

OpenScholar Results

Question 1

What are the main approaches for retrieval-augmented generation in large language models?

Answer

The main approaches for retrieval-augmented generation in large language models include RAG (Retrieval-Augmented Generation) [0], REALM (Retrieval-Augmented Language Model Pre-Training) [1], and Self-RAG (Self-Reflective Retrieval-Augmented Generation) [2]. These methods combine pre-trained parametric and non-parametric memory for language generation by using a dense vector index of Wikipedia or other text datasets, accessed with a pre-trained neural retriever.

References

[0] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks □ - Citations: 2500 - Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability t...

[1] REALM: Retrieval-Augmented Language Model Pre-Training □ - Citations: 1800 - Language model pre-training has been shown to capture a surprising amount of world knowledge, crucial for NLP tasks such as question answering.

[2] Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection □ - Citations: 450 - Despite their remarkable capabilities, large language models (LLMs) often produce responses containing factual inaccuracies due to their sole reliance on the parametric knowledge they encapsulate.

Question 2

How do transformer models handle long-range dependencies in sequences?

Answer

Transformer models can handle long-range dependencies in sequences through various techniques such as using self-attention [0], which allows them to attend to all positions in the input sequence simultaneously, and incorporating local windowed attention with task-motivated global attention in architectures like Longformer [1]. This enables transformer models to capture relationships between distant tokens in a sequence without being limited by sequential processing or position-wise interactions.

References

[0] Attention Is All You Need □ - Citations: 95000 - The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder.

[1] Longformer: The Long-Document Transformer □ - Citations: 3200 - Transformer-based models are unable to process long sequences due to their self-attention operation, which scales quadratically with the sequence length.
