Kore.ai is an AI company focused on delivering conversational and generative AI solutions to enterprises in a responsible and safe manner. Founded in 2014, Kore.ai has deep expertise in building enterprise-grade virtual assistants and AI-powered customer engagement solutions.

Kore.ai's goal is to simplify AI adoption for businesses by providing an end-to-end platform, no-code tools, and solutions to drive business outcomes with conversational and generative AI.

Key Offerings:

Kore.ai Platform:

The Kore.ai Platform consists of two main components:

1. **XO** - For building conversational AI assistants and experiences

XO is a no-code platform to create intelligent virtual assistants. It provides omni-channel orchestration to deliver consistent experiences across conversation channels like web, mobile, and voice assistants.

Key capabilities:
- Drag-and-drop bot builder with pre-built templates
- Contextual dialog management
- Omnichannel support across web, mobile, and voice
- Generative AI powered Co-pilot features for rapid development of virtual assistants
- Build human-like experiences with LLM Orchestration
- Built-in contact center capabilities
- Analytics for conversation insights
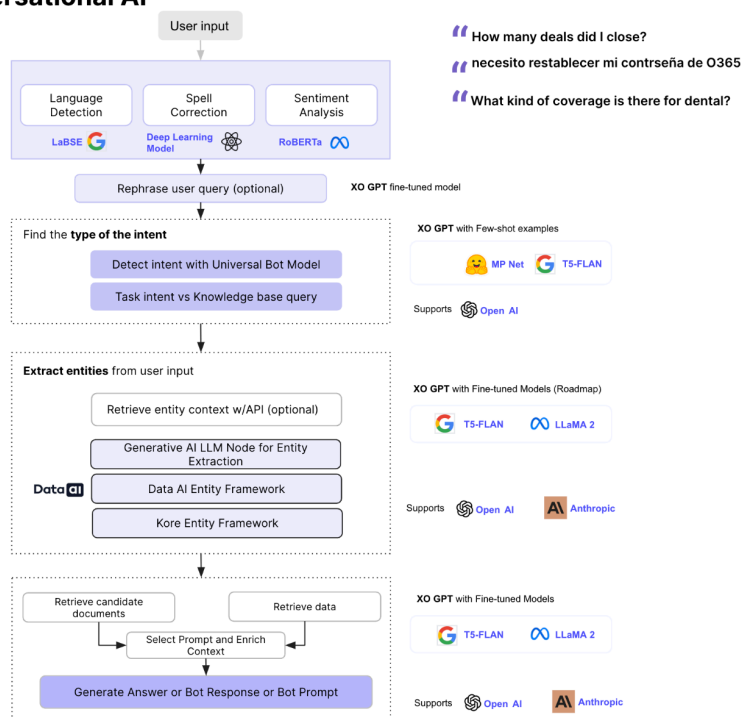- Integrations with business apps and data sources

Orchestrating multiple LLMs for conversational AI

User input

How many deals did I close?

necesito restablecer mi contrseña de O365

What kind of coverage is there for dental?

Language Detection

Spell Correction

Sentiment Analysis

Language Detection — LaBSE

Spell Correction — Deep Learning Model

Sentiment Analysis — RoBERTa

Rephrase user query (optional)

XO GPT fine-tuned model

Universal Bot – Bot Identification

Intent Recognition

Task vs intent-less query

Find the **type of the intent**

Detect intent with Universal Bot Model

Task intent vs Knowledge base query

XO GPT with Few-shot examples

MP Net   T5-FLAN

Supports Open AI

Entity Extraction

Data AI for Entity Extraction and API Integrations

Extract entities from user input

Retrieve entity context w/API (optional)

Generative AI LLM Node for Entity Extraction

Data AI Entity Framework

Kore Entity Framework

Data AI

XO GPT with Fine-tuned Models (Roadmap)

T5-FLAN   LLaMA 2

Supports Open AI   Anthropic

Answer generation

Response generation

Conversational response

Retrieve candidate documents   Retrieve data

Select Prompt and Enrich Context

Generate Answer or Bot Response or Bot Prompt

XO GPT with Fine-tuned Models

T5-FLAN   LLaMA 2

Supports Open AI   Anthropic

Kore.ai's orchestration layers enable customizable and optimized conversational experiences by integrating multiple AI models:

- Smaller, efficient models handle tasks like sentiment analysis and language detection.
- Custom models equipped with Few-shot intent recognition are optimized for managing conversational AI flows.
- Generative AI nodes and dialogue flows are used to integrate and chain calls from various Large Language Models (LLM).
- Built-in guardrails redact personally identifiable information (PII), validate against bias, toxicity and hallucinations.
- Custom fine-tuned models are optimized for query rewriting and response generation in conversational AI contexts.
- Retrieval augmentation from Kore's XO Answers boosts accuracy by ingesting enterprise knowledge (web pages, documents, tabular data such as catalog, support tickets etc.) into a semantic index to generate responses using retrieval augmented generation.
- Multilingual models support global use cases.

With Kore.ai, you control:

- Business logic to align experiences with goals
- Accuracy via custom models tuned for specific tasks

- The overall experience by orchestrating AI models for optimal cost, latency and accuracy

Kore's analytics and insights tools deliver transparency into conversational AI performance, including:
- Self-service usage showing successful automated query resolution
- Conversation flow analytics revealing drop-off points or transfers to human agents
- Bot performance metrics identifying areas for improvement
- Customizable dashboards that allow tagging of user demographics to reveal any biases

---

The *Search and Answers* module of XO Platform enables creating intelligent virtual assistants that can provide accurate responses to customer queries by retrieving and summarizing the most relevant information from your knowledge sources.
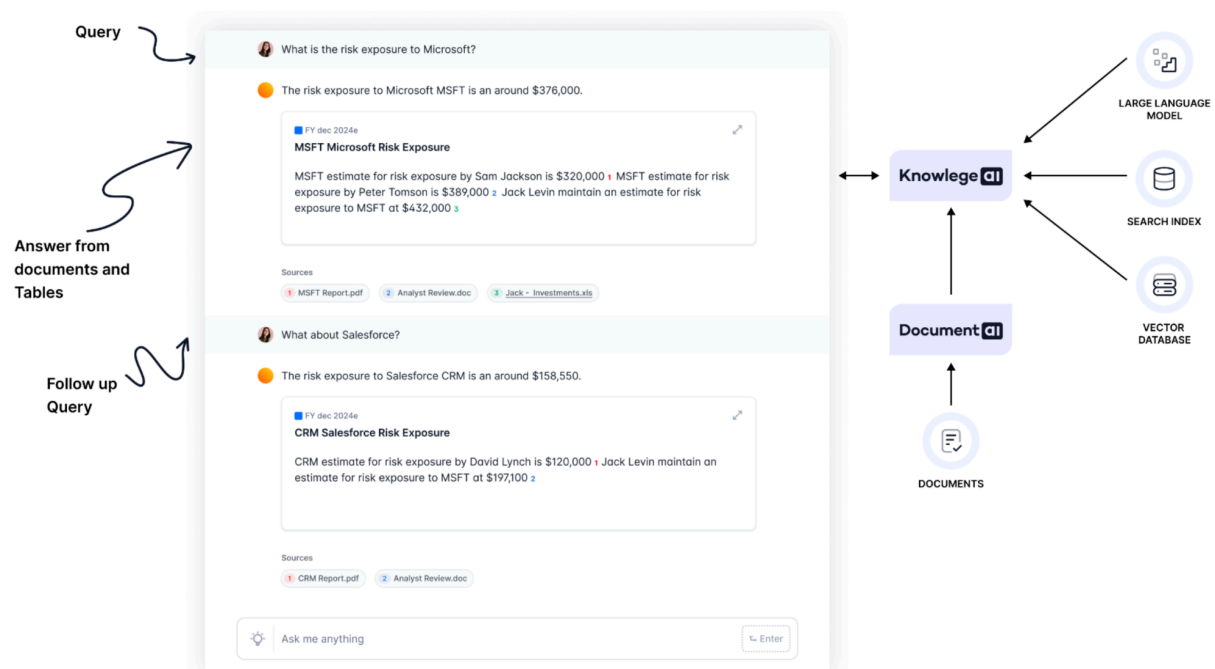
Key features include:
- Vector similarity search - User questions are encoded into embeddings and compared against indexed documents and FAQs to find semantic matches. Relevant snippets are extracted and ranked.

- Multi-source knowledge ingestion - Upload documents, websites, product catalogs, FAQs. Connect to databases, CMS systems like SharePoint and Zendesk to index both structured and unstructured data.
- Document AI - Understands complex PDF layouts and tables to extract structured data for indexing.
- Retrieval augmented generation - Relevant snippets are summarized and combined using generative AI to produce an answer grounded in evidence.
- Contextual answers - Follow-up questions are understood in context to refine answers iteratively.
- Insights - Conversation analytics reveal common questions and gaps.
- Relevance tuning - Fine-tune vector search and re-rankers using feedback.
- Customization - Build tailored query pipelines, UI widgets and journeys.

With Search and Answers, easily ingest both structured and unstructured knowledge into a unified index. Customers get accurate, consistent answers across channels by enhancing your FAQs with website content, documents, catalogs and databases. The module combines the power of vector similarity search with generative AI to produce complete answers with relevant context.

Continuously improve answer relevance using conversational analytics and query feedback. Search and Answers reduces customer effort by providing fast, accurate self-service powered by your knowledge universe.

XO enables enterprises to rapidly build and deploy AI-powered customer service, employee support, and personalized engagement use cases. The no-code tools allow business users to create and continuously improve experiences without dependency on IT teams.

---

2. GALE - For building generative AI applications

GALE enables enterprises to build custom AI applications leveraging the power of large language models (LLMs) like GPT-4, Anthropic's Claude, and community models like FLAN-T5, LLaMA and Mistral. It provides the tools and infrastructure to develop generative AI features and flows tailored to your business goals and data universe.

With GALE, you can:

- Craft optimized prompts using the Prompt Engineering Studio while collaborating with your enterprise teams

- Evaluate and monitor performance of the prompts
- Access leading LLMs like GPT-3, Claude, FLAN-T5, LLaMA through a model marketplace with one-click deployment
- Orchestrate workflows chaining multiple LLMs to mimic complex human reasoning
- Integrate LLMs with internal and external APIs, databases, microservices through code integration
- Ingest enterprise documents, data into LLMs for fine-tuning using reinforcement learning
- Build custom models optimized for specific tasks like sentiment analysis, topic tagging, summarization
- Continuously improve models through retraining on new data
- Test prompts extensively to avoid hallucinations
- Monitor for bias, toxicity using Responsible AI guardrails
- Deploy fine-tuned models via API for integration into apps
- Analyze model analytics through BI dashboards

GALE enables collaborative development of AI solutions spanning business SMEs, data scientists, and developers. With its end-to-end platform, you can rapidly build AI applications like:
- Analyze emails, extract information and take intelligent decisions
- Generate conversation summarization of customer and agent transcripts
- AI-powered search and recommendations
- Sentiment analysis on customer surveys
- Automated report generation from data
- Intelligent campaign generators
- Insights from unstructured text using topic modeling
- Personalized product configurators

GALE unlocks the power of LLMs for the enterprise - making it easier to leverage these models safely and effectively to solve business challenges. With GALE, you control the models and data - rather than relying on generic LLM APIs.

You get the best of human-like LLMs and customized enterprise logic tailored to your needs. GALE provides the tools to build robust, reliable generative AI applications aligned to your business goals.

The Kore.ai Platform is cloud and LLM agnostic, offering flexibility in deployment and choice of AI providers. It ensures scalability, security, and compliance for enterprise rollout. Pre-built integrations provide out-of-the-box connectivity with 100+ enterprise apps.

Assist Solutions for Conversational AI:

Kore.ai provides domain-specific solutions built on the XO platform to accelerate time-to-value for conversational AI projects.

Key solutions for customer experience:
- BankAssist - Virtual assistant for banking
- HealthAssist - Virtual assistant for healthcare
- RetailAssist - Conversational commerce and customer support
- TravelAssist - Virtual assistant for travel and hospitality

These solutions come pre-loaded with industry knowledge, use case templates, and integrations tailored for the vertical. Enterprises can customize them further using XO's no-code bot builder.

Key solutions for employee experience:
- HRAssist - Automates HR enquiries and case management
- ITAssist - Handles IT help desk tickets
- ProcureAssist - Procurement assistant for managing purchases
- SalesAssist - Engage with prospects and enhance conversions. Supports sales reps with finding answers, and automating tasks.

The Assist Solutions use pre-trained models that provide higher accuracy out-of-the-box for industry-specific use cases. They reduce deployment time from months to weeks.

Eva:

Eva is an AI-powered Enterprise virtual assistant that not only provides natural language answers by searching the user's data, but also takes smart actions and generates insights tailored to each user.

Key features include:
- Answering natural language questions by searching relevant sources like emails, documents, transcripts and returning the most relevant snippets
- Smart actions based on answers, like sharing insights via email, Slack, Teams etc.

- Generating custom charts, dashboards and visualizations based on answers from connected business systems
- Connecting to business systems like Salesforce, Jira etc. to generate visualizations from natural language questions
- Finding answers to questions from any tabular data through SQL queries
- Scanning PDFs, images, and semi-structured data to extract answers
- Having natural conversations and refining responses based on follow-up questions and context
- Building user profiles based on usage to provide personalized, context-aware responses
- Allowing easy integration with custom systems to connect them to Eva

The key innovation is using AI to interpret natural language questions, transform them into API calls, SQL queries or other backend actions, execute these actions on the user's data sources, and present the most relevant insights through generated text, visualizations and smart actions - all tailored to each user's unique data universe.

Eva goes beyond a conversational search engine, acting as an AI-powered personal analyst that not only finds answers but also takes data-driven actions and generates custom analytics views.

Kore.ai is committed to providing tools, services and solutions that allow companies to realize business value from AI while mitigating risks. With a full stack offering across conversational and generative AI, Kore.ai aims to simplify and accelerate AI adoption for enterprises globally.

Does XO platform run on GALE or closed source LLMs such as Open AI, BARD, MP Net, FLAN-T5 etc?

The XO platform leverages both open source large language models (LLMs) like GPT-3.5 and GPT-4, as well as proprietary fine-tuned models that are customized and optimized specifically for conversational AI capabilities.
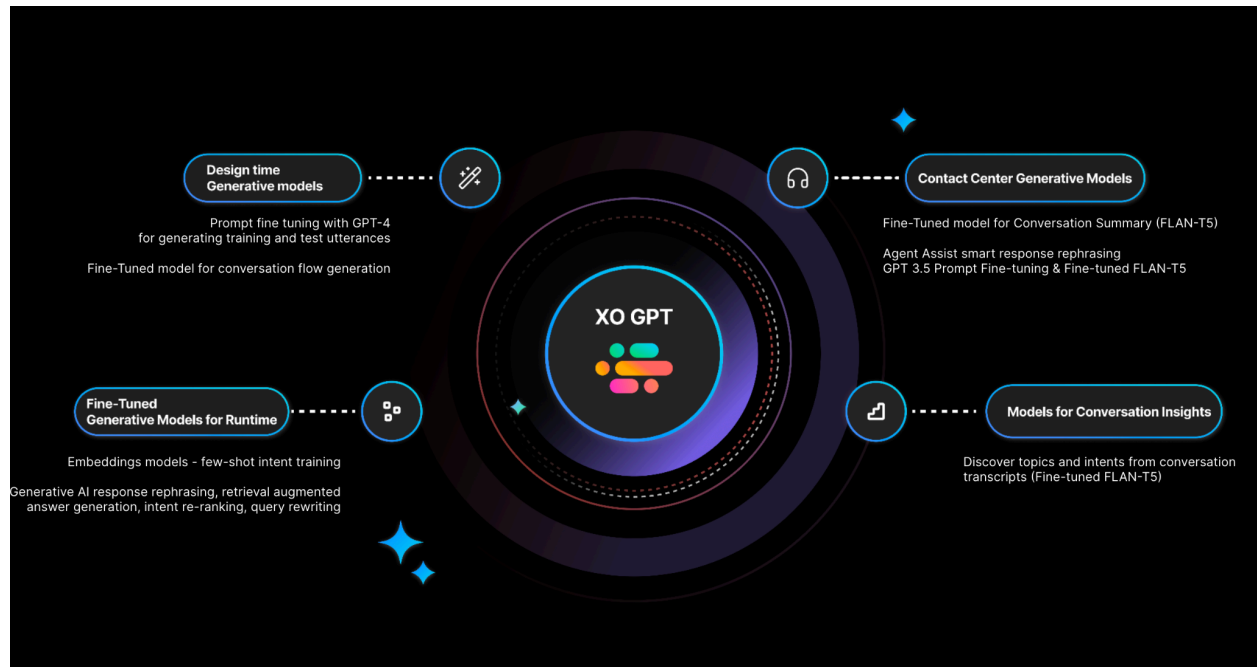
During conversational AI bot design time, XO makes use of advanced generative LLMs including GPT-4 and Anthropic's Claude. These models are used for prompt fine-tuning to generate training utterances, test utterances, and full conversation flows that define the bot logic. Prompt fine-tuning involves providing the model with a few examples of desired outputs like training phrases or conversation flows. This allows the model to learn the pattern and generate additional high-quality outputs. Using powerful models like GPT-4 results in more natural, human-like training data and conversation flows compared to what a human could manually create. However, prompt programming requires extensive iterative testing and refinement to generate optimized outcomes. The XO platform streamlines this with the prompt fine-tuning process - that involves extensive optimized prompt templates, chain of thought instructions, optimized few-shot examples and iterative testing.

At runtime when the bot is conversing with users, XO leverages additional proprietary fine-tuned NLP models for key tasks like intent recognition, entity extraction, and response generation. For intent recognition, XO uses embeddings from open source models like MPNet, FLAN-T5, and Open AI's Ada to represent the user's input text. This text embedding is then fed into a fine-tuned ranking model that selects the most likely matching intent for the user's query from a defined set of possible intents. The fine-tuned model is specialized to distinguish between intents and purposes in complex enterprise scenarios. For entity extraction, XO trains custom models like a food ordering entity extractor to adhere to specialized business rules that generic models struggle with.

Response generation is handled by fine-tuned models based on open source LLM FLAN-T5, and commercial models like Open AI's GPT-3.5 , GPT-4 and Anthropic's Claude. Fine-tuned models are customized to generate natural conversational responses from templates, while correctly incorporating context, co-referencing, empathy, and other nuances. Fine-tuning significantly improves response relevance, reduces hallucination, and provides better control compared to general purpose LLMs.

XO also employs fine-tuned retrieval augmented generation models that can provide answers by searching over a provided context document. This avoids hallucination by grounding responses in given evidence.

Additional proprietary runtime models include user query rewriting for co-referencing, topic modeling for conversation analytics, and summarization for agent handover use cases. Each model is fine-tuned on representative datasets to produce satisfactory performance on live customer conversations.

Customers can optimize the fine-tuned models or build custom fine-tuned models using Kore.ai's GALE platform using their enterprise data. This allows adapting large language models like FLAN-T5, LLaMA 2 to understand company terminology, business rules and user interactions. Fine-tuned models are integrated into XO's conversational platform via APIs. For the contact center, GALE models power assistants that summarize interactions, generate empathetic agent responses, and provide coaching insights.

By orchestrating fine-tuning on enterprise data with deployment through GALE APIs, enterprises can leverage LLMs for contextual assistants without exposing customer data during development. GALE training followed by integration with XO enables capable, customized conversational AI tailored to each industry.

By combining state-of-the-art LLMs like GPT-4 with specialized fine-tuned models for each subtask, the XO platform provides a comprehensive conversational AI solution tailored to complex enterprise needs. The design time models enable easy bot authoring through prompt programming while the runtime models ensure robust, controlled performance during live interactions.

Some key advantages of the fine-tuned proprietary models include:
- Optimized for specific conversational AI subtasks like intent recognition and response generation
- Learn to adhere to business rules and ontology constraints
- Control against hallucination and over consistency compared to general LLMs
- Deployed on smaller, efficient models which improves latency and reduces costs
- Developed with representative enterprise data and scenarios for better real-world performance

- Continuously improved through accumulation of conversation logs and feedback

In summary, the XO platform taps into the best of open source AI through LLMs like GPT-4 and Claude, as well as customized proprietary models fine-tuned for enterprise-scale conversational AI across the complete lifecycle from design to runtime. Combining these technologies provides the accuracy, naturalness, robustness and precision needed for real-world deployment. The fine-tuned models in particular allow adapting the leading LLMs to work reliably on specialized enterprise use cases with complex requirements.