
# Diplomarbeit

## Artificial Intelligence in the Industry and Education Environment

SUBTITITLE

Eingereicht von

**Gabriel Mrkonja**
**Florian Prandstetter**
**Luna Schätzle**

Eingereicht bei

**Höhere Technische Bundeslehr- und Versuchsanstalt Anichstraße**

Abteilung für Wirtschaftsingenieure/Betriebsinformatik

Betreuer

Greinöcker
Egger

Projektpartner

HTL Anichstraße

Innsbruck, April 2025

Abgabevermerk:

Betreuer/in:

Datum:

Gabriel Mrkonja
Florian Prandstetter
Luna Schätzle

# SPERRVERMERK

Auf Wunsch der Firma

HTL Anichstraße

ist die vorliegende Diplomarbeit
für die Dauer von drei / fünf / sieben Jahren
für die öffentliche Nutzung zu sperren.
Veröffentlichung, Vervielfältigung und Einsichtnahme sind ohne
ausdrückliche Genehmigung der Firma *** und der Verfasser
bis zum TT.MM.JJJJ nicht gestattet.

Innsbruck, TT.MM.JJJJ

Verfasser:

Vor- und Zuname Unterschrift

Vor- und Zuname Unterschrift

Firma: Firmenstempel



Gabriel Mrkonja
Florian Prandstetter
Luna Schätzle

# Kurzfassung /Abstract

Eine Kurzfassung ist in deutscher sowie ein Abstract in englischer Sprache mit je maximal einer A4-Seite zu erstellen. Die Beschreibung sollte wesentliche Aspekte des Projektes in technischer Hinsicht beschreiben. Die Zielgruppe der Kurzbeschreibung sind auch Nicht-Techniker! Viele Leser lesen oft nur diese Seite.

Beispiel für ein Abstract (DE und EN)

Die vorliegende Diplomarbeit beschäftigt sich mit verschiedenen Fragen des Lernens Erwachsener – mit dem Ziel, Lernkulturen zu beschreiben, die die Umsetzung des Konzeptes des Lebensbegleitenden Lernens (LBL) unterstützen. Die Lernfähigkeit Erwachsener und die unterschiedlichen Motive, die Erwachsene zum Lernen veranlassen, bilden den Ausgangspunkt dieser Arbeit. Die anschließende Auseinandersetzung mit Selbstgesteuertem Lernen, sowie den daraus resultierenden neuen Rollenzuschreibungen und Aufgaben, die sich bei dieser Form des Lernens für Lernende, Lehrende und Institutionen der Erwachsenenbildung ergeben, soll eine erste Möglichkeit aufzeigen, die zur Umsetzung dieses Konzeptes des LBL beiträgt. Darüber hinaus wird im Zusammenhang mit selbstgesteuerten Lernprozessen Erwachsener die Rolle der Informations- und Kommunikationstechnologien im Rahmen des LBL näher erläutert, denn die Eröffnung neuer Wege zur orts- und zeitunabhängiger Kommunikation und Kooperation der Lernenden untereinander sowie zwischen Lernenden und Lernberatern gewinnt immer mehr an Bedeutung. Abschließend wird das Thema der Sichtbarmachung, Bewertung und Anerkennung des informellen und nicht-formalen Lernens aufgegriffen und deren Beitrag zum LBL erörtert. Diese Arbeit soll

einerseits einen Beitrag zur besseren Verbreitung der verschiedenen Lernkulturen leisten und andererseits einen Reflexionsprozess bei Erwachsenen, die sich lebensbegleitend weiterbilden, in Gang setzen und sie somit dabei unterstützen, eine für sie geeignete Lernkultur zu finden.

This thesis deals with the various questions concerning learning for adults – with the aim to describe learning cultures which support the concept of live-long learning (LLL). The learning ability of adults and the various motives which lead to adults learning are the starting point of this thesis. The following analysis on self-directed learning as well as the resulting new attribution of roles and tasks which arise for learners, trainers and institutions in adult education, shall demonstrate first possibilities to contribute to the implementation of the concept of LLL. In addition, the role of information and communication technologies in the framework of LLL will be closer described in context of self-directed learning processes of adults as the opening of new forms of communication and co-operation independent of location and time between learners as well as between learners and tutors gains more importance. Finally the topic of visualisation, validation and recognition of informal and non-formal learning and their contribution to LLL is discussed.

Gliederung des Abstract in **Thema**, **Ausgangspunk**, **Kurzbeschreibung**, **Zielsetzung**.

**Projektergebnis** Allgemeine Beschreibung, was vom Projektziel umgesetzt wurde, in einigen kurzen Sätzen. Optional Hinweise auf Erweiterungen. Gut machen sich in diesem Kapitel auch Bilder vom Gerät (HW) bzw. Screenshots (SW). Liste aller im Pflichtenheft aufgeführten Anforderungen, die nur teilweise oder gar nicht umgesetzt wurden (mit Begründungen).

# Erklärung der Eigenständigkeit der Arbeit

**EIDESSTATTLICHE ERKLÄRUNG**

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche erkenntlich gemacht habe. Meine Arbeit darf öffentlich zugänglich gemacht werden, wenn kein Sperrvermerk vorliegt.

Ort, Datum Verfasser 1

Ort, Datum Verfasser 1

# Inhaltsverzeichnis

# Teil I

# Introduction

# 1 Einleitung

In der Einleitung wird erklärt, wieso man sich für dieses Thema entschieden hat. (Zielsetzung und Aufgabenstellung des Gesamtprojekts, fachliches und wirtschaftliches Umfeld)

## 1.1 Vertiefende Aufgabenstellung

### 1.1.1 Schüler*innen Name 1

### 1.1.2 Schüler*innen Name 2

## 1.2 Dokumentation der Arbeit

Es werden die Projektergebnisse dokumentiert

- Grundkonzept
- Theoretische Grundlagen
- Praktische Umsetzung
- Lösungsweg
- Alternativer Lösungsweg
- Ergebnisse inkl. Interpretation

Weitere Anregungen:

- Fertigungsunterlagen
- Testfälle (Messergebnisse...)
- Benutzerdokumentation
- Verwendete Technologien und Entwicklungswerkzeuge

# 2 Introduction: AI in the Industry and Education Environment

# Teil II

# Hardware

# 3 Raspberry PI

# 4 Server

# Teil III

# Theoretical background

# 5 Used Technologies

## 5.1 Introduction

## 5.2 Visual Studio Code

## 5.3 Vue.js

## 5.4 firebase

### 5.4.1 Securety

## 5.5 Github

## 5.6 Docker

## 5.7 VPN Tunnel (TailScale)

# 6 Operating Systems used

# 7 Used Programming Languages

## 7.1 Python

**Gabriels Part**

## 7.2 HTML, CSS, and JavaScript in Combination with Vue.js

In this project, the languages HTML, CSS, and JavaScript were used in conjunction with Vue.js to create an interactive and dynamic user experience. For more information about Vue.js, see Chapter **??**, Section "Vue.js."

### 7.2.1 HyperText Markup Language (HTML)

HyperText Markup Language (HTML) is the standard markup language for creating and structuring content on the web. It serves as the backbone of web pages by organizing content through elements represented by tags. Key features of HTML include:

- **Structure Definition:** Tags such as `<html>`, `<head>`, and `<body>` define the structural hierarchy of a web page.
- **Content Organization:** Elements like headings, paragraphs, links, images, and tables provide a clear and user-friendly layout.
- **Web Compatibility:** HTML is universally supported, ensuring seamless integration across browsers and devices.

As one of the core technologies of the World Wide Web, alongside CSS and JavaScript, HTML enables the creation of interactive and visually appealing websites. Its simplicity and adaptability make it an essential tool for web development.

**?**

### 7.2.2 Cascading Style Sheets (CSS)

Cascading Style Sheets (CSS) is a style sheet language designed to control the visual presentation of web pages. CSS enhances the user experience by allowing developers to define the look and feel of a website. Key functionalities of CSS include:

- **Design Customization:** Control over layout, colors, fonts, and spacing for a cohesive visual identity.
- **Responsive Design:** Ensures consistent and optimized appearance across different devices and screen sizes.
- **Cascading Rules:** Allows styles to be applied at element, class, or global levels, offering flexibility in design.

As a foundational technology of the web, CSS plays a vital role in creating modern, responsive, and aesthetically pleasing websites.

**?**

### 7.2.3 JavaScript

JavaScript is a high-level programming language used to add interactivity and dynamic content to web pages. It works seamlessly alongside HTML and CSS to create rich and engaging user experiences. Key features of JavaScript include:

- **Dynamic Content:** Enables animations, form validation, and real-time updates.

- **Client and Server-Side Usage:** Runs in web browsers via JavaScript engines and supports server-side applications through platforms like Node.js.
- **Extensive Ecosystem:** Offers libraries, frameworks, and tools for building feature-rich web applications.

JavaScript's flexibility and versatility have established it as a cornerstone of web development, making it essential for developing interactive and responsive applications.

**?**

## 7.3 Type Script

**Flos Part**

# Teil IV

# Implementation of Artificial Intelligence

# 8 Overview and Integration of AI Models

For the Diploma thesis, there are many different AI models that are in use. There are different Types of AI models, such as:

- LLMs (Large Language Models)
- Defusion Models (Models that are used to create images)
- Object Detection Models (Models that are used to detect objects in images)
- Face Recognition Models (Models that are used to recognize faces in images)

In the following chapters, the different Types and the used models will be explained in more detail.

## 8.1 Large Language Models (LLMs)

Large Language Models (LLMs) represent a significant advancement in artificial intelligence, enabling machines to process and generate natural language. LLMs are built on the concept of deep learning, utilizing neural networks with billions of parameters to understand and generate text in a contextually accurate and coherent manner. These models are trained on vast datasets encompassing diverse topics, allowing them to handle a wide range of tasks, such as translation, summarization, content generation, and conversational AI.

### 8.1.1 Key Characteristics of LLMs

- **Scale and Complexity:** LLMs are distinguished by their immense size, often containing billions of parameters, enabling them to capture intricate patterns in language.
- **Transfer Learning:** These models benefit from pretraining on large datasets, followed by fine-tuning for specific tasks, making them highly versatile.
- **Contextual Understanding:** LLMs excel at understanding context, which allows them to generate coherent and contextually appropriate responses.
- **Multilingual Capabilities:** Many LLMs are trained on datasets in multiple languages, enabling them to process and generate text in various languages.

### 8.1.2 Applications of LLMs

- Text summarization and paraphrasing.
- Question answering and information retrieval.
- Conversational agents and chatbots.
- Code generation and debugging assistance.
- Creative writing, including story and poetry generation.

### 8.1.3 Examples of Popular LLMs

- **GPT Models:** Developed by OpenAI, these models include GPT-3, GPT-4, and ChatGPT, known for their state-of-the-art performance in text generation and comprehension.
- **BERT (Bidirectional Encoder Representations from Transformers):** Developed by Google, BERT focuses on understanding context by analyzing text bidirectionally.
- **LLama Models:** Created by Meta, these models are designed for efficient natural language understanding and generation.
- **Mistral Models:** Aimed at specialized tasks with high precision and multilingual capabilities.

### 8.1.4 Advantages and Challenges of LLMs

**Advantages:**

- High accuracy in generating and understanding text.
- Adaptability to a variety of domains and languages.
- Ability to process complex and context-rich queries.

**Challenges:**

- High computational and memory requirements.
- Potential biases due to the training data.
- Difficulty in maintaining factual accuracy in generated content.

**?**

## 8.2 Utilized Large Language Models

In the context of this diploma thesis, various free and commercial large language models (LLMs) were evaluated to determine their suitability for integration. Leveraging the Ollama application, we were able to test and compare several LLMs. Additionally, we explored different ChatGPT models available through the OpenAI API. OpenAI offers a range of models that vary in terms of size and complexity, with more advanced models incurring higher usage costs.

**? ?**

## 8.3 Ollama Application Overview

The Ollama application is an advanced, locally hosted platform designed to provide a versatile environment for deploying and interacting with a wide array of artificial intelligence models. It offers a comprehensive solution for both text and image processing tasks, facilitating the integration, fine-tuning, and management of models in a secure and scalable manner.

### 8.3.1 Ollama Features

Ollama is distinguished by several key features that enhance its functionality and usability:

- **Multi-Model Support:** The platform supports a variety of AI models, each optimized for specific tasks such as natural language processing and image analysis.
- **Local API Hosting:** The API is hosted on a local server, ensuring rapid and secure processing of requests while maintaining full control over data.
- **Image Processing Capabilities:** In addition to textual data, certain models within Ollama are capable of processing images. These models can analyze visual content, thereby extending the application's utility.
- **Model Customization and Fine-Tuning:** Users can fine-tune existing models to suit their specific needs. Once customized, these models can be re-uploaded to the Ollama server, allowing for continuous improvement and adaptation.

### 8.3.2 Ollama Architecture

The architecture of Ollama is modular and designed to support high performance and scalability:

1. **Model Management Layer:** This layer is responsible for deploying, fine-tuning, and updating the various AI models. It provides a structured approach to manage model versions and customizations.
2. **API Service Layer:** Hosted locally, this layer facilitates communication between client applications and the AI models. It exposes endpoints for both text and image processing, ensuring secure and efficient data exchange.
3. **Integration Interfaces:** These interfaces enable seamless connectivity with external services and applications, promoting interoperability and flexibility in diverse operational environments.

This layered design supports efficient resource management while enabling rapid response times and scalability to handle increasing user demands.

### 8.3.3 Ollama Models

Ollama provides a diverse selection of models, each tailored to specific application domains:

- **Text Generation Models:** Optimized for tasks such as dialogue generation, summarization, and other natural language processing applications.
- **Image Analysis Models:** Developed for image recognition, generation, and related tasks.

Furthermore, the platform allows users to fine-tune these models based on their particular requirements. Customized models can be re-uploaded to the server, enabling a continuous cycle of refinement and performance enhancement.

### 8.3.4 Ollama API

The Ollama API is the primary interface through which client applications interact with the hosted models. It provides robust and secure endpoints for processing both textual and visual data:

- **Data Exchange:** The API facilitates structured data exchange between client applications and the backend, ensuring that requests and responses are handled efficiently.
- **Security and Performance:** Designed with stringent security protocols, the API ensures that all interactions are encrypted and managed in a way that maximizes performance while minimizing latency.
- **Extensibility:** The API's modular design allows for the easy addition of new endpoints and functionalities as the platform evolves.

### 8.3.5 Ollama Integration

Integrating the Ollama API into external applications is straightforward. For instance, a Python-based client can send HTTP requests to the API to perform tasks such as generating text or processing images. This section

is further elaborated in the chapter dedicated to the hosted Flask Service, where detailed examples and implementation guidelines are provided. In brief, the integration involves:

- Establishing a connection to the local API endpoint.
- Sending appropriately formatted requests (e.g., JSON payloads) that include user inputs.
- Handling responses from the API, which may include generated text or URLs to processed images.

### 8.3.6 Benefits and Challenges of Ollama

Ollama presents several benefits:

- **Ease of Use:** The platform is user-friendly, with intuitive APIs that simplify deployment and integration.
- **Versatility:** A wide array of models enables the application of Ollama to diverse tasks, from natural language processing to image analysis.
- **Multilingual Support:** The models are capable of processing multiple languages, thereby broadening the scope of potential applications.
- **Customization:** Users can fine-tune models to meet specific needs and update them on the server, ensuring tailored performance.

However, several challenges must be addressed:

- **Performance Limitations:** Larger models may experience slower response times due to higher computational demands.
- **API Request Management:** Ensuring that the API can handle a high volume of requests efficiently requires robust load balancing and error handling mechanisms.
- **Model Management Complexity:** Coordinating updates, fine-tuning, and deployment of multiple models demands an effective management strategy.
- **Concurrency:** Managing simultaneous user requests, as discussed in the chapter on the hosted Flask Service, is critical to maintaining system performance under high load.

In summary, while Ollama offers a flexible and powerful platform for AI model deployment and interaction, addressing its inherent challenges is crucial for optimizing performance and ensuring long-term scalability in practical applications.

## 8.4 Evaluation of Models via the Ollama Platform

In this project, we conducted an evaluation of various models accessible through the Ollama application, which are available for download from the Ollama server.

Given that Ollama operates locally, it was imperative to select models that align with specific criteria to ensure optimal performance. Consequently, we assessed models of diverse sizes and complexities to determine their suitability for local deployment. This evaluation encompassed both the efficacy and efficiency of the models within a local environment.

**?**

### 8.4.1 Model Selection Criteria

The selection of models was guided by the following criteria:

- **Model Size:** The model must be capable of running on the server without exceeding available memory capacity.
- **Performance Speed:** The response time of the model, i.e., how quickly it can generate output.
- **Complexity:** The model's ability to handle complex prompts and generate coherent, contextually accurate text.
- **Accuracy:** The overall precision of the model's responses, particularly in terms of factual correctness and linguistic quality.
- **Language Support:** The model's proficiency in understanding and generating text in multiple languages, particularly English and German.
- **User Experience:** The model's overall usability and user-friendliness, including ease of integration and customization.

There is often a trade-off between these criteria. Larger models tend to exhibit higher accuracy and greater contextual understanding but are generally slower and require more computational resources.

### 8.4.2 Challenges in Model Testing and Updates

One significant challenge lies in the rapid development and frequent release of new models, which complicates the process of continuous integration and comprehensive evaluation of recent advancements. Regular testing and updates are imperative to ensure the incorporation of state-of-the-art models while maintaining system reliability and relevance.

Another challenge involves achieving an optimal balance between performance, accuracy, and resource efficiency, ensuring that the chosen model meets the application's functional requirements without compromising the overall user experience.

During the evaluation process, we encountered several obstacles. For instance, identifying standardized questions that could be uniformly answered by all models proved challenging. Some smaller models demonstrated limitations in addressing certain questions comprehensively. Additionally, specialized models, while excelling in niche areas, often lacked the ability to provide detailed answers across a broader range of topics.

For the final evaluation phase, we employed a diverse question set consisting of self-crafted questions, publicly available questions from online sources, and queries generated by ChatGPT-4 to ensure a comprehensive assessment covering a wide spectrum of queries.

### 8.4.3 Model Selection for the Final Application

In the final implementation, users are provided with a curated list of recommended models from which they can select their preferred option. This list was carefully compiled based on our comprehensive testing and reflects the models that demonstrated the best balance between performance, accuracy, and resource efficiency.

For the production version of the application, this list must be updated periodically to include newly released models and maintain optimal performance.

## 8.5 Ollama Model Testing and Evaluation

Based on the aforementioned criteria, we conducted an extensive evaluation of the following models:

### 8.5.1 Quantitative Evaluation Methods

For the quantitative evaluation, we focused on key performance metrics to assess the efficiency and reliability of each model:

- **Response Time:** The time taken by the model to generate a response after receiving input.
- **CPU Usage:** The percentage of CPU resources utilized during model execution.
- **GPU Usage:** The extent to which GPU resources were leveraged to enhance performance.
- **Memory Usage:** The amount of RAM consumed while the model was running.
- **Multiple Choice Question Answering:** The accuracy of the model when answering structured multiple-choice questions.
- **Translation Quality:** Measured using the BLEU (Bilingual Evaluation Understudy) score, which evaluates the similarity between the model-generated translation and a human reference translation. **?**
- **Text Generation Quality:** Assessed using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score to measure the overlap between the generated text and reference texts.

**?**

### 8.5.2 Qualitative Evaluation Methods

Qualitative evaluation requires human judgment and is inherently resource-intensive. Therefore, our primary focus was placed on the quantitative evaluation. Nevertheless, we conducted qualitative assessments for specific criteria where human input was indispensable:

- **Contextual Understanding:** The model's ability to interpret and respond accurately to the context of the input text.
- **Factual Correctness:** The accuracy of the factual information provided in the model's output.
- **Linguistic Quality:** The grammatical correctness, fluency, and coherence of the generated text.
- **Emotional Intelligence:** The extent to which the model's output mimics human-like empathy, tone, and emotional nuance.

By combining both quantitative and qualitative evaluation methods, we obtained a comprehensive understanding of each model's strengths and limitations.

### 8.5.3 Models Evaluated During Testing

For the evaluation process, we selected the most popular models available in the Ollama application and conducted extensive testing on each.

- **qwen2.5-coder:0.5b** - A compact model with 0.5 billion parameters, specifically designed for coding tasks.
- **qwen2.5-coder:7b** - A small-scale model with 7 billion parameters, optimized for coding tasks.
- **qwen2.5-coder:14b** - A medium-sized model with 14 billion parameters, tailored for coding-related tasks.
- **llama3.2:1b** - A small model with 1 billion parameters, designed for general-purpose tasks by Meta.
- **llama3.2:2b** - A medium model with 2 billion parameters, developed for general-purpose tasks by Meta.

- **mistral:7b** - A medium-sized model with 7 billion parameters, created for general tasks by Mistral AI, a European AI company.
- **phi4:14b** - A medium-sized model with 14 billion parameters, intended for general tasks by Microsoft.
-

**?**

### 8.5.4 Data Collection

For the Data collection we used the School AI Server wich was provided by the HTL Anichstraße, to run the different models in the Evaluation phase. For the later Data collection we used our own Server to run the different models.

To colletct the data we used a varaitey of different python scripts. For the quantitative data we used the following python script:

```python
import time
import json
import psutil  # For CPU, memory usage
from ollama import chat

# Prompts for testing
prompts = [
    "Explain the theory of relativity in simple terms.",
    "Create a short story about a knight.",
    "What are the advantages of open-source projects?",
    "Write a Python function that outputs prime numbers up to 100.",
    # ...
]

# Model name
model_name = "qwen2-math"

# Store results
results = []

# Function to get GPU usage if available
def get_gpu_usage():
    try:
        import torch
        if torch.cuda.is_available():
            gpu_memory = torch.cuda.memory_allocated() / (1024 ** 2)  # Convert to MB
            gpu_utilization = torch.cuda.utilization(0) if hasattr(torch.cuda, 'utilization') else "N/A"
            return gpu_memory, gpu_utilization
        else:
            return 0, "No GPU detected"
    except ImportError:
        return 0, "torch not installed"

# Loop through prompts
for prompt in prompts:
    try:
        # Measure system usage before model execution
        cpu_before = psutil.cpu_percent(interval=None)
        memory_before = psutil.virtual_memory().used / (1024 ** 2)  # Convert to MB
```

```
40
41          start_time = time.time()
42          # Ollama chat request
43          response = chat(model=model_name, messages=[{'role': 'user', 'content': prompt}])
44          end_time = time.time()
45
46          latency = end_time - start_time
47
48          # Measure system usage after model execution
49          cpu_after = psutil.cpu_percent(interval=None)
50          memory_after = psutil.virtual_memory().used / (1024 ** 2)  # Convert to MB
51
52          cpu_usage = cpu_after - cpu_before
53          memory_usage = memory_after - memory_before
54          gpu_memory_usage, gpu_utilization = get_gpu_usage()
55
56          # Extract content from the Message object
57          if response and hasattr(response["message"], "content"):
58              response_text = response["message"].content  # Accessing the attribute of the Message object
59          else:
60              response_text = "No content returned or unexpected format"
61
62          print(f"Prompt: {prompt}\nResponse Time: {latency:.2f} seconds\n")
63
64          # Save the result
65          results.append({
66              "Prompt": prompt,
67              "Response Time (seconds)": latency,
68              "Response": response_text,
69              "CPU Usage (%)": cpu_usage,
70              "Memory Usage (MB)": memory_usage,
71              "GPU Memory Usage (MB)": gpu_memory_usage,
72              "GPU Utilization (%)": gpu_utilization
73          })
74      except Exception as e:
75          print(f"Error with prompt '{prompt}': {e}")
76          results.append({
77              "Prompt": prompt,
78              "Response Time (seconds)": "Error",
79              "Response": f"Error: {str(e)}",
80              "CPU Usage (%)": "N/A",
81              "Memory Usage (MB)": "N/A",
82              "GPU Memory Usage (MB)": "N/A",
83              "GPU Utilization (%)": "N/A"
84          })20.01.2025
85
86  # Save to JSON file
87  json_file_name = model_name + "_response_time_results_ressours_usage.json"
88  with open(json_file_name, "w") as file:
89      json.dump(results, file, indent=4)
90
91  print(f"The results have been saved in {json_file_name}.")
```

Listing 8.1: Python-quantitative-data-collection

For this Pyhton Skript we used ChatGPT to help with the psutil library to get the CPU and Memory usage. We also used ChatGPT to get more Questions for the data collection.

The results were saved as JSON files for the later analysis. One Problem we encountered were the tourch librarys, witch diden't function probally on the School AI Server, so we couldn't get the GPU usage for the models.

#### Used python libaries

**psutil libarie**

The `psutil` library is a Python module that provides an interface for retrieving information on system utilization, including CPU, memory, disk, network, and processes. It is commonly used for monitoring and managing system performance and is highly efficient due to its low overhead. `psutil` is cross-platform, supporting major operating systems like Windows, Linux, and macOS. It enables developers to create scripts for system diagnostics, process control, and resource management, making it an essential tool for performance optimization and system administration in Python-based projects.

**?**

**Ollama**

The `ollama` library is a Python package designed to provide a seamless interface for interacting with the Ollama application. It allows users to easily access and leverage various AI models for natural language processing tasks. By simplifying the integration of AI models into Python applications, the library supports a wide range of functionalities, making it an efficient tool for developing AI-powered solutions.

**?**

### 8.5.5 Testresults and Analysis

Explain how the data was analyzed and interpreted. implement nice Graphics

### 8.5.6 Model Comparison for different Use Cases and Scenarios

Explain wich modell is best for what

### 8.5.7 Model Selected for the Final Application

Explain why the selected model was selected and what the criteria were and explain the different models.

### 8.5.8 Model Integration and Deployment

Explain how the model was integrated and deployed

## 8.6 Integration of OpenAI's API

In this work, we integrated the OpenAI API to leverage proprietary, high-performance AI models that are hosted on dedicated servers with advanced hardware capabilities. The utilization of external computing power allows for the concurrent execution of multiple models, thereby enhancing both scalability and efficiency in our application.

The decision to adopt the OpenAI API was influenced by its widespread adoption, robust performance, and extensive documentation. Numerous examples, tutorials, and community resources are available, which greatly facilitate the integration process and ensure that best practices are followed in scientific and industrial applications.

### 8.6.1 Overview of the OpenAI API

The OpenAI API provides access to state-of-the-art AI models developed by OpenAI, including various iterations of the ChatGPT model. These models are capable of generating human-like text, answering queries, and engaging in complex conversations. The API supports a range of models with different sizes and capabilities, allowing users to select the model that best fits the requirements of their specific use cases.

Designed with user accessibility in mind, the API comes with comprehensive documentation and a wealth of code samples, which significantly streamline the process of embedding advanced AI functionalities into diverse applications and platforms. Furthermore, the API utilizes a token-based pricing model, which charges users according to the number of tokens processed during interactions. This pricing structure is not only transparent but also aligns closely with the computational effort required to generate responses.

Before accessing the API's full functionality, users must pre-fund their accounts by depositing a specified amount of money. This account-based billing system enables users to manage their expenditures effectively, including the option to set monthly spending limits. In addition to text generation, the OpenAI ecosystem also includes DAL-E, an image-generation model that creates visuals based on textual input, thus broadening the spectrum of applications available through the API.

**?**

### Tokens in Large Language Models

Tokens are the fundamental units of text that large language models (LLMs) process and generate. In this context, a token represents the smallest segment of text that a model can understand, which may correspond to an entire word, a fragment of a word, or even an individual character or punctuation mark.

The process of tokenization involves converting raw text into these discrete units. This approach enables LLMs to efficiently capture complex patterns in both syntax and semantics, even when encountering new or out-of-vocabulary terms. Techniques such as subword tokenization are particularly valuable, as they break down words into meaningful components, thereby reducing the overall vocabulary size and enhancing the model's ability to manage linguistic variability.

Moreover, tokens are closely related to the concept of a context window, which defines the span of tokens a model can consider during text generation or prediction. Typically, one token is estimated to average around four

characters in English or roughly three-quarters of a word. This estimation is crucial for determining computational requirements and understanding the limitations imposed by the model's finite context window.

In summary, tokens are indispensable for the operation of LLMs, providing a structured means to process language. Their effective management through advanced tokenization strategies is essential for optimizing both the computational efficiency and the overall performance of these models.

**?**

### 8.6.2 Data Security and Privacy in Compliance with Austrian and EU Regulations

The integration of OpenAI's API into our systems necessitates a thorough examination of data security and privacy considerations, particularly in the context of Austrian and European Union (EU) regulations. The General Data Protection Regulation (GDPR) serves as the cornerstone of data protection within the EU, imposing stringent requirements on the processing of personal data.

OpenAI has implemented several measures to safeguard user data and align with GDPR mandates. Notably, they support compliance with privacy laws such as the GDPR and the California Consumer Privacy Act (CCPA), offering a Data Processing Addendum to customers. Their API and related products have undergone evaluation by an independent third-party auditor, confirming alignment with industry standards for security and confidentiality.

**?**

Despite these measures, concerns have been raised regarding data handling practices. For instance, data transmitted through the OpenAI API could potentially be exposed, and compliance with GDPR remains a complex issue. Additionally, data may be accessible to third-party subprocessors, introducing further privacy considerations.

**?**

To address these concerns, we have proactively informed our user community through a notice on the school website. This notice outlines the data handling practices associated with the OpenAI API and provides guidance on how users can manage their data when interacting with our systems. By maintaining transparency and offering clear instructions, we aim to uphold the highest standards of data security and privacy in our academic environment.

In light of the evolving regulatory landscape, it is imperative to remain vigilant and responsive to any changes in data protection laws within Austria and the broader EU. Continuous monitoring and adaptation of our data handling practices will ensure ongoing compliance and the safeguarding of user privacy.

### 8.6.3 OpenAI API Implementation in Vue.js

This section details the integration of the OpenAI API within a Vue.js application framework, with a focus on both text and image generation capabilities. The implementation not only illustrates the interaction between the Vue.js frontend and the OpenAI API but also demonstrates adherence to security best practices and modular code design. The following discussion is supported by annotated code examples and an explanation of the libraries used.

#### Overview of the Implementation

The implementation is structured as a Vue.js component that facilitates the following functionalities:

- Accepting user input via a text area.
- Initiating API calls for generating text responses (using ChatGPT models) and creating images (via the DALL-E endpoint).
- Displaying the results (generated text and images) dynamically within the user interface.

The component is designed with a clear separation between presentation and business logic, ensuring that the code remains both maintainable and scalable.

### Explanation of the Used Libraries

**OpenAI Library** The `openai` library is employed as the primary interface to interact with OpenAI's API endpoints. This library abstracts the complexities of HTTP communication and provides a user-friendly API to access advanced AI functionalities such as natural language generation and image synthesis. Its integration simplifies the process of constructing API requests and handling responses, which is critical for developing robust AI-driven applications.

**API Key Management** To ensure secure handling of sensitive credentials, the OpenAI API key is imported from an external module (i.e., `OPENAI_API_KEY` from the `secrets` file). This approach adheres to security best practices by preventing the direct embedding of API keys within the source code, thereby mitigating the risk of unauthorized exposure.

### Code Example: Vue.js Component for OpenAI API Integration

Below is an illustrative example of a Vue.js component that integrates the OpenAI API for both text and image generation. The code is presented in two parts: the HTML template and the JavaScript logic.

Listing 8.2: Vue.js Template for OpenAI API Integration

```
<template>
  <div class="openai-container">
    <h1>OpenAI API Integration in Vue.js</h1>
    <textarea
      v-model="userInput"
      placeholder="Enter your prompt here..."
```

```
        rows="4"
        cols="50">
      </textarea>
      <div class="action-buttons">
        <button @click="generateText">Generate Text</button>
        <button @click="generateImage">Generate Image</button>
      </div>
      <div v-if="generatedText" class="output-section">
        <h2>Generated Text</h2>
        <p>{{ generatedText }}</p>
      </div>
      <div v-if="generatedImage" class="output-section">
        <h2>Generated Image</h2>
        <img :src="generatedImage" alt="Image␣generated␣by␣OpenAI␣API" />
      </div>
    </div>
</template>
```

Listing 8.3: Vue.js Script for OpenAI API Integration

```
<script>
import OpenAI from "openai";
import { OPENAI_API_KEY } from "../secrets";

export default {
  name: "OpenAIComponent",
  data() {
    return {
      userInput: "",
      generatedText: "",
      generatedImage: ""
    };
  },
  methods: {
    async generateText() {
```

```
      // Initialize OpenAI client with API key
      const openai = new OpenAI({ apiKey: OPENAI_API_KEY });
      try {
        const response = await openai.chat.completions.create({
          model: "gpt-3.5-turbo",
          messages: [{ role: "user", content: this.userInput }]
        });
        // Extract and assign the generated text
        this.generatedText = response.choices[0].message.content;
      } catch (error) {
        console.error("Error during text generation:", error);
      }
    },
    async generateImage() {
      // Initialize OpenAI client for image generation
      const openai = new OpenAI({ apiKey: OPENAI_API_KEY });
      try {
        const response = await openai.images.generate({
          prompt: this.userInput,
          n: 1,
          size: "512x512"
        });
        // Extract and assign the URL of the generated image
        this.generatedImage = response.data[0].url;
      } catch (error) {
        console.error("Error during image generation:", error);
      }
    }
  }
};
</script>
```

### Discussion

The presented component exemplifies how modern web applications can seamlessly integrate AI capabilities while maintaining a secure and modular architecture. Key points of consideration include:

- **Modularity:** The separation of the UI (HTML template) and the business logic (JavaScript methods) facilitates easier maintenance and potential scalability.
- **Security:** By importing the API key from an external secrets module, the risk of credential leakage is minimized. This practice is crucial in academic and production environments where data security is paramount.
- **Extensibility:** The design allows for further expansion, such as additional error handling mechanisms or the integration of more advanced functionalities provided by the OpenAI API.

In conclusion, this integration not only demonstrates the practical application of AI APIs in modern web development but also reflects best practices in secure and maintainable code design. Such an approach is essential for building reliable applications in both academic research and industrial contexts.

## 8.7 Conclusion

# 9 hosted Flask Service

## 9.1 Server structure

# 10 Studen AI Website

# Teil V

# Evaluations

# 11 Open source evaluation on Economics

## 11.1 Introduction

### 11.1.1 Chapter Overview

This chapter introduces the concept of Open Source and highlights its significance in the modern economy. Key aspects such as the advantages and disadvantages of Open Source, as well as the challenges associated with its adoption and creation, are discussed. Additionally, the chapter explores revenue models within the Open Source ecosystem and its role in economic systems. Finally, the chapter concludes by presenting the Open Source tools utilized in this project, alongside a reflection on the experiences gained through their application.

### 11.1.2 What is Open Source?

Open Source represents a collaborative and transparent approach to software development and distribution, where the source code is made publicly accessible. This philosophy empowers users not only to utilize the software but also to modify, improve, and redistribute it freely. By fostering an environment of openness and collaboration, Open Source drives innovation and democratizes access to technology.

Linus Torvalds, the creator of the Linux operating system, encapsulated this spirit of freedom and collaboration with his famous remark:

*"Software is like sex: it's better when it's free."*

**?**

This statement highlights the fundamental ethos of Open Source—the belief that open access and shared knowledge result in better, more impactful solutions.

The development process for Open Source software is often a collective effort, with contributions from diverse communities of developers, users, and organizations. These collaborative efforts enhance the software's functionality, security, and usability, resulting in products that are robust and adaptable. Prominent examples include the Linux operating system, the Apache web server, and the Firefox web browser, all of which have significantly influenced technological innovation and market dynamics.

**?**

### 11.1.3 Advantages of Open Source

Open Source software offers a wide range of benefits, making it a cornerstone of modern technology:

- **Cost Efficiency:** Open Source software is typically free of charge, helping organizations and individuals save on licensing and maintenance costs.
- **Flexibility:** Users can access the source code, enabling them to tailor the software to their specific needs and requirements.
- **Security:** The open nature of the source code allows for peer review, ensuring vulnerabilities are identified and addressed promptly.
- **Community Support:** Open Source projects often benefit from vibrant developer communities, providing updates, patches, and user assistance.
- **Innovation:** The collaborative ecosystem of Open Source encourages creativity, leading to groundbreaking solutions and advancements.
- **Compatibility:** Many Open Source projects are designed to integrate seamlessly with existing systems, reducing technical barriers.

- **Transparency:** Open access to the source code ensures that users can understand and verify how the software operates.
- **Freedom:** Users are granted the liberty to use, modify, and share the software without restrictive licensing agreements.

**? ?**

### 11.1.4 Why Do People Use Open Source?

The adoption of Open Source software is motivated by several compelling factors:

- **Control:** Users gain full control over the software, enabling customization and optimization for specific use cases.
- **Cost Savings:** The absence of licensing fees significantly reduces expenses, making Open Source particularly attractive for startups and educational institutions.
- **Security:** Transparency in the source code allows for thorough auditing, enhancing trust and reliability.
- **Community:** The collaborative spirit of Open Source connects users with knowledgeable communities that share resources and support.
- **Stability:** Many Open Source projects offer long-term support and regular updates, ensuring reliability over time.
- **Skill Development:** Learning and using Open Source tools are valuable in educational and professional contexts, equipping individuals with in-demand skills.

## 11.2 What is and isn't Open Source?

### 11.2.1 Definition and Guiding Principles

Open Source, as defined by the Open Source Initiative (OSI), is a development approach that prioritizes accessibility and transparency of software

source code. It allows users to view, modify, and distribute the code freely, fostering collaboration and innovation.

The OSI outlines several key principles that define Open Source software:

- **Free Redistribution:** The software can be freely shared and distributed without restrictions.
- **Source Code Access:** Users must have access to the source code to study, modify, and improve the software.
- **Modification and Sharing:** Users are allowed to create and share modified versions, as long as they follow the license terms.
- **No Discrimination:** The software must be available for everyone, regardless of individual characteristics or professional field.
- **Neutrality and Compatibility:** The license must not favor specific technologies or restrict the use of other software.

These principles ensure that Open Source remains a transparent, inclusive, and adaptable approach to software development, enabling innovation and collaboration across industries and communities.

**?**

### 11.2.2 Misconceptions About Open Source

Open Source is often misunderstood and confused with other software distribution models, which can lead to misconceptions about its nature, functionality, and benefits. It is crucial to distinguish Open Source from other types of software:

- **Open Source:** Software that is freely accessible, modifiable, and redistributable under an Open Source license, adhering to principles such as transparency and collaboration.
- **Freeware:** Software available at no cost but typically without access to the source code, meaning users cannot modify or redistribute it.
- **Proprietary Software:** Software owned and controlled by a single entity, restricting access to the source code and preventing users from making modifications or redistributions.

- **Commercial Software:** Software sold for profit, which may be either Open Source or proprietary, depending on the licensing terms.

Understanding these distinctions helps users make informed choices about software selection and ensures their expectations align with the capabilities and freedoms provided by the chosen software.

To verify whether a software is truly Open Source, it is essential to examine the license agreement and confirm the availability of the source code. Software with an OSI-approved license is a reliable indicator that it adheres to Open Source principles, providing transparency, freedom, and collaboration opportunities.

One common misconception about Open Source software arises from the phrase "free as in freedom"versus "free as in free beer."While "free as in freedomëmphasizes the liberty to access, modify, and share the software, "free as in free beerßimply denotes that the software is free of cost. Although Open Source software is often available without charge, its true value lies in the freedoms it grants to users, developers, and organizations. This distinction highlights the broader significance of Open Source as a philosophy, not just a pricing model.

**?**

## 11.3 The Role of Open Source in Economics

Cost efficiency, innovation, and collaboration are key factors that have positioned Open Source as a cornerstone of modern economic systems. M any industries and organizations utilize Open Source software to reduce costs, increase flexibility, and promote creativity, thereby driving economic growth and sustainability.

### 11.3.1 Driving Innovation and Shaping Market Dynamics

Open Source software fosters a culture of experimentation, creativity, and knowledge sharing, leading to the rapid development of new technologies

and solutions. By granting users access to modify and redistribute the source code, Open Source encourages collaboration and innovation, enabling individuals and organizations to build upon existing software to create new products and services.

A distinctive strength of Open Source is its inclusivity—anyone, regardless of their affiliation with a company, can contribute to its development. This openness lowers barriers to entry for innovation and allows passionate individuals to make meaningful contributions.

Companies also play a significant role in advancing Open Source projects. With greater resources and structured teams, organizations can contribute in a more organized and impactful manner, accelerating development and enhancing software quality.

The collaborative nature of Open Source facilitates cross-industry partnerships, allowing organizations from diverse sectors to share knowledge, resources, and best practices. This cross-pollination of ideas not only enhances software development but also fosters innovation across industries, ultimately shaping market dynamics and driving economic progress.

The study **?** by Mike Hendrickson, Roger Magoulas, and Tim O'Reilly underscores that Open Source is not only a catalyst for small business growth but also a driver of future success for many startups today. By providing cost-effective and flexible solutions, Open Source enables small and medium-sized enterprises to strengthen their online presence and enhance their economic performance.

### 11.3.2 Supporting Startups and small Enterprises

The impact of Open Source on startups and small enterprises is both profound and transformative. For these businesses, Open Source software provides a highly cost-effective alternative to proprietary solutions, granting access to advanced tools and technologies without the financial burden of high licensing fees typically associated with commercial software. This

affordability allows startups and small enterprises to allocate their limited resources more strategically, fostering innovation and growth while maintaining financial flexibility.

**?**

### 11.3.3 Enabling Cross-Industry Collaboration and Open Innovation

## 11.4 Advantages and Disadvantages of Open Source

### 11.4.1 Advantages

Open Source software offers numerous advantages for users, developers, and businesses. It can vary from cost savings to increased innovation and flexibility for customization.

- Cost savings.
- Flexibility for customization.
- Increased innovation due to open collaboration.

### 11.4.2 Disadvantages

- Reliance on community support.
- Potential security vulnerabilities.
- Compatibility issues with other systems.

## 11.5 Challenges of Using or Creating Open Source

There are many challanges that come with using or creating Open Source software. These can range from technical to economic and social challenges. Understanding these challenges is crucial for successful Open Source adoption and development.

### 11.5.1 Technical Challenges

- Maintaining quality and long-term compatibility.
- Managing security and privacy risks.

### 11.5.2 Economic Challenges

- Monetization and sustainability concerns.
- Balancing free access with profitability.

### 11.5.3 Social Challenges

- Effective community management and governance.

### 11.5.4 Legal Issues

- Navigating complex licensing models (e.g., GPL, MIT).

## 11.6 Revenue Models in Open Source

Open Source projects can generate revenue through various business models, each with its own advantages and challenges.

- Common business models:
  - Freemium.

  - Support and maintenance services.
  - Dual licensing.
  - Crowdfunding and donations.
- Real-world examples of successful Open Source businesses (e.g., Linux, Red Hat, MySQL).

## 11.7 Open Source in Key Industries

- The role of Open Source in transforming:
  - Information Technology (e.g., operating systems, tools).
  - Artificial Intelligence (e.g., TensorFlow, PyTorch).
  - Education (e.g., Moodle, Jupyter Notebooks).
- Governmental and policy support for Open Source adoption.

## 11.8 Reflexion

- Answering the research question based on the above analysis.
- Evaluating the broader implications of Open Source for economic systems.
- Connecting Open Source's potential with sustainability and global development.

## 11.9 Open Source in Practice: A Personal Experience

- Open Source tools and technologies used in the project:
  - Python, Flask, Vue.js, Linux, wttr.in API, LLaMA API.
- Challenges and solutions encountered:
  - Technical hurdles.
  - Why Open Source alternatives were chosen or rejected.

- Comparison of Open Source and closed-source software used:
  - Reasons for choosing closed-source alternatives where applicable.

## 11.10 Open Source in Our Project & Licensing

### 11.10.1 Project

- Description of the project.
- How Open Source principles were applied.
- Benefits and challenges of Open Source in the project.

### 11.10.2 License

- Choice of license and rationale.
- How the license aligns with the project's goals.
- The license problems of the project.
- Future plans for the project's development and licensing.

## 11.11 Conclusion

- Summary of Open Source's economic impact.
- Reflections on its potential to drive future innovation and growth.
- Final thoughts on your personal experience and insights gained.

# Teil VI

# Conclusion

# 12 Conclusion

# 13 Proplems that occured

# 14 Outlook

# Appendix

# Tabellenverzeichnis

# Abbildungsverzeichnis

# Listings

# Literaturverzeichnis

*10 biggest advantages of open-source software* (2022). [Online; accessed 2025-01-28].
**URL:** *https://www.rocket.chat/blog/open-source-software-advantages*

Ankush (2024), 'What is ollama? everything important you should know'. [Online; accessed 2025-01-13].
**URL:** *https://itsfoss.com/ollama/*

Forbes Technology Council (2024), 'Misconceptions about open source solutions clarified by tech experts'. Accessed: 2024-12-04.
**URL:** *https://www.forbes.com/councils/forbestechcouncil/2024/10/09/misconceptions-about-open-source-solutions-clarified-by-tech-experts/*

Fortis, S. (2024), 'Openai hit with privacy complaint in austria, potential eu law breach'. [Online; accessed 2025-02-07].
**URL:** *https://cointelegraph.com/news/openai-privacy-complaint-austria-potential-eu-law-breach*

Foy, P. (2024), 'Understanding tokens & context windows'. [Online; accessed 2025-02-07].
**URL:** *https://blog.mlq.ai/tokens-context-window-llms/*

Hendrickson, M., Magoulas, R. and O'Reilly, T. (2012), *Economic Impact of Open Source on Small Business: A Case Study*, O'Reilly Media.
**URL:** *https://www.oreilly.com/library/view/economic-impact-of/9781449343408/*

*HTML - Wikipedia* (2001). [Online; accessed 2025-01-23].
**URL:** *https://en.wikipedia.org/wiki/HTML*

IBM (n.d.), 'What are large language models (llms)? | ibm'. [Online; accessed 2025-01-21].
**URL:** *https://www.ibm.com/think/topics/large-language-models*

Initiative, O. S. (2007), 'The open source definition', https://opensource.org/osd. Accessed: 2024-12-02.

*Introducing data residency in Europe | OpenAI* (n.d.). [Online; accessed 2025-02-07].
**URL:** *https://openai.com/index/introducing-data-residency-in-europe/?utm$_s$ource = chatgpt.com*

*Ollama* (n.d.*a*). [Online; accessed 2025-02-07].
**URL:** *https://ollama.com/search*

*Ollama* (n.d.*b*). [Online; accessed 2025-01-20].
**URL:** *https://ollama.com/search*

*ollama/ollama-python: Ollama Python library* (n.d.). [Online; accessed 2025-01-21].
**URL:** *https://github.com/ollama/ollama-python*

OpenSource.com (2024), 'What is open source?', https://opensource.com/resources/what-open-source. Accessed: 2024-12-02.

*Overview - OpenAI API* (n.d.*a*). [Online; accessed 2025-01-13].
**URL:** *https://platform.openai.com/docs/overview*

*Overview - OpenAI API* (n.d.*b*). [Online; accessed 2025-02-07].
**URL:** *https://platform.openai.com/docs/overview*

*psutil · PyPI* (2024). [Online; accessed 2025-01-21].
**URL:** *https://pypi.org/project/psutil/*

Santhosh, S. (2023), 'Understanding bleu and rouge score for nlp evaluation | by sthanikam santhosh | medium'. [Online; accessed 2025-01-13].
**URL:** *https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb*

StudioLabs (2024), 'Open source for startups: Lower costs, higher growth'. Accessed: 2024-12-04.
**URL:** *https://www.studiolabs.com/open-source-for-startups-lower-costs-higher-growth/*

*The Pros and Cons of Open-Source Software: A Guide for Developers and Executives* (2023). [Online; accessed 2025-01-28].
**URL:** *https://www.bairesdev.com/blog/the-pros-and-cons-of-open-source-software-a-guide-for-developers-and-executives/*

to Wikimedia projects, C. (2001*a*), 'Css - wikipedia'. [Online; accessed 2025-01-23].
**URL:** *https://en.wikipedia.org/wiki/CSS*

to Wikimedia projects, C. (2001*b*), 'Javascript - wikipedia'. [Online; accessed 2025-01-23].
**URL:** *https://en.wikipedia.org/wiki/JavaScript*

to Wikimedia projects, C. (2006), 'Bleu - wikipedia'. [Online; accessed 2025-01-13].
**URL:** *https://en.wikipedia.org/wiki/BLEU*

Torvalds, L. (2024), 'Linus torvalds quotes', https://www.brainyquote.com/quotes/linus_torvalds_135583. Accessed: 2024-12-02.

Tran-Thien, V. (n.d.), 'Key criteria when selecting an llm'. [Online; accessed 2025-01-13].
**URL:** *https://blog.dataiku.com/key-criteria-when-selecting-an-llm*