LLM Evaluation Report

Model: Mistral-7B-Instruct

Test Corpus: 20 startup PDFs

Top Retrieval Quality: 87.5%

Hallucination Rate: 5.2%

Inference Cost (4-bit): ~1.3GB RAM, 120ms latency

Conclusion: Best balance for edge deployment.