Abstract

Large Language Models (LLMs) have achieved remarkable success in natural language processing tasks, particularly in question-answering (QA). However, their integration into real-world systems is often hindered by two key challenges: first, the need for structured, concise answers and reliable confidence estimation to support aggregation across multiple reasoning paths; and second, the large model size and computational cost, which limit deployment efficiency.

To address these challenges, this thesis presents two complementary contributions. First, we introduce ANSPRE, a structured answer generation framework that guides LLMs to produce concise answers with reliable confidence scores. ANSPRE improves aggregation across retrievals and reasoning chains, enhances answer quality, and generalizes across multilingual and vision-language QA tasks. Extensive experiments on open-domain, multilingual, and visual QA benchmarks demonstrate that ANSPRE significantly improves Exact Match (EM) and F1 scores while providing well-calibrated confidence estimates.

Second, we investigate the behavior of pruning techniques across varying sparsity levels and identify key findings that inform optimal pruning strategies. Building on these insights, we propose OPTIPRUNE, a method that dynamically selects the most suitable pruning approach for each sparsity regime. Empirical evaluation shows that OPTIPRUNE consistently outperforms state-of-the-art pruning methods across multiple architectures, benchmarks, and language-specific calibrations, enabling efficient deployment without significant performance degradation.

Together, ANSPRE and OPTIPRUNE advance LLMs toward high-quality, reliable, and deployment-efficient systems, addressing critical gaps in structured reasoning, confidence estimation, and model compression.

[**Keywords**] Model Reliability, Model Efficiency, Retrieval-Augmented Generation, Model Pruning, Question-Answering, Large-Language-Model