AI Buzzwords & Jargon — With Open■Source & Vendor Products (2024–2025)

1. RAG (Retrieval-Augmented Generation)

The dominant architecture for enterprise GenAI applications.
Products: LangChain, LlamaIndex, Vertex AI Search, OpenAI RAG, Elastic RAG.

2. Embeddings

Numerical vectors that capture semantic meaning of text.
Products: SentenceTransformers, HuggingFace models, Vertex Embeddings, OpenAI Embeddings.

3. Vector Databases

Stores embeddings and performs fast semantic + hybrid search.
Products: Qdrant, Weaviate, Pinecone, Milvus, Chroma, OpenSearch.

4. ANN (Approximate Nearest Neighbor Search)

Algorithms for retrieving nearest vectors efficiently.
Products: FAISS (Meta), ScaNN (Google), HNSWlib, Qdrant ANN engine.

5. Hybrid Search (BM25 + Vector)

Combines lexical and semantic retrieval for higher precision & recall.
Products: Weaviate Hybrid, Qdrant Hybrid, Vespa.ai, OpenSearch Hybrid.

6. Chunking Strategy

Splitting documents into optimized segments for retrieval quality.
Products: LangChain TextSplitter, LlamaIndex NodeParser.

7. Cosine Similarity

Measures the angle between vector embeddings for semantic similarity.
Products: NumPy, SciPy, FAISS.

8. LLM Context Window

Maximum tokens an LLM can process; larger windows increase reasoning.
Products: Gemini 1.5, Claude 3, GPT■4.1, Llama■3.

9. Prompt Engineering

Designing prompts to optimize LLM behavior and outputs.
Products: LangChain prompt templates, Vertex Prompt Manager, OpenAI assistants.

10. Grounding / Hallucination Reduction

Ensures answers stay aligned with enterprise truth sources.
Products: Google Vertex Grounding, Azure RAI filters, NeMo Guardrails.

11. Fine-Tuning / LoRA

Adapting models to domain-specific data using lightweight training.
Products: HuggingFace PEFT, Axolotl, OpenAI Fine-Tuning, Vertex Tuning.

12. Cross-Encoder Reranking

Re-ranks retrieved candidates for higher precision than ANN alone.
Products: ColBERT, MonoT5, Nvidia Reranker, Cohere Reranker.

13. vLLM / Model Serving

High-throughput model inference optimized for production latency.
Products: vLLM, Triton Inference Server, HuggingFace TGI, Ray Serve.

14. Quantization (INT8 / INT4)

Compressing model weights for faster and cheaper deployment.
Products: GGUF, GPTQ, AWQ, TensorRT■LLM.

15. Agentic Workflow / Tool Use

LLMs performing multi-step tasks and calling external tools.
Products: LangGraph, CrewAI, OpenAI Agents, Google Agent Builder.

16. Mixture of Experts (MoE)

Model architecture where only expert subsets activate per token.
Products: Mixtral 8x7B, DeepSeek MoE, Google Switch Transformer.

17. Safety Guardrails / Moderation

Ensures compliance, safety, and prevention of harmful output.
Products: Azure Content Safety, Google Safety Filters, NeMo Guardrails.

18. Knowledge Cutoff

Latest date included in model training data.
Products: GPT■4.1, Gemini 1.5, Claude 3 — each with distinct cutoffs.

19. AI Governance / Compliance

Enterprise controls for risk, privacy, and regulatory compliance.
Products: IBM watsonx.governance, Azure Policy, Google Model Evaluation.

20. Latency Budgeting

Managing time across retrieval, ranking, and generation.
Products: Redis Cache, vLLM batching, Qdrant ANN tuning.