/mlst

The scientist who coined retrieval augmented generation

March 24, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: d5ba9701

Topic headings include retention indicators: ● above average, ● average, ● below average (vs. similar YouTube videos)

Interested in licensing ReScript for your podcast? Get in touch with Tim Scarfe.

Contents

Overview 2

Intro to RAG 3

RAG Evaluation: Poll framework and model performance 3

Data Quality: Cleanliness vs scale in AI training 3

Human-AI Collaboration: Research agents and UI design 3

RAG Origins: Open-domain QA to generative models 3

RAG Challenges: Info retrieval, tool use, faithfulness 3

Dense vs Sparse Retrieval: Techniques and trade-offs 3

RAG Applications: Grounding, attribution, hallucination prevention 4

UI for RAG: Human-computer interaction and model optimization 4

Word Embeddings: Word2Vec, GloVe, and semantic spaces 4

Language Model Evolution: BERT, GPT, and beyond 4

AI and Human Cognition: Sequential processing and chain-of-thought 4

References 5

Overview

Dr. Patrick Lewis, who coined the term Retrieval Augmented Generation and now works at Cohere, provides an insider account of how RAG came to be and where it is heading. Starting from his early work on the LAMA probe and the KILT benchmark, Lewis traces the intellectual path from open-domain question answering to the creation of RAG as a framework for grounding language model outputs in retrieved evidence.

The conversation covers the full technical landscape: evaluation challenges across faithfulness, fluency, and perceived utility; the tension between data quality and scale; dense versus sparse retrieval methods and when each excels; and the practical difficulties of deploying RAG in enterprise settings where link structure and training data are scarce. Lewis offers candid perspectives on prompt engineering, Cohere's approach to building retrieval-aware models through post-training, and why current evaluation benchmarks have been outpaced by model quality.

The discussion takes a historical turn through the evolution of language representations, from Mikolov's Word2Vec – which famously rediscovered the periodic table from chemistry literature – through GloVe and BERT to modern generative models. Tim Scarfe and Lewis close with reflections on the parallels between AI sequential processing and human cognition, and what chain-of-thought reasoning reveals about both.

Intro to RAG ●

(No transcript segments for this topic)

RAG Evaluation: Poll framework and model performance ●

(No transcript segments for this topic)

Data Quality: Cleanliness vs scale in AI training ●

(No transcript segments for this topic)

Human-AI Collaboration: Research agents and UI design ●

(No transcript segments for this topic)

RAG Origins: Open-domain QA to generative models ●

(No transcript segments for this topic)

RAG Challenges: Info retrieval, tool use, faithfulness ●

(No transcript segments for this topic)

Dense vs Sparse Retrieval: Techniques and trade-offs ●

(No transcript segments for this topic)

RAG Applications: Grounding, attribution, hallucination prevention ●

(No transcript segments for this topic)

UI for RAG: Human-computer interaction and model optimization ●

(No transcript segments for this topic)

Word Embeddings: Word2Vec, GloVe, and semantic spaces ●

(No transcript segments for this topic)

Language Model Evolution: BERT, GPT, and beyond ●

(No transcript segments for this topic)

AI and Human Cognition: Sequential processing and chain-of-thought ●

(No transcript segments for this topic)

References

[1] **Patrick Lewis Homepage**
https://www.patricklewis.io/
person

[2] **Cohere Command Models**
https://cohere.com/command
tool

[3] **Language Models as Knowledge Bases?**
https://arxiv.org/abs/1909.01066
paper

[4] **KILT: A Benchmark for Knowledge Intensive Language Tasks**
https://arxiv.org/abs/2009.02252
paper

[5] **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**
https://arxiv.org/abs/2005.11401
paper

[6] **Efficient Estimation of Word Representations in Vector Space**
https://arxiv.org/abs/1301.3781
paper

[7] **GloVe: Global Vectors for Word Representation**
https://nlp.stanford.edu/projects/glove/
paper

[8] **BERT: Pre-training of Deep Bidirectional Transformers**
https://arxiv.org/abs/1810.04805
paper

[9] **The Language Game**
https://amzn.to/4grEUpG
book