2024 AI Technology Trends Report

Executive Summary

This report analyzes major trends in artificial intelligence technology as of 2024. It focuses on the development of RAG (Retrieval-Augmented Generation) systems and the practical application of Large Language Models (LLMs).

1. Evolution of Large Language Models

In 2024, next-generation language models such as GPT-4, Claude 3, and Gemini 1.5 have been commercialized. These models exhibit the following characteristics:

- Extended Context Window: Processing up to 1M tokens
- Multimodal Support: Integrated processing of text, images, and audio
- Enhanced Reasoning: Solving complex mathematical and logical problems
- Code Generation: Strengthened programming assistance

Notably, hallucination phenomena have significantly decreased. This enables more reliable answer generation through integration with RAG systems.

2. Advancement of RAG Systems

RAG (Retrieval-Augmented Generation) is an innovative architecture combining information retrieval and generation. Key developments in 2024:

2.1 Hybrid Search
Search accuracy has greatly improved by combining Dense and Sparse vector searches. Models like BGE-M3 and Cohere Embed support hybrid embeddings.

2.2 Chunking Strategies
Various strategies for effective document segmentation:
- Recursive Chunking: Maintains semantic units through recursive splitting
- Semantic Chunking: Meaning-based segmentation
- Late Chunking: Preserves context by splitting after embedding

2.3 Reranking
Cross-Encoder based rerankers reorder search results to improve final answer quality.

3. Vector Databases

Vector databases are the core infrastructure of RAG systems.

Qdrant:
- High-performance vector DB based on Rust
- Hybrid search support (Named Vectors)
- Enhanced filtering and payload search
- Clustering and sharding support

Pinecone:
- Fully managed cloud service
- Auto-scaling
- Metadata filtering

Weaviate:
- GraphQL API provision
- Multimodal vector search
- Built-in modular architecture

2024 AI Technology Trends Report

4. Real-World Applications

4.1 Customer Support Chatbots
Major e-commerce companies have adopted RAG-based chatbots to respond accurately and promptly to customer inquiries.

4.2 Legal Document Analysis
Law firms search vast case laws and legal documents with RAG systems, significantly improving lawyer efficiency.

4.3 Medical Information Systems
Healthcare institutions support doctors' diagnoses by integrating patient records, papers, and clinical guidelines.

4.4 Internal Knowledge Management
Companies integrate internal documents, wikis, and emails into RAG systems to improve information accessibility.

5. Performance Evaluation Methodology

Metrics for quantitatively evaluating RAG system performance:

Retrieval Performance Metrics:
- Precision@K: Proportion of relevant documents in top K results
- Recall@K: Proportion of retrieved among all relevant documents
- NDCG@K: Normalized Discounted Cumulative Gain considering ranking
- MRR: Mean Reciprocal Rank of first relevant document
- Hit Rate: Whether relevant documents were retrieved

Generation Quality Metrics:
- BLEU: N-gram similarity with reference answers
- ROUGE: Summary quality evaluation
- BERTScore: Semantic similarity
- Human Evaluation: Expert assessment

Recently, LLM-as-a-Judge methods using GPT-4 as evaluators are also widely used.

6. Conclusion and Outlook

AI technology in 2024 has entered the practical application phase. RAG systems have become core technologies capable of solving hallucination problems and providing accurate information.

Future Outlook:
- Multimodal RAG: Integrated search including images and videos
- Real-time Updates: Dynamic knowledge bases
- Personalization: Customized search and generation per user
- Multilingual Support: Information access without language barriers

The development of RAG technology will continue, expected to bring innovation to various industries.