Business AI

2025

This case example presents a realistic scenario created by ATR Soft, based on our extensive knowledge of industrial support needs. While the situation described is imaginary and not an actual client case, the analysis has been conducted as if it were a real case.

Analysis: Utilizing AI to help with support

Problem

Description

Our support receives thousands of support tickets yearly. While the response times are good, they could of course be always better. When support personnel is answering the tickets, they have to search answers from multiple documentation sources, both public and internal. The data sources include: several documentation sources (web page, confluence, pdfs and other files in internal infra), internal databases, code, discussions from previous support tickets, as well as support chats with customers. The amount of data is not huge, but, it being spread in multiple places makes manually searching it really time consuming.

Roadmap

Initial goal is to make work of support personnel easier, but, if the solution proves to be good, we might also consider integrating it to our existing support solution, to work as a AI chatbot when support personnel is unavailable.

Prioritization

If possible, automatically finding duplicate support tickets would yield the most savings.

Budget

in best case scenario, the solution would make it possible not to hire more support personnel. Thus, some monthly costs would be acceptable.

Limitations

The code, and discussions inside the ticket system are considered to be secrets, and cannot be sent to any cloud. In addition, our existing, custom built customer support system should integrate to this solution, not be replaced with it.

Analysis

Overview

There are many existing (AI) tools that are designed for this problem: Kustomer, Forethought, Guru, Zendesk, Solvvy, Capacity, Talla, ... While some of these can be run on-premises, they would be replacements for the existing support solution, so not applicable here. Fortunately, in addition to these, there are multiple general purpose semantic search systems (Elasticsearch, Microsoft Azure AI search), that can be integrated with various data sources, and which would solve the search part of the problem. However, chat-like features require additional LLMs, and semantic search over code is in itself a difficult problem.

Categorization

Difficult. While partially solvable with existing tools, the on premise -requirement means that some development work is always needed. Full solution requires both software development, custom integrations, maintenance and some AI development. Analysis will be split to 2 scenarios based on which is considered more important, scalability to large datasets/amounts of users, or semantic search quality. The cheapest solution might actually be the third, to replace the existing custom-built tool with one that already has the functionality.

EU AI Act risk category

The resulting AI system would be designed to help support staff find information in tickets and documentation using semantic search. The closest possible high risk AI system area would be Employment, workers management and access to self-employment (EU AI Act, Annex III). However, the resulting system would not make or automate decisions about employment terms, promotions, or task allocation, nor does it monitor or evaluate employee performance. Its function is to assist with information retrieval, which is a narrow, procedural, and preparatory task that supports human work without materially influencing outcomes.

When an AI system is performing such tasks, it is not considered high risk even if the use case area could be seen as one of those mentioned in Annex III (EU AI Act, Chapter III, Section 1, Article 6). Because this system only improves the efficiency of a previously completed human activity and does not replace or influence human assessment, it falls under the limited risk category.

As a result, the system is not subject to the strict requirements for high-risk AI systems. Instead, general transparency obligations apply, such as informing users when they interact with AI if an AI chat system is implemented (EU AI Act, Chapter IV, Article 50). No additional compliance steps specific to high-risk systems are required.

Case 1: scalability & cost efficiency is more important

Elastic provides self-hosted deployments via Elastic Stack. Elasticsearch is widely used for creating custom (potentially AI-enhanced) search engines that can integrate data from various internal sources like knowledge bases, CRMs, and helpdesks. While originally a highly scalable full-text search engine instead of a semantic search (or vector search) engine, it can be extended to handle also vector searches as well.

Solution

On premises Elastic search (+ local open source LLM, if semantic search is really needed). Elastic is highly customizable and can be configured to index data from various sources, such as PDFs, web pages, Confluence, internal databases, and support ticket systems. Chat solution will need to be custom built.

Pros:

- scales well
- large community
- inexpensive
- allows also keyword & hybrid search

Cons:

- semantic search more difficult

POC: how well would elastic's full text search help with finding similar tickets?
workload estimate: 5+months
POC workload estimate: 3 weeks

Case 2: semantic search quality is more important

Finding duplicate tickets, using chat-like interface to run searches and searching answers from codebases can all be achieved with LLMs and vector databases. Without them, these parts of the problem are difficult. There are tools that help building solutions like this, but a lot of development is still needed.

Solution

Open source LLM running on premises. Vector database running on premises. LLM used to calculate embedding vectors, for all different data sources. Chat-like interface, natural language questions transformed to vectors using LLM. Vector database used to retrieve contextually relevant documents/duplicate tickets.

Pros:

- llamaindex offers out of the box semantic search & LLM support
- solves all parts of the problem
- extendable
- possibility to easily achieve natural language responses

Cons:

- costly to develop
- costly to maintain, requires dedicated hardware
- doesn't scale well to big numbers of users (LLMs are costly)

POC: semantic search over support tickets
suggested tech: llamaindex, ollama, codellama
requirements: dedicated infra for running the (small) LLM locally
workload estimate: 6+months
POC workload estimate: 4 weeks

Get your AI to the next level

Business AI Service is designed for large industrial companies.

This is the right service for you if:

✓ You are a large industrial company.
✓ You have AI ideas but are unsure about feasibility or impact.
✓ You want to prioritize AI projects that deliver real business value.
✓ You want to test ideas quickly before full-scale investment.
✓ You want guidance on using existing tools vs. building custom solutions.
✓ You want to reduce risk and maximize ROI from AI initiatives.

Schedule with us a quick 15-minute intro call:

Book a meeting

Mastering industrial complexity may be hard,

but working with us isn't.

Business AI Service
ATR Soft Oy
atrsoft@atrsoft.com
Joukaisenkatu 1
20520 Turku
Finland