NeurIPS 2026

40th Conference on Neural Information Processing Systems

SAGE: Semantic Ambiguity Guided Capacity Expansion for Retrieval-Augmented Generation

Xunlei Chen, Jinyu Guo, Yi Gong, Qirui Ye, Zhaokun Wang*, Wenyi Li, Wenhong Tian*

School of Information and Software Engineering
University of Electronic Science and Technology of China

Why Similar Passages Are Hard to Rank

Where ranking breaks down

- **RAG needs precise evidence**

  Multi-hop questions often require supporting passages from several sources.

- **Similar passages compete**

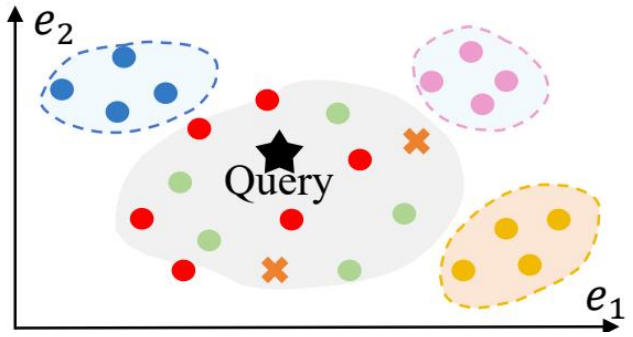
  Fixed-dimensional scoring can give close scores to evidence and distractors.

- **Routing can add ambiguity**

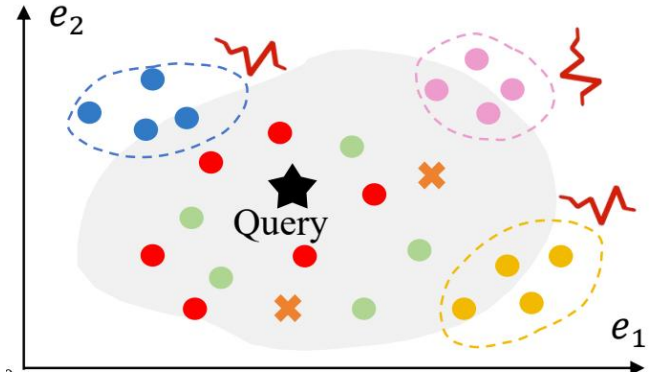
  Trees and graphs narrow the search, but routing or summaries can introduce noise.

Semantic entanglement

Flat dense retrieval



After structured routing



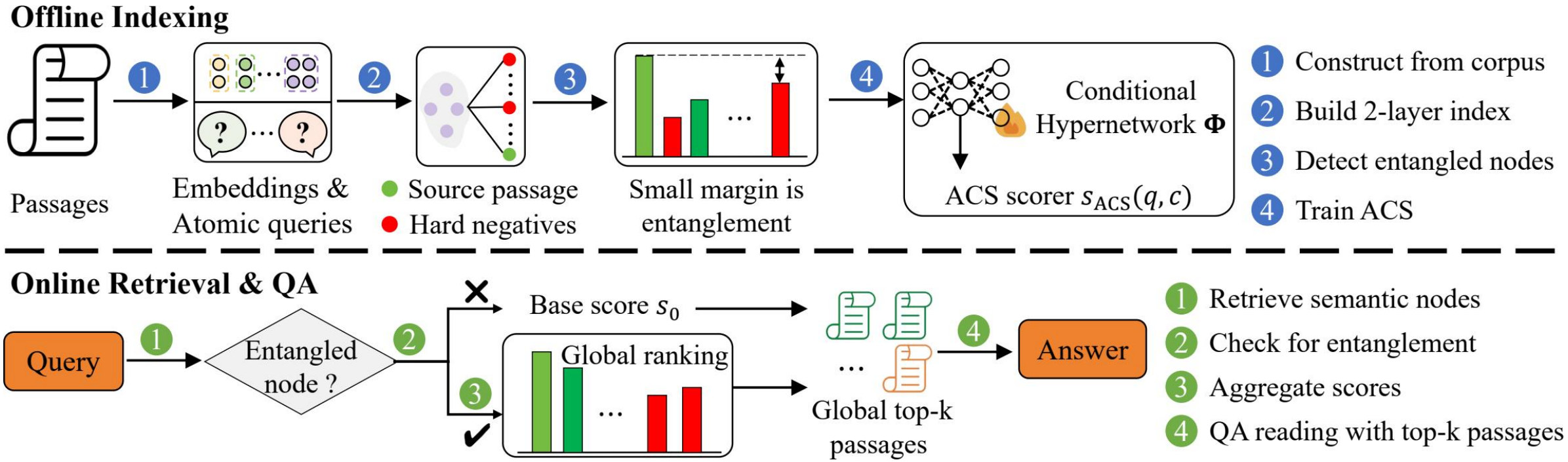
Relevant Irrelevant Ambiguous

Even within a routed node, the base scorer can struggle to separate the right passage from close competitors.

Which semantic neighborhoods need a stronger scorer?

SAGE Overview

Offline diagnosis and online retrieval



Offline diagnosis with LSD

Local Separability Deficit (LSD) flags nodes where source-to-negative score margins fall short.

Online scoring with ACS

Use the Ambiguity-Conditioned Scorer (ACS) in flagged nodes and cosine scoring elsewhere.

LSD Measures Where Cosine Scoring Struggles

An offline probe generated from a passage

An atomic question from the corpus

Compare the source with the strongest filtered hard negative in the same node.

Source passage 0.72

Passage that answers the question

Hard negative 0.68

Top-scoring local distractor

Illustrative cosine scores

From one probe to a node score

Observed gap 0.72 − 0.68 = 0.04

Shortfall = max(0, target margin − gap)

Local Separability Deficit (LSD)

Average the shortfall over the atomic questions from this node.

Flag the top 20% of nodes by LSD before serving user queries

ACS Adapts Scoring to the Query and Node

Ambiguity-Conditioned Scorer (ACS)

1. Condition

Query + node summary

The query embedding tells us what to look for. The summary describes the local context.

→

2. Generate weights

Shared hypernetwork

Generate the weights of a small nonlinear scorer for this query and node.

→

3. Score passages

Local ACS

Apply that scorer to frozen passage embeddings in the flagged node.

Calibration uses the same kind of corpus-derived triplets

A hinge loss rewards a sufficient lead for the source over its hard negatives. Only the hypernetwork is updated.

ACS ranks passages in flagged nodes. Other nodes use cosine.

A Brief Retrieval Example

Who directed the 2015 film based on Andy Weir's debut novel?

Author node

Cosine finds the novel

Andy Weir's debut novel is The Martian.

Flagged film node

ACS promotes the right film

The passage on The Martian (2015) names Ridley Scott as its director.

Min-max normalize per node, keep the best score per passage, and return the global top-k.

Answer Ridley Scott

SAGE Improves Retrieval and Answer Quality

Passage recall (%) and downstream token-level F1 (%)

Method	Avg. Recall@2	Avg. Recall@5	2Wiki Recall@5	Avg. QA F1
NV-Embed-v2	58.0	**72.3**	75.5	**54.3**
HippoRAG 2	59.3	**75.4**	88.2	**55.2**
SAGE	**61.2**	**76.7**	**89.4**	**56.7**

Retrieval across five benchmarks

SAGE leads on average Recall@5 by 1.3 points over HippoRAG 2.

Answer quality across six benchmarks

All systems use Llama-3.3-70B-Instruct as the answer reader.

Where We Spend the Extra Capacity Matters

Same two-layer index. Different policies for activating ACS.

Scoring policy	ACS nodes	Avg. Recall@5 (%)	End-to-end QA time
Cosine only	None	**72.9**	**0.27 s**
Random ACS	Random 20%	**73.6**	**0.66 s**
All-node ACS	All nodes	**74.4**	**1.90 s**
SAGE	**LSD 20%**	**76.7**	**0.70 s**

Same 20% budget

Selecting with LSD adds 3.1 Recall@5 points over random node selection.

Compared with ACS on every node

SAGE gains 2.3 recall points and cuts mean QA time from 1.90 s to 0.70 s.

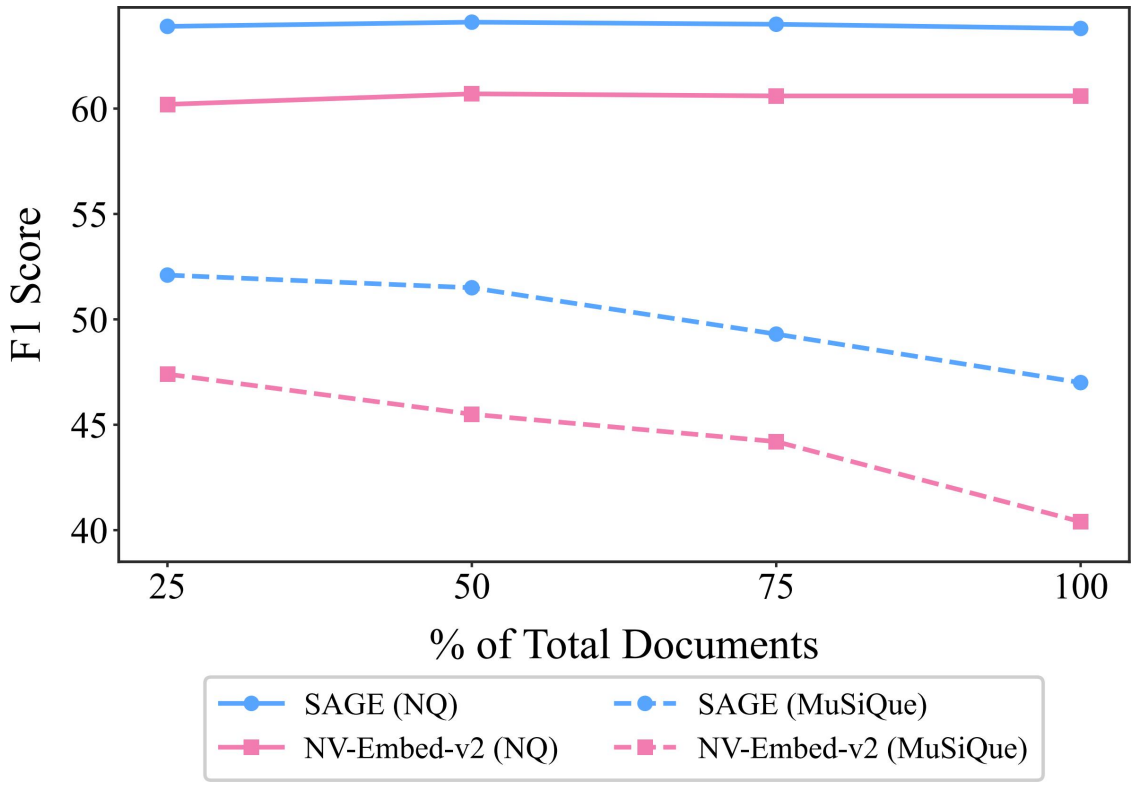
LSD Tracks Retrieval Difficulty

LSD alignment on benchmark queries

LSD percentile	Share (%)	Error@5 (%)	Gain (pp)
0–20%	12.0	20.6	**+0.4**
20–40%	15.2	25.6	**+1.0**
40–60%	18.8	30.1	**+2.1**
60–80%	24.2	35.9	**+4.2**
80–100%	**29.8**	**43.0**	**+7.3**

Error is measured with two-layer cosine scoring.
Gain is from separate, bin-specific ACS calibration.

Robustness as the passage pool grows



SAGE stays ahead of NV-Embed-v2 on both datasets as the pool grows

Thank you for listening

Add scoring capacity where cosine struggles

Questions and discussion

Xunlei Chen

University of Electronic Science and Technology of China

202321090308@std.uestc.edu.cn