformly across systems; its only role is to prevent verbosity-driven precision collapse from contaminating cross-reader-size comparisons. It is not a contribution of this paper.

I Pareto Frontiers — Full Visualizations

Figure 4 replicates the body's six-panel Pareto grid (Figure 1) with the y-axis switched from Ext-F_1 to raw F_1 (the same MuSiQue/SQuAD kernel applied to the reader's full output, no extraction step). Under the canonical RAE L1+L2 pinned uniformly, RAE sits on the Pareto frontier in $4/6$ cells under raw F_1. The two off-frontier cells under raw F_1 are the small-reader cells on both benchmarks (`qwen3.5:0.8B` MuSiQue and HotpotQA), where the truncated dense baselines hold the cheap edge AND verbosity-driven precision collapse pushes RAE's raw-F_1 y-value below HippoRAG's — the same cells invert under Ext-F_1 once extraction normalizes reader output. Absolute scores are uniformly lower under raw F_1 for every system, with the gap most pronounced at `qwen3.5:0.8B` where reader verbosity is highest. The gap closes at `qwen3.5:4B` and 9B, where readers follow terseness instructions more reliably; this reader-size dependence of the Ext-F_1 vs. raw-F_1 gap is itself empirical evidence that the extraction protocol is doing what it should.

J RT-RAG Per-Query Token Cost: Per-Cell Measurements

RT-RAG's per-query token cost is reader-dependent on MuSiQue because the recursive retrieve-reason tree (branching 5, depth 4) expands further when the inner reader fails to converge on hard queries. We therefore report per-cell measurements rather than a single representative figure.

Per-cell measurements (this work). On the MuSiQue-200 split that RAE is evaluated on, the RT-RAG pipeline at the three local reader tiers logged the following mean per-query token totals: 10.5 M at `qwen3.5:0.8B`, 0.9 M at `4B`, and 3.6 M at `9B`. The same RT-RAG pipeline on HotpotQA-500 is essentially reader-independent at ∼318 k tokens/q (272 k–308 k prompt, 8 k–16 k completion across reader rows) because HotpotQA's two-hop questions converge faster and the tree rarely re-expands. The body tables (Table 1, Table 5) report the per-reader cell value at the row's reader; cost ratios versus RAE follow accordingly, with the qwen3.5:0.8B MuSiQue cell giving the maximum (∼9,500×).

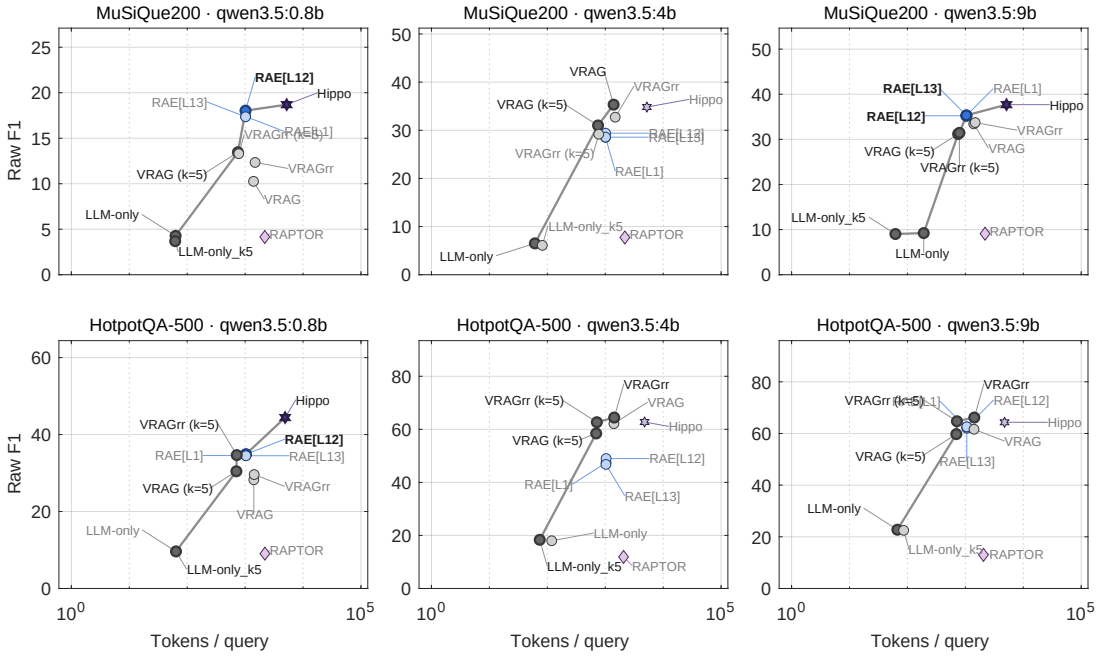

Figure 4: Multi-reader Pareto frontier under *raw* F_1 (no extraction step). Same systems, same axes, same cells as Figure 1; only the y-axis differs. Plot labels: VRAG/VRAGrr/Hippo = Vector-RAG / Vector-RAG + Reranker / HippoRAG 2; RAE[L12] = canonical RAE L1+L2; RAE[L3] = L3-only ablation (Appendix C). With the canonical RAE L1+L2 pinned uniformly, RAE sits on the Pareto frontier in $4/6$ cells under raw F_1 (off-frontier at the two small-reader cells: `qwen3.5:0.8B` on both MuSiQue and HotpotQA). The pattern is consistent with the verbosity-driven precision collapse the extraction protocol is designed to neutralize: at `qwen3.5:0.8B` MuSiQue, HippoRAG and Vector-RAG + Reranker are ahead of RAE on the raw-F_1 y-axis (HippoRAG 0.187 vs RAE L1+L2 0.040); the Ext-F_1 ranking inverts the same cell because the RAE reader's output is verbose-but-gold-bearing while HippoRAG's is terse-and-direct (App. H). Absolute scores are uniformly lower than Ext-F_1 (extraction can only narrow the predicted span), with the gap shrinking as readers strengthen.

Structural sanity check on the 10.5 M figure. The RT-RAG tree configuration (branching $B = 5$, depth $D = 4$, sampling $S = 5$ per node, max-variants $V = 2$, retry-on-ATTEMPT-2) produces an upper bound of $B^D = 625$ leaf trajectories per query. Each trajectory packs a context window of retrieved chunks plus generated reasoning at every node (typically 5–15 k tokens of context per node). At the small qwen3.5:0.8B reader the inner reader frequently fails to converge on the harder MuSiQue 4-hop questions, triggering ATTEMPT-2 re-expansion of the inner tree. The resulting mean per-query total of 10.5 M tokens is therefore consistent with ~ 600 trajectories $\times \sim 16$ k tokens each on the long tail of unconverged 4-hop queries, averaged over the full 200-query batch. Per-query JSON traces with ATTEMPT-1 and ATTEMPT-2 token-counts are in the anonymized supplementary materials.

Per-query cost variability. Per-query latency and cost are heavy-tailed across the eight-reader slate: the long tail of 4-hop questions triggers ATTEMPT-2 re-expansion and pushes individual queries into the multi-million-token regime. The mean is what we report; the median understates the practical cost differential. RT-RAG cost accounting and per-query latency CDFs are in the anonymized supplementary materials.

K Connection to the Controlled-Hallucination Model of Perception

The RAE architecture was conceived under the controlled-hallucination model of perception (Seth, 2021, 2014; Friston, 2010; Hohwy, 2013), which holds that perception is an iterative inferential process: the brain generates a top-down prediction — an imagined or hallucinated explanation — compares it to incoming sensory evidence, and revises the prediction until prior and evidence converge. The resulting percept is a fixed point of this closed-loop error-minimisation process. Recent presentations of this view (Seth, 2021) make the closed-loop structure explicit: a hierarchy of generative

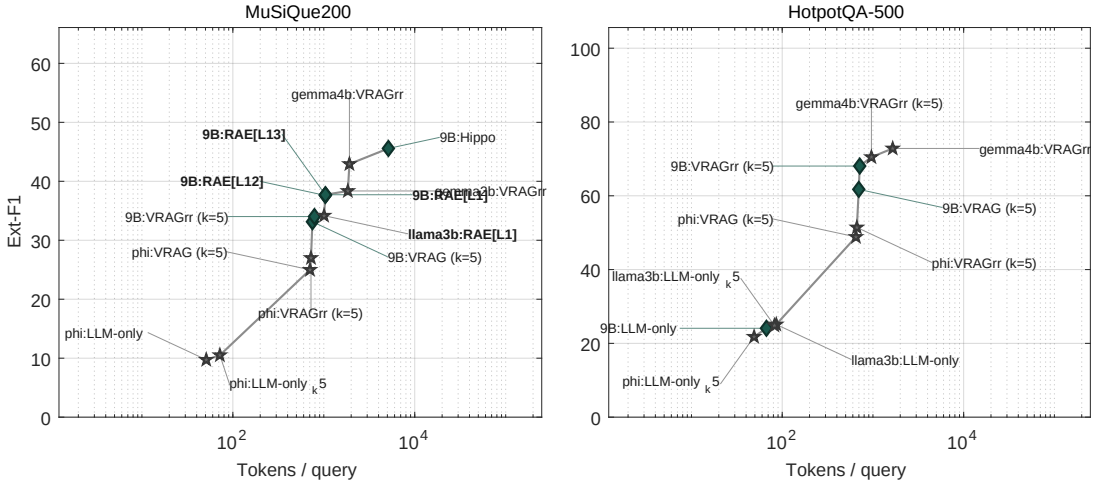

Figure 5: Cross-reader Pareto frontier (Ext-F_1) on MuSiQue (left) and HotpotQA (right), drawn across the eight-reader slate on a single log-x canvas. Companion to the body grid (Figure 1). On MuSiQue, RAE-cfg2 at the qwen tier sits on the frontier together with HippoRAG 2 (high-cost quality alternative) and budget-trimmed dense baselines (low-cost edge); on HotpotQA, the budget-trimmed dense methods hold the frontier and RAE sits inside it, consistent with the regime split discussed in Section 5.2. The composite view exposes the cost separation across families: RAE clusters around 1 000 tokens/q regardless of reader; HippoRAG around 5 000; RAPTOR around 2 100; RT-RAG above 10^5.

priors at progressively finer levels, sensory likelihoods that pass evidence upward, and prediction-error channels that propagate disagreement downward.

We translate this view into RAE as a hierarchical controller in which the LLM acts as the generative-prior subsystem, the semantic-equilibrium servo of Theorems 1–3 acts as the evidence-exploration (sensory-likelihood) subsystem, and prediction-error feedback closes the loop between them over a knowledge representation aligned with the query.

Generative prior. The LLM plays the role of the *generative-prior subsystem*. The planner emits a top-down hypothesis about the structure of the answer — the chain of hops with their predicates and expected entity types. This is a hallucination in the technical sense: the LLM generates a candidate decomposition before any retrieval is performed, drawing on its parametric prior over question-decomposition patterns. The hallucinatory framing is consistent with fundamental impossibility theorems on hallucination control (Xu et al., 2024): rather than try to suppress the prior, we use it as the planning substrate and probe it against external evidence. The chain is what the system *expects* the answer to look like, prior to evidence; the hops are the probes the rest of the architecture uses to test the prior. A lightweight *presearch* pass (a single dense-retrieval query against the corpus that returns a short list of attested entities used to constrain the planner's typed slots) optionally grounds the prior in corpus-attested entities before the planner runs, reducing prior–evidence mismatch at the first hop without altering the controller's structure.

Evidence exploration. The servo plays the role of the *sensory-likelihood subsystem*, exploring evidence available in the external knowledge graph. Given a hop's predicate, the servo deterministically computes the equilibrium semantic potential $\mathbf{x}_\infty = \mathbf{R}(\mathbf{u})\mathbf{r}$, which ranks chunks by their conductance-weighted relevance to the queried ideas. For any reference $\mathbf{r}$, the servo returns a unique non-negative semantic-potential vector with explicit spectral structure (Theorems 1–2). Two refinements modulate the output: *chunk seeding* augments $\mathbf{r}$ with the top-k chunks retrieved by dense similarity against the raw query (so the servo starts from a lexically-anchored boundary, not from idea-only seeds), and *BM25 conductance* rescales components of $\mathbf{u}$ by lexical-relevance scores (so the edges of high-BM25 chunks carry more current). The reranker then truncates and re-orders the top chunks. The servo's role is bottom-up — the corpus's response to the planner's probe.

Prediction-error channel. The verifier closes the loop as a local *prediction-error check*. Without access to external verification, the system is limited to encoded LLM knowledge and the corpus-

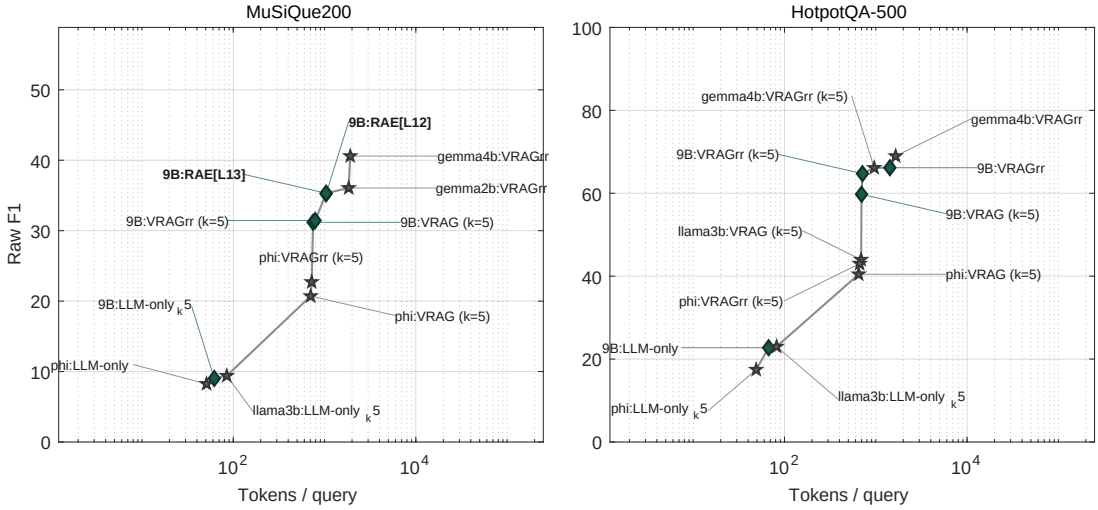

Figure 6: Cross-reader Pareto frontier under *raw* F_1 on MuSiQue and HotpotQA, drawn across the eight-reader slate. Companion to Figure 5: same systems and cells, raw F_1 replaces Ext-F_1 on the y-axis. Absolute values are lower under raw F_1 at small readers because verbose answer styles depress token overlap before the extraction step normalizes them; the qwen3.5 family converges on the two metrics by 4B and 9B. The cost separation across families is unchanged from Ext-F_1: RAE around 1 000 tokens/q, dense baselines 700–1 500, HippoRAG $\sim$ 5 000, RT-RAG well above 10^5.

derived structure. The verifier tests whether the extracted entity has the type predicted by the planner's hop; a type mismatch is one straightforward prediction error and is the feedback signal for the outer loop. Richer error estimators (e.g. entailment checks, multi-source agreement) are admitted by the design but generally require external knowledge.

Why this connection matters. Beyond architectural inspiration, the controlled-hallucination framing reads RAE as a small computational instance of a much-discussed cognitive-science position: that intelligent inference under uncertainty is necessarily a closed-loop dialogue between a generative prior and an evidence channel, with prediction error driving revision. Recent neurosymbolic work has begun to formalize this connection in adjacent domains — closed-loop verification of LLM outputs in process control (Galitsky and Rybalov, 2026), deterministic fact ledgers as evidence-channel surrogates (Agand, 2026), and the broader agentic-RAG taxonomy that distinguishes single-pass active retrieval from true closed-loop workflows (Mishra et al., 2026). RAE is, in this lineage, the closed-form deterministic instance: the generative prior is bounded to question planning; the evidence channel is a polynomial-time graph servo with three closed-form theorems; the prediction-error check is a typed verifier. The formal claims of the paper (Theorems 1–3) do not depend on any neuroscientific commitment, but the architecture's shape — specifically the two-level structure, the deterministic inner servo, and the prediction-error gate — becomes intelligible only against the broader perception-and-consciousness backdrop.