GRACE: Graph-Grounded Reflective Agent Copilot Engine for Expert-in-the-Loop Knowledge Expansion

BOSTON UNIVERSITY · MassMutual

John Seon Keun Yi[1] · Joshua R. Minot[2] · Dokyun Lee[1]

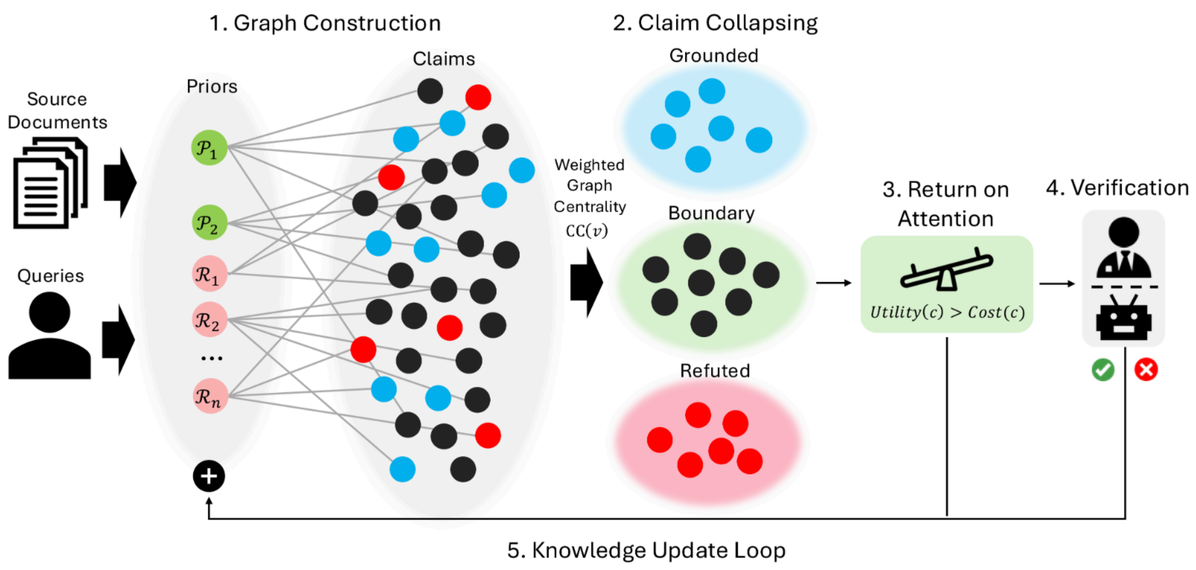
EMNLP2026 BUDAPEST

1 · MOTIVATION

- LLMs **confabulate rather than abstain** on out-of-knowledge queries, producing fluent but ungrounded claims.
- **RAG** retrieves isolated passages and cannot track how evidence across documents relates to each claim.
- **Uncertainty estimation** relies on the LLM's own consistency, so it is blind to *shared hallucinations*: false claims repeated confidently across samples.
- **Our claim:** ground outputs against trusted external priors, and spend expert attention only where it pays off.

2 · METHOD: GRACE

Graph-Grounded Reflective Agent Copilot Engine



- **Decompose** source documents and LLM responses into atomic claims; anchor them against trusted knowledge priors.
- **Weighted bipartite graph** of priors, responses and claims with four edge types:

  Anchor > *Agreement* > *Neutral* > *Contradictory*

- **Claim collapsing:** weighted closeness centrality CC(v) classifies each claim:
  - **Grounded** – anchored to priors
  - **Boundary** – novel or contested: the knowledge frontier
  - **Refuted** – contradicts priors
- **Return on Attention (RoA):** defer a boundary claim to an expert **only if** its priority-weighted uncertainty exceeds the cost of verification:

$$Priority(c) \cdot (1 - \sigma(CC(c))) > Cost(c)$$

- Priority combines **centrality** (verifying it collapses clusters of related claims) and **novelty** (it expands the frontier); claims are picked greedily under a budget B.
- **Knowledge update loop:** expert-confirmed claims are promoted to new priors and edges are re-weighted, so each round resolves more claims automatically.
- **Grounded retrieval:** at query time, passages are ranked by query similarity *and* closeness centrality, so retrieved context is both relevant and well-grounded.

3 · RETRIEVAL

Can GRACE serve as a reliable retrieval foundation?

	QASPER [(Dasigi et al., 2021)]		QuALITY [(Pang et al., 2022)]
	Accuracy (%)	Answerable (%)	Accuracy (%)
GPT-4o-mini [(OpenAI, 2024)]			
Zero-shot + Part	10.8	26.6	56.2
Zero-shot + Full	29.0	53.6	79.6
RAG	19.8	45.2	66.5
GraphRAG	27.8	90.4	85.3
GRACE (Ours)	**28.1**	**50.6**	**77.0**
LLaMA 3.1 8B [(Dubey et al., 2024)]			
Zero-shot + Part	10.7	26.6	41.1
Zero-shot + Full	29.0	29.0	45.1
RAG	18.5	35.6	58.9
GraphRAG	29.0	95.2	29.1
GRACE (Ours)	**28.1**	**50.5**	**44.2**

- Matches **full-context accuracy** (28.1 vs 29.0 on QASPER) from only the top-5 passages; **+8–10 pts over RAG**.
- Calibrated: answers ~50% of questions, tracking its accuracy, while GraphRAG answers >90%.

4 · KNOWLEDGE EXPANSION

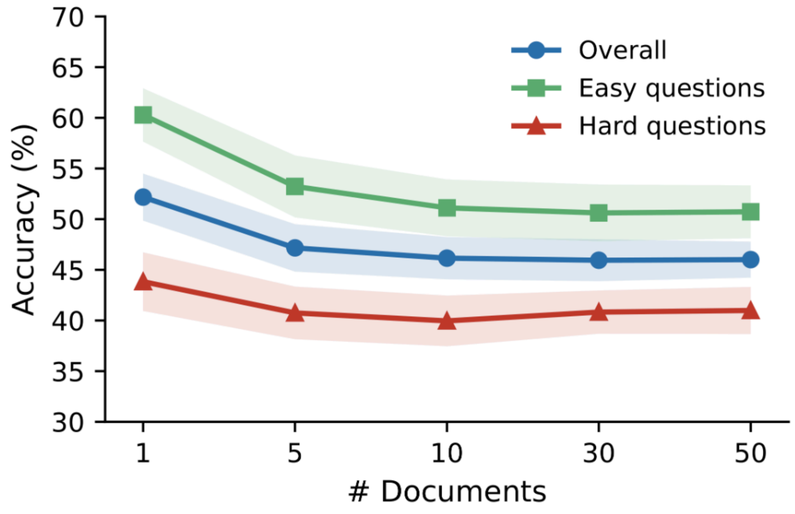
Do verified claims compound into a better knowledge base?

Context	Accuracy (%)	Δ
Abstract only	0.1413	—
Round 0 (Abstract + Intro)	0.2225	+8.1
Round 1 (Batch 1)	0.2447	+10.3
Round 2 (Batch 2)	0.3716	+23.0
Round 3 (Batch 3)	0.4913	+35.0
Full paper	0.3360	+19.5

- QA accuracy climbs 14.1% → 49.1% in three rounds, +15.5 pp over the full paper as context.
- **97.6% grounding precision**; human verifiers (15 NLP grad students) match the LLM verifier on coverage and accuracy.

5 · SCALABILITY

Does GRACE hold up as the knowledge base grows?



- Claim pool grows **50×** (91 → 4,554 claims), yet accuracy drops **< 6 pp** and plateaus after 5 documents.

6 · TAKEAWAYS

- **Ground, don't self-check:** anchoring claims to trusted priors separates facts from shared hallucinations.
- **Spend attention where it pays:** RoA-guided expansion builds a knowledge base that beats the full paper as QA context.
- **Reliable at scale:** full-context accuracy with calibrated answerability, steady as the corpus grows 50×.