Integration of Large Language Models with microcontrollers to optimize industrial automation processes

Nuru Dashdamirli

1 Doctoral student; Azerbaijan Technical University; Republic of Azerbaijan

Abstract.
Integration of Artificial Intelligence (AI) systems into microcontroller-based industrial automation systems can reshape the manufacturing, transportation, and energy production industries. This study explores different Large Language Model (LLM) applications in microcontroller based industrial automation processes; such as real-time control, adaptive decision-making, advanced sensor integration. The study also discusses challenges related to LLM implementation in microcontrollers and proposes potential solutions for overcoming them to drive innovation within modern industries.

Keywords:
artificial intelligence
microcontroller
automation
**Introduction**

Progression of AI systems has led to significant achievements in many different fields, including computer vision, natural language processing, and autonomous control systems. AI systems continue to evolve rapidly, and future AI technologies are expected to make even greater progress in new fields. It will play an increasingly significant role in shaping our world, as several types of automation systems have already started to take advantage of AI powered technologies. However, running most AI powered systems, especially LLMs, requires a significant amount of processing power and specialized hardware, limiting its direct applications on low-end hardware like single board computers or microcontrollers. Microcontrollers are excellent for managing and controlling all types of electronic devices and sensors. They are especially useful in small-scale systems or embedded applications due to their small size, low price, and energy efficiency. Microcontrollers can also integrate many unique peripherals in a single chip, making them ideal for automation and robotics applications. As mentioned, the implementation of AI powered solutions in any type of microcontroller system is a challenging task, and limited computing capabilities of microcontrollers makes it impossible to run any significant LLM application directly in a microcontroller. This paper will explore a potential solution for use of LLMs with microcontroller-based systems, such as Internet of Things (IoT) networks and industrial machinery.

**Research methodology**

It is well known that microcontrollers are an inseparable part of industrial automation due to their ability to control and monitor diverse types of equipment within industrial systems. Microcontrollers are especially useful for battery powered equipment because of their low power requirements. As researchers find new applications for AI systems each day, it is not surprising that there were already considerable achievements for application of AI in microcontrollers. One of the most well-known AI technologies in microcontrollers is TensorFlow Lite project, which allows machine learning models to be deployed in resource-constrained microcontrollers. Due
to limited working memory and storage limitation of microcontrollers it only allows to run exceedingly small models for specific applications, however, this technology can still be particularly useful for uses such as text detection, motion analysis, pathfinding and more [1].

Most used AI-powered tools today are Large Language Models. An LLM refers to an AI system that is designed for natural language processing, such as text generation and understanding. LLMs are typically trained on immense amounts of data, allowing them to learn the patterns in human languages. LLMs can generate coherent and contextually appropriate responses based on training data, making them valuable tools for text summarization, translation between different languages, code generation, and conversational AI systems. Since the first popularization of ChatGPT from OpenAI in 2023, open-source communities put considerable amount of effort into creating alternative open LLMs. Some of these open-source models can even outperform original ChatGPT models in one or multiple areas. Compared to commercial only models, open-source LLMs are especially advantageous in privacy and independency aspects, as they can be deployed locally even without an internet connection and there are no worries about confidential information collected by third party companies [2]. Unfortunately, it is not possible to fit any kind of useful LLM in a microcontroller for any purpose, as smallest low quality LLMs comes with hundreds of megabytes in size and requires computational power that even the most powerful microcontrollers would not come close to. As such, the only option to run LLMs in microcontrollers is to run them externally, so microcontrollers can make requests to them.

This paper proposes a solution to use microcontrollers with locally or remotely configured LLM processing servers, where controlled equipment can be a part of an IoT system, providing means for communication with microcontrollers over the network. Other viable solutions such as connecting microcontrollers to hardware that runs LLMs using low-level communication protocols are also analyzed. In conclusion, the study seeks to bring attention to the role that LLMs can play within microcontroller-based industrial automation processes.
by discussing challenges, proposing solutions, and exploring various applications.

**Integrations of LLM systems with microcontrollers**

The first approach that comes to mind when talking about LLM integration is using ChatGPT over internet with special API key obtained from OpenAI. However, this requires paid subscription for commercial usage and constant internet connection. It is also not suitable for confidential work, as data is being collected for further training of the models. Instead, LLM can be deployed and run on a local server or personal computer that is connected to the network. If there is an IoT network locally, network-supported microcontrollers or development boards based on microcontrollers such as ESP8266 and ESP32 could be used to be able to connect to the IoT network directly (Fig. 1). This can be done without the need for an internet connection, as not all IoT networks need to connect to the internet; even though IoT gets its name from the internet, it does not need the internet unless cloud services are used. However, the proposed method can still be applied in environments where computing resources are away and require internet connection, while fully allowing control over data confidentiality and server infrastructure.

![Diagram](image_url)

**Figure 1**

*Using LLMs on microcontrollers using IoT in industrial setting*
There are many open-source applications for running local LLM models, such as GPT4All, Ollama, LMStudio, AnythingLLM, JanAI and Faraday. Currently, one of the most user-friendly ways for running LLMs locally is using GPT4All, which is an open-source application with intuitive GUI interface that lets users with or without technical knowledge to run LLMs with ease [3]. GPT4All is a cross-platform application that provides a wide range of open-source models for download and use, some of which are also licensed to be used commercially. It also provides special API and bindings for high level programming languages like Python and JavaScript, allowing software developers to implement custom LLM solutions for their needs [4]. Ollama is another backend for running LLMs, which does not provide a GUI by itself, but instead has a CLI support with OpenAI compatible API scheme. Ollama also provides its own repository for ease of downloading latest LLMs for local use. Ollama APIs can be called over network without any special libraries, which makes it easier to integrate to any kind of application without limited by programming language choice [5].

After choosing a backend to run an LLM, appropriate model should be selected for the application, as using application specific or fine-tuned models increases quality and reliability of the responses. LLMs can be categorized for various criteria, such as training data, architecture type, file format, parameter size, quantization, context length, prompt template - these parameters should be considered carefully when selecting appropriate model. As requests and responses should be in plain text format, prompt engineering is also particularly important for getting results that can be handled without any issues.

In April of 2024, two very capable, general-purpose, small-sized open models were released: Phi-3-mini, a 3.8B parameter language model trained on 3.3 trillion tokens by Microsoft [6]; Llama3, 8B parameter language model trained on 15 trillion tokens by Meta [7]. These models are incredibly capable for today’s technology, and being smaller in size allows them to run on cheaper consumer hardware with very fast responses. These LLMs can be very useful for locally run intelligent automation systems. Applications for such LLMs in
microcontroller powered industrial systems can include:
- allowing process control with text-based commands in human languages;
- analyzing sensor data to predict equipment failures and schedule maintenance tasks;
- identifying defects in manufactured products based on inspection;
- assisting with collecting, tracking, and forecasting data for automation and fine-tuning;
- tracking key performance indicators and alerting for deviations from expected values;
- identifying potential issues with machinery and recommending corrective actions;
- detecting potentially dangerous conditions and alerting users.

Proposed techniques and approaches mentioned above can also be used with very small models that can be used on single board computers like Raspberry Pi, making it possible to implement limited LLM capabilities on small-scale industrial automation processes. Such models will have 1 to 3 billion parameters, planned to be used with at most 4-bit quantization. One example of such models is TinyLlama, which has 1.1 billion parameters trained with 3 trillion tokens [8]. TinyLlama 4-bit quantized GGUF model can run on Raspberry Pi 5 8GB RAM version with generation speed of approximately 10 tokens per second. Microcontroller development boards that have no networking abilities built-in such as STM32 Blue Pill board can connect to Raspberry Pi over low-level protocols and use LLMs like TinyLlama, controlling and monitoring equipment connected to it (Fig. 2).
However, as TinyLlama model is sufficiently small, it has much greater chance to hallucinate and make contextual mistakes. This makes it risky to use for critical equipment and automation processes. In such cases, Phi-3-mini could be used in the same hardware, for slower but more accurate responses, as 8GB RAM is enough to load this model without any issues.

**Conclusion**

Proper implementation of AI technologies in microcontroller-based systems is effective for industrial automation and its processes. However, since the energy and computing power of microcontroller-based systems are limited compared to traditional computers, the direct application of AI in microcontrollers is extremely limited. Therefore, a new locally or remotely configured AI server approach is shown in the research and a structure for its implementation is defined. This approach allows for the effective implementation of AI technologies such as LLMs and neural networks on microcontrollers with limited capabilities and resources. LLMs are especially advantageous for these applications as they are easy to deploy and integrate to various industrial processes, compared to other AI technologies, such as deep learning neural networks, which require extensive training. LLMs can also be fine-tuned and with the usage of prompt engineering they can be irreplaceable tools. However, it should be noted that it is not optimal to always use LLMs, as there are many instances that using plain neural networks or computer vision would be more feasible.

**References:**


data distillation from gpt-3.5-turbo. GitHub.


