GRO-RAG: Gradient-Driven Subset Selection for Multi-source Retrieval-Augmented Generation

Abstract

Retrieval-augmented generation (RAG) increasingly relies on evidence pools gathered from diverse repositories. In practice, pipelines often either blend sources uniformly or commit to a single source beforehand—choices that suppress complementary semantics and inflate redundant overlap. Prevailing re-ranking modules also emphasize proxy relevance, leaving a gap between ranking scores and the generator's training objective. GRO-RAG addresses these issues with a training-free mechanism that uses loss-sensitive gradient signals from the language model to govern subset selection. From one backward pass, we estimate each candidate document's marginal influence on the next-token negative log-likelihood and choose a k-size subset accordingly. Thus, ordering is determined by direct feedback from the generation objective rather than heuristic similarity. At the pool level, GRO-RAG first composes a source mixture by balancing query affinity against cross-source redundancy, before document-wise selection takes place. We show that this greedy, gradient-informed strategy approximates the loss-minimizing subset and optimizes a surrogate upper bound for leave-one-out risk. Across multi-source QA and open-domain generation settings, GRO-RAG yields consistent gains over uniform aggregation and static-source baselines, reinforcing the value of generation-aware retrieval selection.