RAG e Suas Alternativas: Um Framework para o Aprimoramento de Grandes Modelos de Linguagem

Maria C. X. de Almeida, Anna P. V. L. B. Moreira, Evellyn M. R., Sávio S. T. de Oliveira, Fernando M. Federson[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)

{mariaalmeida2, annapietra, nicole}discente.ufg.br

{savioteles, federson}ufg.br

Abstract. *To address limitations in LLM static knowledge, this work proposes a taxonomy based on five temporal categories and a decision framework for LLM enhancement. We demonstrate that technique selection depends on specific problem characteristics, establishing that no single solution is universally superior.*

Resumo. *Para abordar limitações de conhecimento estático em LLMs, este trabalho propõe uma taxonomia baseada em cinco categorias temporais e um framework de decisão para aprimoramento de LLMs. Demonstramos que a seleção da técnica depende das características específicas do problema, não havendo solução universalmente superior.*

1. Introdução

Grandes Modelos de Linguagem (LLMs) revolucionaram o processamento de linguagem natural, mas enfrentam limitações críticas relacionadas ao conhecimento paramétrico estático, incorporado apenas durante o pré-treinamento. Isso resulta em dois problemas inter-relacionados: a obsolescência imediata do conhecimento após o treinamento e a ausência de acesso a dados proprietários ou especializados que não estão presentes no corpus original. A Geração Aumentada por Recuperação (RAG), proposta por Lewis e colaboradores em 2020[1], emergiu como uma solução promissora ao combinar capacidade generativa com recuperação externa em tempo real, evitando o retreinamento do modelo. A técnica rapidamente ganhou adoção prática, inclusive pela comunidade brasileira, com aplicações em domínios jurídicos e educacionais. Entretanto, desenvolvimentos tecnológicos recentes — notadamente a expansão das janelas de contexto para milhões de tokens, a maturação de APIs de busca especializadas e técnicas eficientes de Fine-Tuning (LoRA, QLoRA) — questionam a universalidade da RAG como solução preferencial. Esse cenário de proliferação de alternativas cria um desafio prático sobre como escolher a técnica mais adequada para cada problema específico. A literatura atual tende a analisar as técnicas isoladamente, carecendo de frameworks comparativos que orientem decisões arquiteturais de forma sistemática.

[1]Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In Advances in Neural Information Processing Systems (NeurIPS).

2. Objetivo e Metodologia

O objetivo principal deste trabalho é apresentar uma ferramenta prática de apoio à decisão sobre técnicas de aprimoramento de LLMs, tendo a RAG como referência. A metodologia baseia-se em três pilares: a síntese sistemática da literatura atual sobre técnicas de aprimoramento de LLMs; desenvolvimento de uma taxonomia organizacional estruturada de acordo com o momento de disponibilização do conhecimento ao modelo; e a construção de um framework de decisão baseado em critérios objetivos extraídos da análise comparativa.

3. Resultados

A taxonomia temporal proposta organiza as técnicas em cinco categorias fundamentais: C1 - Internalização de Conhecimento (Fine-Tuning completo, LoRA, QLoRA), onde o conhecimento é incorporado aos pesos do modelo durante o treinamento, oferecendo independência de infraestrutura externa, mas com alto custo de retreinamento; C2 - Recuperação de Conhecimento Externo (RAG tradicional, Hybrid Retrieval, Graph-RAG, Search-First, Tool-Augmented), implementando acesso just-in-time a fontes externas com máxima flexibilidade para dados dinâmicos; C3 - Expansão da Janela de Contexto (Long Context), fornecendo todo o conhecimento diretamente no prompt através de janelas de 1M+ tokens; C4 - Otimização do Raciocínio (Chain-of-Thought), melhorando o processamento da informação existente sem adicionar conhecimento novo; e C5 - Protocolos e Frameworks de Habilitação (Knowledge Graphs, Model Context Protocol), viabilizando a infraestrutura para outras categorias. O framework de decisão estrutura-se em componentes integrados: taxonomia temporal organizando técnicas em cinco categorias (C1-C5); tabela comparativa multidimensional analisando volume de dados, frequência de atualização, privacidade, recursos computacionais, latência, orçamento e expertise técnica; árvore de decisão baseada nas características do problema; e matriz de adequação técnica para cenários de aplicação. A análise comparativa oferece uma base objetiva para a seleção técnica fundamentada em contextos específicos. Os resultados obtidos, assim como o artigo completo, estão disponíveis em um Repositório GitHub.

4. Conclusão

Este trabalho apresentou uma análise estruturada das técnicas de aprimoramento de LLMs, tendo a RAG como referência central. A contribuição principal reside na transformação do debate teórico sobre técnicas de aprimoramento em uma ferramenta prática de apoio à decisão, particularmente relevante para a comunidade brasileira de pesquisa em IA. Os resultados demonstram que não existe uma solução universalmente superior, sendo a seleção apropriada fundamentalmente dependente das características específicas do problema. Direções futuras incluem estudos longitudinais sobre a evolução das técnicas e a adaptação a diferentes domínios ao longo do tempo, bem como a manutenção contínua do framework frente ao cenário tecnológico em rápida evolução.