Multi-Modal Embeddings: Comparative Study 2026

Research Overview

Comprehensive evaluation of multi-modal embedding approaches for document understanding systems. This study compares text-only, vision-only, and combined embedding strategies for RAG applications processing documents with mixed content types.

Key Findings
• Combined text+vision embeddings improve retrieval by 31%
• Vision-language models excel at diagram understanding
• Text embeddings remain superior for pure textual content
• Computational overhead is manageable for production use

Recommended Architecture
For OneNote RAG systems: Use text-embedding-3-large for text content and combine with vision embeddings for image-heavy documents. This hybrid approach maximizes both accuracy and efficiency while maintaining reasonable computational costs.