Good to have prereq's

Embeddings	website/blog	https://jalammar.github.io/illustrated-word2vec/
Attention is all you need	paper	https://arxiv.org/abs/1706.03762
Illustrated transformer by jay alammar	blog	https://jalammar.github.io/illustrated-transformer/
Decoder side of Tranformer(GPT)	Video	https://www.youtube.com/watch?v=kCc8FmEb1nY
Tokenization	Substack article	https://seantrott.substack.com/p/tokenization-in-large-language-models
Positional Encoding	website/blog	https://machinelearningmastery.com/a-gentle-introduction-to-positional-encoding-in-transformer-models-part-1/
Mixture of experts	blog	https://huggingface.co/blog/moe
Open LLM leader board to understand metrics	blog	https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
Finetuning LLMs	Coursera course/project	https://www.coursera.org/projects/finetuning-large-language-models-project

RAG

LLM, RAG and EVALUATION	Medium Blog series(amazing blog)	https://medium.com/@vipra_singh/building-llm-applications-introduction-part-1-1c90294b155b#4d28	
Embeddings and Hybrid Retreival	Blog	https://www.linkedin.com/pulse/guidebook-state-of-the-art-embeddings-information-aapo-tanskanen-pc3mf/	
Prompt Engineering	Coursera course/project	https://www.coursera.org/projects/chatgpt-prompt-engineering-for-developers-project	
Langchain RAG playlist	Youtube	https://youtu.be/wd7TZ4w1mSw?si=BzbT8-P9an1pGnsZ	
Agentic RAG with llama index	Coursera course/project	https://www.coursera.org/projects/building-agentic-rag-with-llamaindex	
Langchain Chat with you data	Coursera course/project	https://www.coursera.org/projects/langchain-chat-with-your-data-project	
Function tools and agents with langchain	Coursera course/project	https://www.coursera.org/projects/functions-tools-and-agents-with-langchain-project	
AI agents in Langgraph	Coursera course/project	https://www.coursera.org/projects/ai-agents-in-langgraph	
RAG Evaulation and monitoring(RAGAS)		https://docs.ragas.io/	
Building and Evaluation of advanced RAG Applications		https://www.deeplearning.ai/short-courses/building-evaluating-advanced-rag/	
Building LLM applications for production	Blog	https://huyenchip.com/2023/04/11/llm-engineering.html	
Pydantic and outlines for LLM output structure	github	https://github.com/outlines-dev/outlines	
Graph RAG	github	https://github.com/microsoft/graphrag	
Knowledge Graphs for RAG	Coursera	https://www.deeplearning.ai/short-courses/knowledge-graphs-rag/	
More to be added on graph rag...			
			Can run and experiment in colab notebooks
Practical experiments with Llama Index (continuously updated by team)	Llama index website	https://docs.llamaindex.ai/en/stable/examples/	
Langchain Integrations	langchain website	https://python.langchain.com/v0.1/docs/integrations/platforms/	

RAG Papers and surveys (continuously updated based on relevance)

A survey on Rag meeting LLMs	Survey Paper	https://arxiv.org/abs/2405.06211	5/10/2024
Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity	Paper	https://arxiv.org/pdf/2403.14403	5/28/2024
Retrieval Augmented Generation for Domain-specific Question Answering	Paper	https://arxiv.org/pdf/2404.14760	5/29/2024
RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems	Paper	https://arxiv.org/pdf/2407.01219	6/25/2024
M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation	Paper	https://www.semanticscholar.org/reader/4d5735c186ddb2430ac9689ccf61fdcbbfc23abc	6/29/2024
LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs	Paper	https://arxiv.org/pdf/2406.15319	6/30/2024
Searching for Best Practices in Retrieval-Augmented Generation	Paper	https://arxiv.org/pdf/2407.01219	7/1/2024
Retrieval-Augmented Generation for Natural Language Processing: A Survey	Paper	https://arxiv.org/pdf/2407.13193	7/19/2024
Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach	Paper	https://arxiv.org/pdf/2407.16833	7/23/2024
Long Context RAG Performance of LLMs	blog(survey) from databricks	https://www.databricks.com/blog/long-context-rag-performance-llms	8/12/2024
Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity	Paper	https://arxiv.org/pdf/2403.14403	5/28/2024
Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity	Paper	https://arxiv.org/pdf/2403.14403	5/28/2024
Retrieval Augmented Generation for Domain-specific Question Answering	Paper	https://arxiv.org/pdf/2404.14760	5/29/2024
RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems	Paper	https://arxiv.org/pdf/2407.01219	6/25/2024
M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation	Paper	https://www.semanticscholar.org/reader/4d5735c186ddb2430ac9689ccf61fdcbbfc23abc	6/29/2024
LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs	Paper	https://arxiv.org/pdf/2406.15319	6/30/2024
Searching for Best Practices in Retrieval-Augmented Generation	Paper	https://arxiv.org/pdf/2407.01219	7/1/2024
Retrieval-Augmented Generation for Natural Language Processing: A Survey	Paper	https://arxiv.org/pdf/2407.13193	7/19/2024
Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach	Paper	https://arxiv.org/pdf/2407.16833	7/23/2024
Long Context RAG Performance of LLMs	blog(survey) from databricks	https://www.databricks.com/blog/long-context-rag-performance-llms	8/12/2024
Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity	Paper	https://arxiv.org/pdf/2403.14403	5/28/2024