Silent Intrusion

Enterprise

$33bn

Conversational AI market by 2029 (Pitchbook)

$80bn

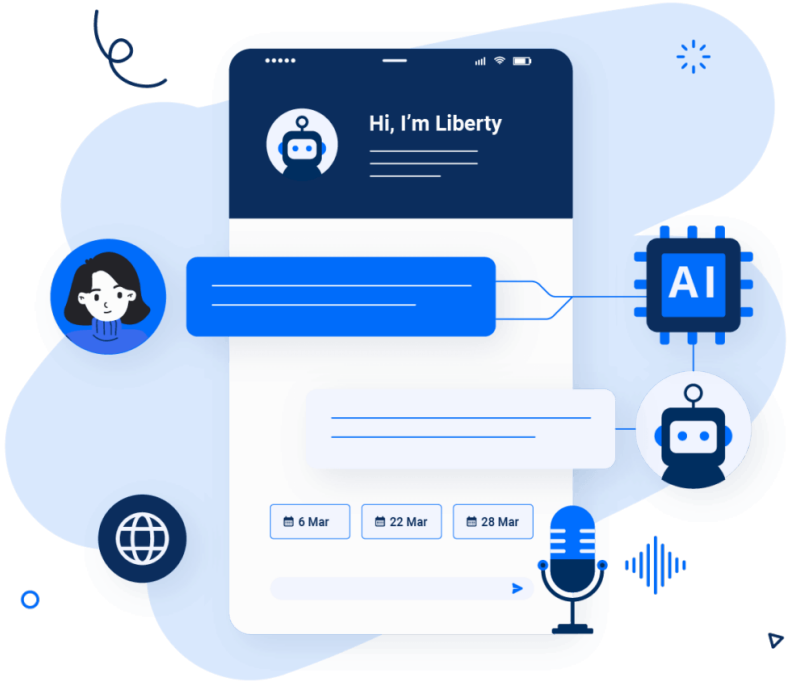
Contact center labor cost reduction driven by Conversation AI by 2026 (Gartner)

$454m

Chatbot revenue by 2026 (Statista)

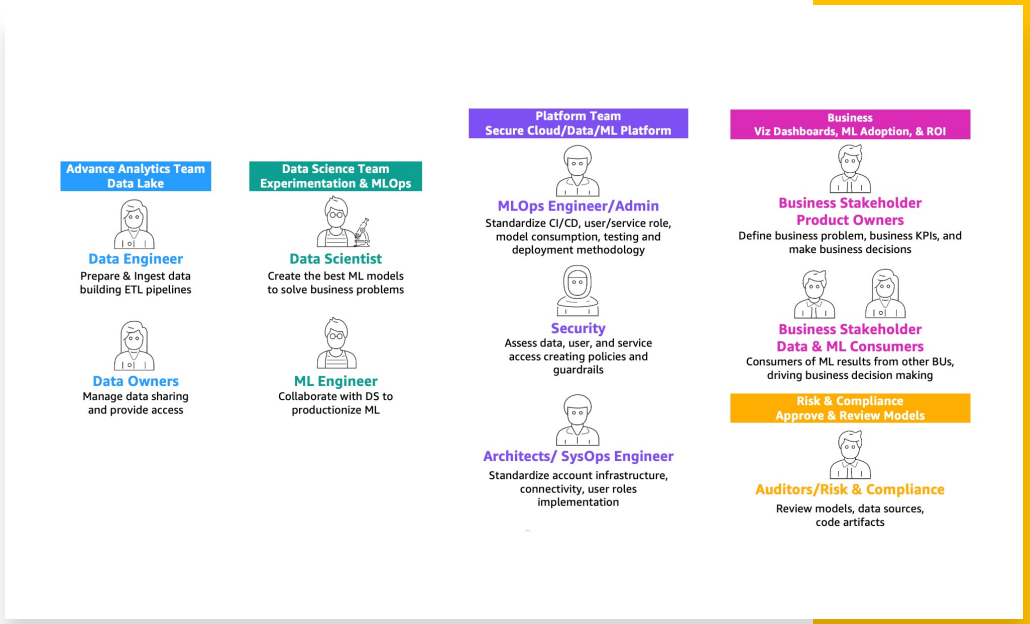
10%

Automated agent interactions By 2026 (Gartner)

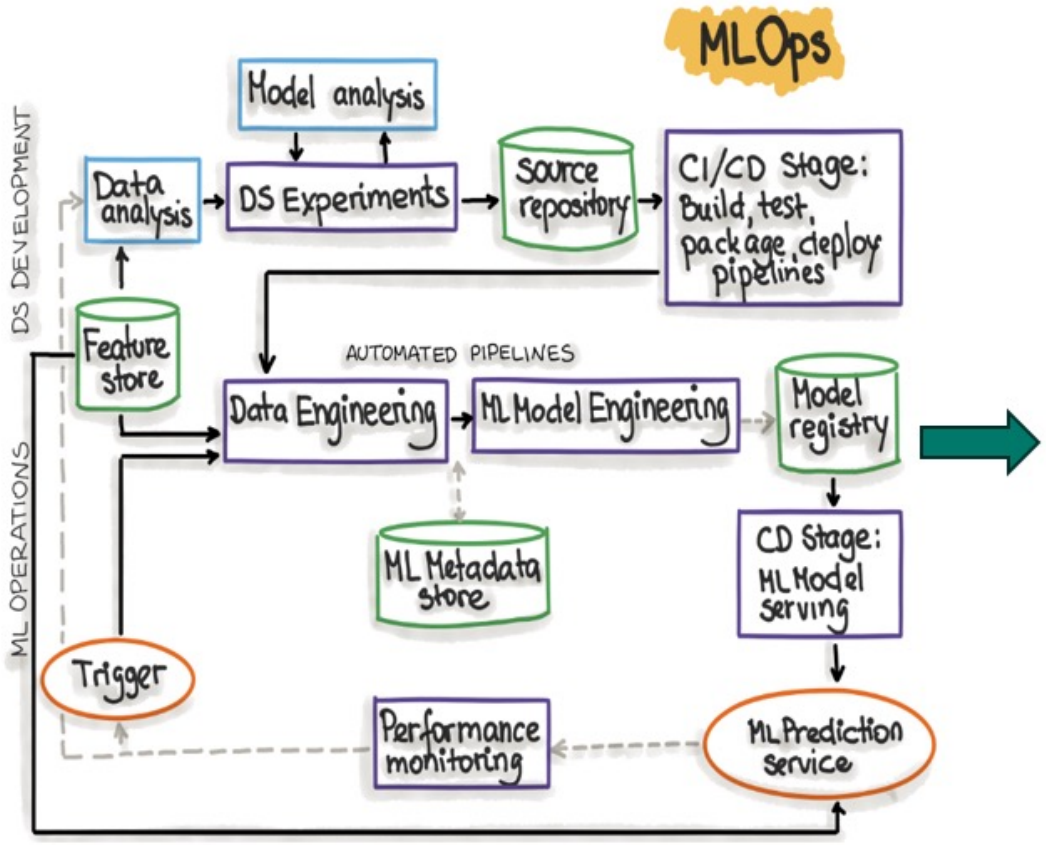


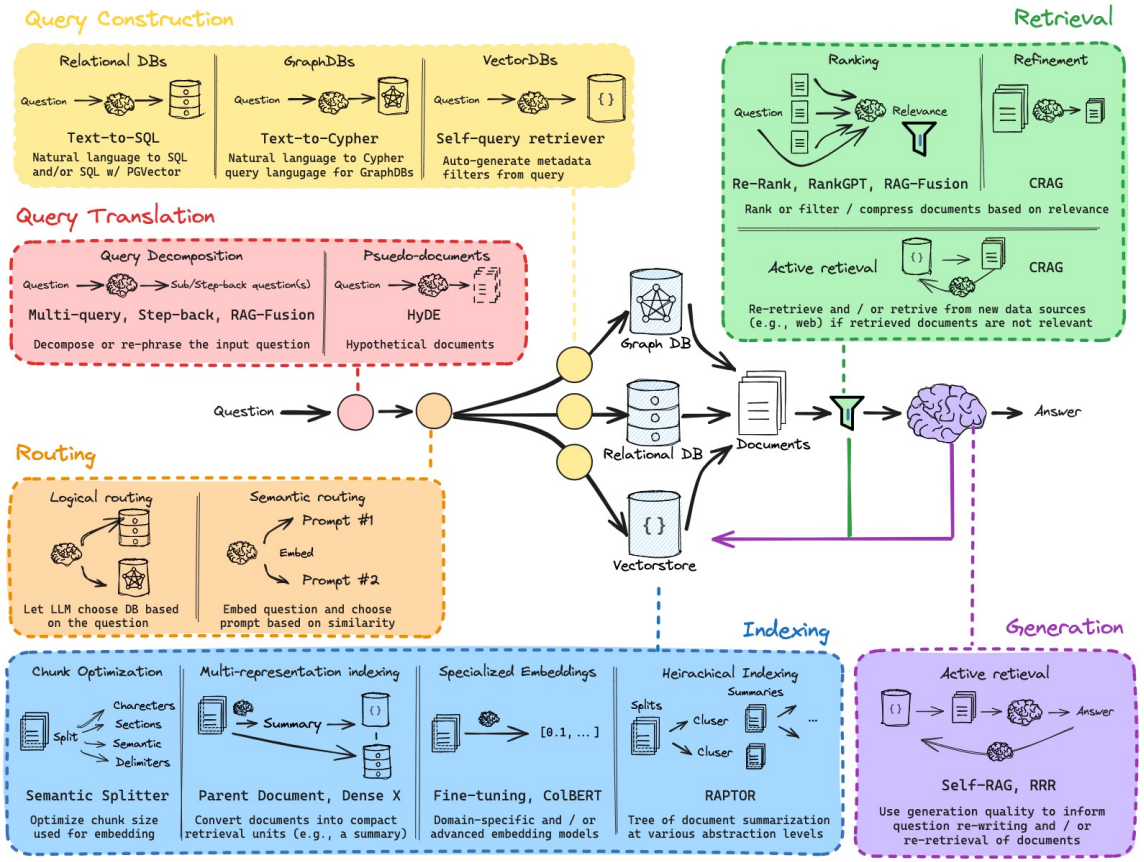
Transformation

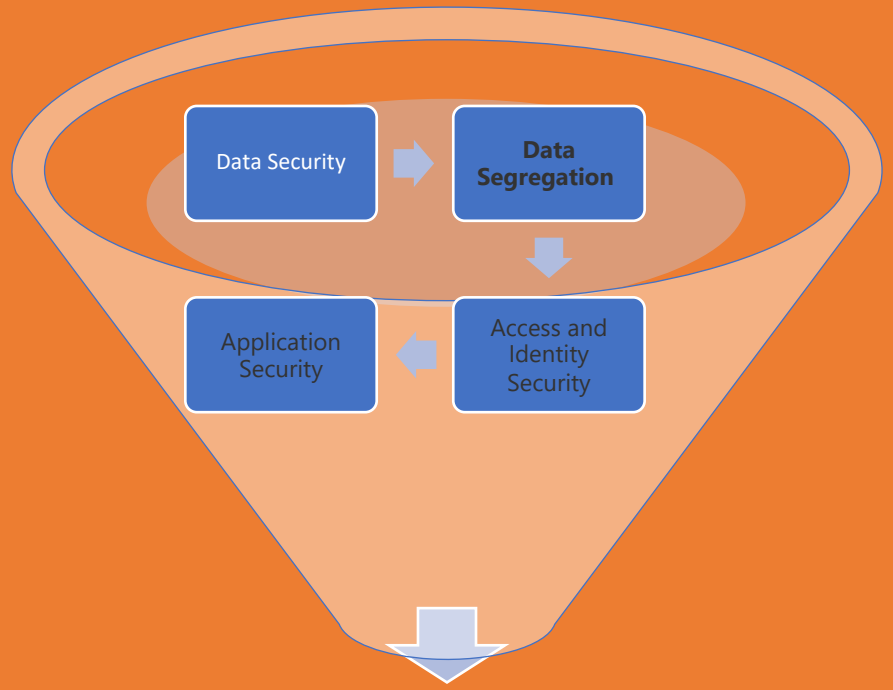
- Product Design
- B2B & B2C – Use Cases
- "Architecture as Code" (AasC)
- Stakeholder Framework
- HR Future Framework



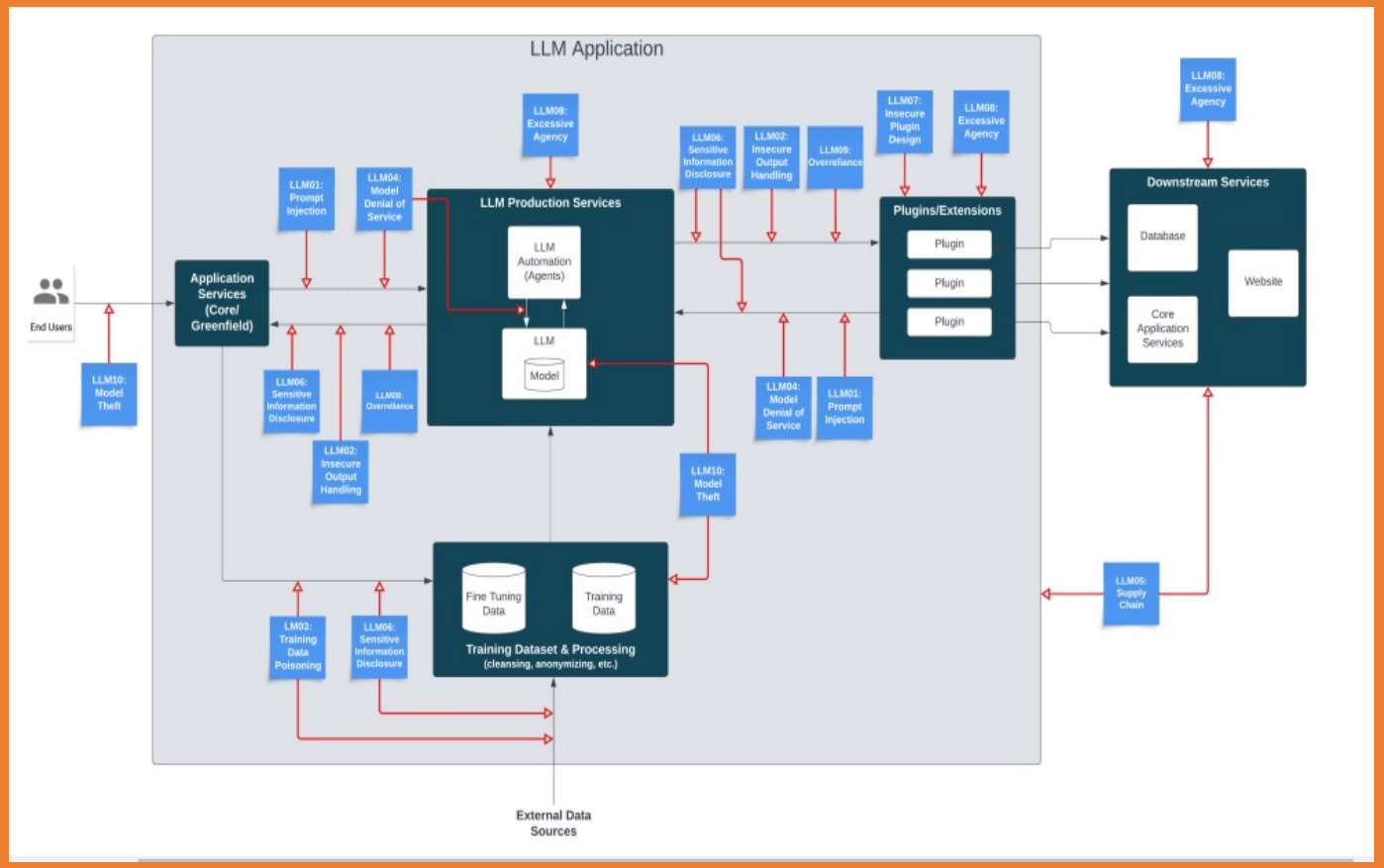
Are We Ready


MLOps
DS DEVELOPMENT
Model analysis
Data analysis
DS Experiments
Source repository
CI/CD Stage: Build, test, package, deploy pipelines
Feature store
AUTOMATED PIPELINES
Data Engineering
ML Model Engineering
Model registry
ML OPERATIONS
ML Metadata store
CD Stage: ML Model serving
Trigger
Performance monitoring
ML Prediction service


Query Construction
Relational DBs
Question
Text-to-SQL
Natural language to SQL and/or SQL w/ PGVector
GraphDBs
Text-to-Cypher
Natural language to Cypher query langugage for GraphDBs
VectorDBs
Self-query retriever
Auto-generate metadata filters from query
Retrieval
Ranking
Refinement
Relevance
Re-Rank, RankGPT, RAG-Fusion
CRAG
Rank or filter / compress documents based on relevance
Active retieval
CRAG
Re-retrieve and / or retrive from new data sources (e.g., web) if retrieved documents are not relevant
Query Translation
Query Decomposition
Sub/Step-back question(s)
Multi-query, Step-back, RAG-Fusion
Decompose or re-phrase the input question
Psuedo-documents
HyDE
Hypothetical documents
Graph DB
Relational DB
Vectorstore
Documents
Answer
Routing
Logical routing
Let LLM choose DB based on the question
Semantic routing
Prompt #1
Embed
Prompt #2
Embed question and choose prompt based on similarity
Indexing
Generation
Chunk Optimization
Split
Charecters
Sections
Semantic
Delimiters
Semantic Splitter
Optimize chunk size used for embedding
Multi-representation indexing
Summary
Parent Document, Dense X
Convert documents into compact retrieval units (e.g., a summary)
Specialized Embeddings
[0.1, ...]
Fine-tuning, ColBERT
Domain-specific and / or advanced embedding models
Heirachical Indexing
Splits
Summaries
cluser
cluser
RAPTOR
Tree of document summarization at various abstraction levels
Active retieval
Answer
Self-RAG, RRR
Use generation quality to inform question re-writing and / or re-retrieval of documents



Security



OWASP Top 10 for LLM Applications

Balance

- Open Source Vs Custom Inhouse
- Finance – Token and HR
- Use Cases Identification
- Do We need LLM
- Evolving Architecture

Hosting an LLM on your cloud

The default instance recommended by AWS is ml. p4de. 24xlarge, with a listed price of almost 33 USD per hour (on-demand). This means that such a deployment would cost at least 23,000 USD per month, assuming it doesn't scale up or down and that no discounts are applied. 28 Feb 2024

TensorOps
https://www.tensorops.ai › post › understanding-the-cos...

Understanding the cost of Large Language Models (LLMs)

CEO says he tried to hire an AI researcher from Meta and was told to 'come back to me when you have 10,000 H100 GPUs' It's only getting harder to hire workers with AI skills. A startup CEO said he couldn't poach a Meta employee because his company didn't have enough GPUs. 11 Mar 2024

Business Insider
https://www.businessinsider.com › News