Advanced RAG 02: Uncover the secrets of PDF parsing

Florian June [AI Study Room](javascript:void(0);) May 13, 2024 11:21

Extracting information from documents is an inevitable step for RAG (Retrieval-Augmented Generation). Ensuring effective content extraction from source documents is crucial to improving the final output quality.

This process should not be underestimated when implementing RAG. Poor information extraction during the parsing process can result in limited understanding and use of the information contained in the PDF file.

In actual work, unstructured data is much richer than structured data. If these large amounts of data cannot be parsed, their huge value cannot be realized.