Hungarian Association of Agricultural Informatics
European Federation for Information Technology in Agriculture, Food and the Environment

Journal of Agricultural Informatics. Vol. 15, No. 2
journal.magisz.org

OpenSource alternatives of Generative Artifical Intelligence for SME's

Róbert Szilágyi[1]

INFO
Received 11/01/2025
Accepted 27/01/2025
Available on-line 04/02/2025
Responsible Editor: László Várallyai

Keywords:
LLM, open-source, KNIME, Ollama, RAG

ABSTRACT

This paper investigates the potential of local Large Language Models (LLMs) for Small and Medium-sized Enterprises (SMEs). While cloud-based LLMs offer powerful capabilities, their associated costs, including subscription fees and token-based pricing, can be prohibitive for many SMEs. This research explores the benefits of developing and deploying custom, local LLM solutions, which offer advantages such as reduced operational costs, enhanced data privacy and security, and greater flexibility for customization and integration. The paper examines viable open-source alternatives to commercial LLMs, including Ollama, LangChain, Gpt4All, and the integration of LLMs within the data science platform KNIME. Furthermore, it explores techniques for improving LLM performance, such as Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA). General properties of local LLM solutions, such as Command Line Interface (CLI) and Graphical User Interface (GUI) options and multi-platform support, are also discussed. By embracing local LLM solutions, SMEs can leverage the power of AI while mitigating the challenges associated with cloud-based services. This approach empowers businesses to gain a competitive advantage, enhance operational efficiency, and drive innovation while maintaining control over their data and minimizing costs.

1. Introduction

AI (artificial intelligence) and large language models like GPT (Generative Pre-trained Transformer) have caused rapid and profound changes in the field of companies and the economy. There are several existing solutions for SMEs for Generative AI applications. Achieving significant business optimization results is easy with LLMs like ChatGPT, Claude, Gemini, and others. If a company wants to see "what is under the hood," there are several free, custom, local LLM alternatives. In this article, I will highlight some lesser-known solutions and explore ways to expand the list of possible applications.

Research questions

In this paper the local custom generative AI is taken as a possible support tool, and I am looking for the answer to the alternatives of not ChatGPT-based solutions. The following issues could be examined here:

- What about the cost of generative AI?
- What are the main factors affecting local customs LLM?
- What about the ChatGPT alternatives?

Methodology

To answer the previously asked questions, I used the following three key methodologies in the research: case analysis, secondary analysis of specialist-related articles, and literature review. In **Case Analysis** I studied specific local custom generative AI-related cases or examples. In the **Secondary Analysis of Specialist-Related articles**, I used the relevant parts to answer our research question. In the **Literature Review**, I examined the existing scholarly works, articles, books, and other sources that are relevant to our research.

[1] Róbert Szilágyi
University of Debrecen
szilagyi.robert@econ.unideb.hu

doi: 10.17700/jai.2024.15.2/733

1.1. The importance and the cost of the Large Language Model

While LLMs offer significant benefits, it's essential to consider their costs:

Subscription-based Pricing: Some LLM providers offer tiered subscription plans, with varying capabilities and costs. OpenAI, for instance, has a free tier, a Plus tier for $20/month, and a Pro tier for $200/month (OpenAI.com, 2025a). However, these costs may vary depending on the provider and the chosen plan.

Token-based Pricing: Many LLM services use a token-based pricing model. Tokens represent units of text processed by the LLM. The cost of each request depends on the number of input and output tokens used. Pricing may differ between input and output tokens (Microsoft, 2025). For instance, OpenAI's gpt-4o-based model charges around $2.50 per 1 million input tokens and $10 per 1 million output tokens (OpenAI.com, 2025b).

Why should we develop a custom local LLM?

To overcome these cost considerations and gain greater control over their AI capabilities, SMEs should explore developing and deploying custom, local LLMs. This approach offers several key advantages. Firstly, it significantly reduces operational costs compared to cloud-based solutions, eliminating the need for recurring subscription fees and providing greater control over predictable expenses. Secondly, running LLMs locally enhances data privacy and security by ensuring that sensitive information remains within the company's infrastructure, crucial for organizations dealing with confidential data and essential for complying with data privacy regulations such as the General Data Protection Regulation (GDPR).

Furthermore, developing local LLMs allows for greater customization and flexibility. SMEs can fine-tune these models to meet their related needs and requirements, resulting in more effective and relevant AI applications. Local APIs enable seamless integration of LLMs into existing applications and workflows, further enhancing flexibility.

Finally, local LLMs provide uninterrupted operations even in situations with limited or unreliable internet connectivity, enhancing the overall reliability and stability of AI-powered applications.

By developing and deploying custom, local LLMs, SMEs can unlock the full potential of AI while addressing critical concerns related to cost, data privacy, and security. This empowers them to gain a competitive advantage, drive innovation, and achieve sustainable growth in today's rapidly evolving business landscape (Medium.com, 2024b)

The future of NLP in the shadow of Generative AI

Generative AI has transformed NLP by enabling the creation of large amounts of text data, allowing for more creative applications beyond traditional parsing and classification tasks. While Large Language Models are powerful, they are also expensive. LLMs are often overused, and applied to problems that simpler methods could solve more effectively. This is problematic due to the high cost of LLMs. *We have to use a proper model for the actual task***.** The *traditional NLP models* are well-suited for tasks that follow clear patterns and have well-defined goals. For example, identifying spam emails is a good fit for standard models. The *LLMs (Large Language Models)* are best used for tasks that require creativity and the generation of new content. However, it's crucial to ensure that the potential benefits of using an LLM outweigh the significant costs associated with them. (Datacamp, 2025)

1.2. What new skills may be needed in businesses?

Prompt design and prompt engineering are essential for getting accurate results from AI systems. Prompts are input instructions or questions that determine the responses generated by the model. To ensure accurate and useful answers from GPT or other language models, it's crucial to create effective prompts. Prompt engineering, as defined by Lo (2023), is the process of crafting

questions or instructions that guide the model to the desired outcomes. To ensure success, users must learn to create efficient and accurate prompts. **Acquiring programming basics** is inevitable, while proficiency in programming is not necessary for using GPT and other AI models, basic programming knowledge can be advantageous. The very high-level visual programming environments like KNIME, and Scratch allow the user to focus mainly on the workflows, but the programming basics are a suggested skill (Ihrmark and Tyrkkö, 2023). (**API (Programming Interface)** use, APIs allow applications to communicate with other services and systems. There are a growing number of APIs in the field of AI that allow developers and companies to easily access AI systems and integrate them into their applications.
These skills and approaches can help companies take advantage of the opportunities offered by AI and large language models. AI is an increasingly widespread technology, and those who properly prepare and apply it can gain a competitive advantage in the age of digitization and automation.

2. Results

In the article, I will highlight several possible ChatGPT alternatives:

- Omni-Layer Learning Language Acquisition Model (Ollama)
- Gpt4All
- LangChain
- Konstanz Information Miner (KNIME)
- Retrieval Augmented Generation (RAG), Low-Rank Adoption (LoRA)

General properties of local LLM solution

Command-line (CLI) and Graphic User Interface (GUI) options: Many local LLM solutions offer both CLI and GUI interfaces, providing flexibility for developers and users with different preferences and skill levels.

Multi-platform support (macOS, Linux, and Windows): To ensure broad compatibility and accessibility, leading local LLM solutions typically support major operating systems such as macOS, Linux, and Windows.

2.1. Omni-Layer Learning Language Acquisition Model (Ollama)

Ollama is an open-source tool that runs large language models (LLMs) directly on a local machine. This makes it particularly appealing to AI developers, researchers, and businesses concerned with data control and privacy.

Table 1. is the summarizes information about the useable Large Language Models for Ollama

Table 1. The useable LLM models for Ollama based on Ollama.com

LLM name	Main tasks	Preferred usage area
Llama	natural language processing (NLP)	text generation, summarization, machine translation
Mistral	code generation and large-scale data analysis	developers working on AI-driven coding platforms
Code Llama	programming-related	writing and reviewing code
LLaVA	processing text and images	generate accurate image captions, answer visual questions, combine text and image analysis
Phi-3	scientific and research-based applications	literature reviews, data summarization, scientific analysis

2.2. Gpt4All

Gpt4All provides three different packages: chat client User Interface (UI), a Python library, and a Python Command Line Interface (CLI) script in a terminal window (Nomic, 2024).

The chat client can be downloaded from gpt4all.io. When the package is installed, we are asked first to download a model. In the GUI, we can see a short explanation of each model. The download size is between 3GB and 7GB depending on the model. After downloading, we can start the chat.

Gpt4All is focused on improving the accessibility of open-source language models. Currently provides native support and benchmark data for over 30 models (such as Replit and Hugging Face). The API support is available for several programming languages including Python, Typescript, C#, and Java. (Anand et al. 2023),

2.3.LangChain

LangChain is a popular framework used in Python and Java. It is designed to simplify the development of applications using large language models (LLMs). It acts as a bridge between the application code and the LLM, providing a structured way to interact with them. There is a component that allows LLMs to access new datasets without retraining. The main components of the LangChain are the following, *models* (LLMs), *prompts* with prompt template, the *chains* (allow to link multiple LLMs calls), *indexes*, and *agents* (this systems use an LLM as a reasoning engine to determine which actions to take). The interoperable building blocks allow to building of end-to-end applications (LangChain, 2024).

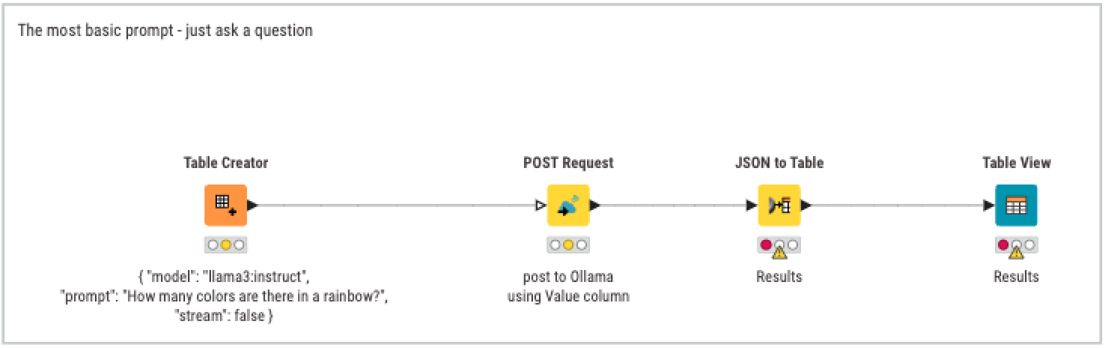
The LangChain framework offers practical guides for building various applications, including autonomous agents, chatbots, and systems for understanding code, extracting information, answering questions from documents, summarizing text, and analyzing structured data (Topsakal, Akinci, 2023). Also, the support of the LangChain Github community makes easier the development. LangChain is a popular framework in Python and Java languages.

2.4.KNIME: the Konstanz Information Miner

KNIME is a data analytics and data science tool that lets us build data workflows of any complexity with highly accessible, no-code, drag-and-drop visual programming (KNIME, 2025).

KNIME could provide effective support in identifying topics from small datasets and thus can be utilized as a support tool in thematic analysis (Fitkov-Norris et al, 2023). The built-in Machine Learning approaches in KNIME have good potential in NLP application (Ihrmark and Tyrkkö, 2023). Ihrmark and Tyrkkö highlighted that KNIME looks like an appealing solution for text analytics also.

Figure 1. is an example of the application of the Ollama in KNIME. The installed Llama3 model is used to answer a prompt.



```
{ "model": "llama3:instruct",
"prompt": "How many colors are there in a rainbow?",
"stream": false }
```

Figure 1. The most basic prompt - just ask a question under KNIME, based on https://hub.knime.com/mlauber71

Figure 2 is also an Llama3 model-based KNIME chat application where we can use the listed files as a corpus. We can set the date/time range to collect the relevant files. The selected and also timestamped files are used as input for the Llma3. The chat is saved to a table, for a future analysis.

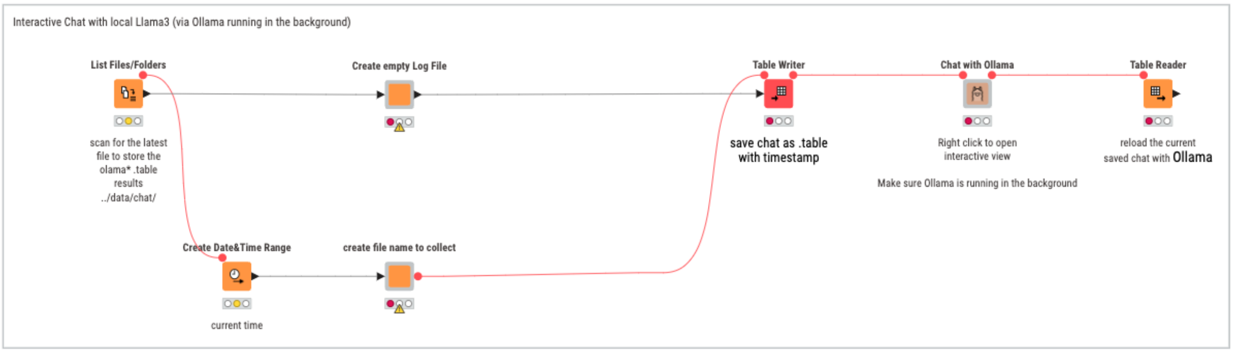

Interactive Chat with local Llama3 (via Ollama running in the background)
List Files/Folders
scan for the latest file to store the olama* .table results ../data/chat/
Create empty Log File
Table Writer
save chat as .table with timestamp
Chat with Ollama
Right click to open interactive view
Make sure Ollama is running in the background
Table Reader
reload the current saved chat with Ollama
Create Date&Time Range
current time
create file name to collect

Figure 2. Interactive Chat with local Llama3 under KNIME, based on https://hub.knime.com/mlauber71

2.5. Retrieval Augmented Generation (RAG), Low-Rank Adoption (LoRA)

As large language models (LLMs) continue to evolve, users are constantly seeking methods to improve their performance. Increasing the amount and variety of data used to train and query NLP models can lead to better results. By providing models with a broader range of information, they can better understand and respond to complex questions and tasks. However, it's important to note that simply adding more data is not always enough. The quality and relevance of the data are also crucial factors in improving model performance (Wu, 2024). Two prominent approaches that have gained significant attention are Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA). While both aim to enhance LLM capabilities, they do so through fundamentally different mechanisms. (Jeong, 2023)

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a popular approach to enhance the capabilities of large language models through external knowledge retrieval. RAG features two core components: the Retriever, which searches and retrieves relevant information from a knowledge base, and the Generator, a language model that produces text based on input and retrieved context. (Jeong, 2023)

Figure 3 the simplified logical structure of the RAG highlights the importance of the corpus-related answer, and the related documents.

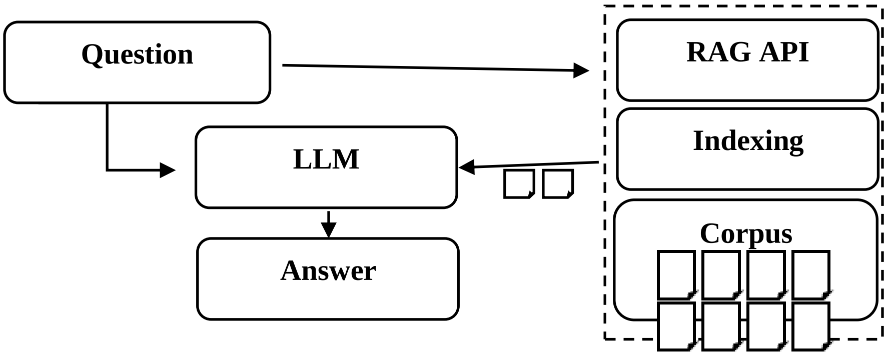


Figure 3. The simplified logical structure of the RAG based on Medium.com 2024a

Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique that adapts pre-trained language models to specific tasks or domains. It works by first freezing the weights of a pre-trained model (the original pre-trained model parameters remain unchanged) (Wu, 2024). The LoRA concept can be utilized in any type of model. The Hugging Face-based library is useable for fine-tuning LLM. Using *keras nlp*, implementing for LoRA is also available.

RAG vs LoRA

doi: 10.17700/jai.2024.15.2/733

RAG focuses on augmenting the model's knowledge through external retrieval. The RAG-Fusion chatbot was able to provide more accurate answers than traditional RAG models, so the RAG has a potential (Rackauckas, 2024). The choice between RAG and LoRA depends on the specific use case. For applications requiring access to current or extensive external knowledge, RAG is often the better choice. For efficient adaptation to specific domains or tasks, especially with limited computational resources, LoRA provides an excellent solution (Wu, 2024).

Table 2. There is a summarization of the mentioned open-source solutions

Table 2. The main conclusion about the enlisted open-source solutions

Name	Type	Main area	Development support
Ollama	Large Language Model (LLM)	General purpose, all round	https://ollama.com/
Gpt4All	Large Language Model (LLM)	General purpose, all round	https://www.nomic.ai/gpt4all
Langhcain	Framework	General purpose, all round	https://www.langchain.com/
KNIME	DataMining / Workflow	Mainly Data Science-related	https://hub.knime.com
RAG, LoRA	Software development technique	Optimization/ Efficiency	https://github.com/ specific Gits

From LLMs utilization (Ollama, Gpt4All) and the application framework LangChain, to the data mining platform KNIME and performance-boosting techniques (RAG, LoRA), this collection represents a diverse set of key technologies within AI and data science.

Conclusion

This article demonstrates the significant potential of local LLM solutions for Small and Medium-sized Enterprises (SMEs). These solutions can provide SMEs with a competitive edge, improve their operations, and open new avenues for growth. By developing and deploying custom, local LLMs, businesses can maintain control over their data, ensuring privacy and security while fostering innovation. Furthermore, the development and deployment of custom, local LLMs will empower businesses to maintain control over their data, ensuring privacy and security while fostering innovation. While challenges remain, such as the ongoing evolution of LLM technology and the need for skilled personnel, the potential benefits for SMEs are substantial. The key to unlocking these benefits lies in a combination of technological advancements, accessible resources, and a supportive ecosystem for SME development. ChatGPT (Datacamp, 2023) offers a non-free alternative, but it's a straightforward process to create a custom GPT. This paper highlights several open-source technologies that can significantly benefit businesses in the rapidly evolving AI landscape. Ollama is usable for local LLM, prioritizing data privacy and control. Gpt4All is an accessible collection of open-source LLM. LangChain is a framework for AI-powered LLM applications. KNIME is a data analytics platform for creating complex data workflows with AI components. The Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA) are advanced techniques for enhancing the performance and efficiency of large language models. Local LLMs offer a clear path for SMEs to harness the potential of AI while circumventing the obstacles of cloud-based solutions. This empowers them to navigate the intricate landscape of modern business and flourish in today's data-driven world.

References

Anand, Y., Nussbaum, Z., Treat, A., Miller, A., Guo, R., Schmidt, B., Duderstadt, B., & Mulyar, A. (2023). GPT4All: An Ecosystem of Open Source Compressed Language Models. *3rd Workshop for Natural Language Processing Open Source Software, NLP-OSS 2023, Proceedings of the Workshop*. https://doi.org/10.18653/v1/2023.nlposs-1.7

Datacamp (2023) https://www.datacamp.com/tutorial/how-to-make-custom-gpts, (Accessed on 12.12.2024)

Datacamp (2025) https://www.datacamp.com/podcast/did-genai-kill-nlp, (Accessed on 16.01.2025)

Fitkov-Norris, E., & Kocheva, N. (2023). Are we There yet? Thematic Analysis, NLP, and Machine Learning for Research. *Proceedings of the European Conference on Research Methods in Business and Management Studies*, *2023-September*, 93–102. https://doi.org/10.34190/ecrm.22.1.1616

Ihrmark, D., & Tyrkkö, J. (2023). Learning text analytics without coding? An introduction to KNIME. *Education for Information*, *39*(2), 121–137. https://doi.org/10.3233/efi-230027

Jeong, C. (2023). A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. *Advances in Artificial Intelligence and Machine Learning*, *3*(4). https://doi.org/10.54364/aaiml.2023.1191

KNIME (2025). https://www.knime.com , (Accessed on 07.01.2025)

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *Journal of Academic Librarianship*, 49(4). https://doi.org/10.1016/j.acalib.2023.102720

LangChain (2024). https://www.langchain.com, (Accessed on 12.12.2024)

Medium.com (2024a) https://medium.com/gitconnected/building-enterprise-ai-apps-with-multi-agent-rag-06356b35ba1a, (Accessed on 12.12.2024.)

Medium.com (2024b) https://medium.com/@stahl950/a-practical-guide-to-using-llms-as-an-sme-714d03e7ee7f (Accessed on 12.20.2024.)

Microsoft.com (2025) https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens, (Accessed on 06.12.2025.)

Nomic.com (2025) https://www.nomic.ai/gpt4all, (Accessed on 24.12.2024.)

Ollama.com (2025) https://ollama.com, (Accessed on 07.01.2025)

OpenAI.com (2025a) https://openai.com/chatgpt/pricing/, (Accessed on 06.12.2024.)

OpenAI.com (2025b) https://openai.com/api/pricing/, (Accessed on 06.12.2024.)

Rackauckas, Z. (2024). Rag-Fusion: A New Take on Retrieval Augmented Generation. *International Journal on Natural Language Computing*, *13*(1). https://doi.org/10.5121/ijnlc.2024.13103

Topsakal, O., & Akinci, T. C. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. International Conference on Applied Engineering and Natural Sciences, 1(1), 1050–1056. https://doi.org/10.59287/icaens.1127

Wu, H. (2024). Large language models capsule: A research analysis of In-Context Learning (ICL) and Parameter-Efficient Fine-Tuning (PEFT) methods. *Applied and Computational Engineering*, *43*(1). https://doi.org/10.54254/2755-2721/43/20230858

Yu, W. (2022). Retrieval-augmented Generation across Heterogeneous Knowledge. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 52–58, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-srw.7