# Resource-Efficient Domain-Specific AI Assistant

## Overview

*Large Language Models (LLMs) showcase exceptional performance across diverse applications. Nevertheless they encounter substantial challenges when operationalized on systems with restricted computational resources, such as single-GPU configurations. While their extensive knowledge base is versatile, it becomes resourse-inefficient when applied to specialized domains. Additionally, utilizing existing cloud solutions from major players raises serious data privacy concerns.*


## Keywords

**LLM (large language models)**
**language generation**
**AI assistant**
**chatbot**
**resource-efficient LLM**
**local LLM**
**domain adaptation**
**natural language processing**
**domain-specific knowledge**
**vector database**

## Situation

CVUT directives serve as official governing documents that necessitate strict adherence from students. However, their formal language and large scale make them difficult to comprehend for the end users. A localized, domain-specific AI assistant that can interpret and navigate through CVUT directives would be invaluable in assisting students to effortlessly find pertinent information within these extensive documents.

## Solution

To address this, an open-source LLM was optimized using various techniques to enable its operation on single GPU systems. Subsequently, a vector database was constructed, incorporating the wealth of information found within CVUT directives. This model served as a dedicated, localized assistant, specifically tailored to assist in navigating and interpreting CVUT directives efficiently, while addressing the challenges of resource constraints and data privacy.

## Requirements

A resource-efficient AI assistant capable of operating locally.

Domain-specific adaptation to understand and follow CVUT directives.

Quick adjustability to incorporate changes in directives.

## Benefits and Results

- Domain Specificity: The assistant, being specialized, offers precise and relevant information from CVUT directives, enhancing accessibility and understanding for the students.
- Resource Efficiency: Optimization allows the model to run on local, single-GPU setups, reducing the need for extensive computational resources.
- Privacy Preservation: Operating locally, the assistant safeguards user data, mitigating the risks associated with cloud-based solutions.
- Quick Adaptability: The model can swiftly incorporate changes in directives, ensuring the provision of up-to-date and accurate information.
- Enhanced Accessibility: Students can effortlessly retrieve essential information from the voluminous directives, making official documents more approachable and understandable.

**This approach not only simplifies the interaction with complex official documents but also sets a precedent for developing localized, domain-specific AI solutions that are resource-efficient and privacy-preserving, addressing the inherent challenges posed by generalized large language models.**