Supporting Foundry Workers with an AI-Based Multimodal System Tunned Using Foundry Knowledge

Javier Nieves*, Andres Perez
Azterlan member of Basque Research Team Alliance (BRTA), Aliendalde Etxetaldea 6, 48200 Durango (BI), Spain
*Corresponding address: e-mail: jnieves@azterlan.es

Abstract: Currently, society is experiencing a new Artificial Intelligence revolution with the creation of Large Language Models (LLMs). However, these tools missed the specific and needed knowledge of a domain to be accurate. This paper presents a novel approach of using LLMs but including the knowledge generated by a research center. Hence, by providing access through a simple human interaction system, we present a possible ecosystem of models that combining textual and image knowledge will support day-by-day work. The system was tested and evaluated by foundry experts, assuring that this kind of tools could be a fundamental future consulting tool.

Keywords: artificial Intelligence; LLM; RAG; foundry knowledge; intelligent systems

1 Introduction

Although we have passed 4 different industrial revolutions, currently, we are in a new one, Industry 5.0., where the human being is the center of the system. The new trend in Artificial Intelligence, Large Language Models (LLMs) [1], makes possible the introduction of the human being into the Artificial Intelligence and Knowledge Ecosystem, providing and easy way to improve human-machine interactions. These models are a good but, if we are involved in a domain specific problem, we need to improve them.

In this way, when we want to adapt the specific knowledge to the aforementioned AI models, it can be done applying one of the following manners.

Create new models. The most expensive. It needs extensive work to get all data for learning and the energy consumption to generate it (i.e., learning the GPT-3 model in 2020 by OpenAI was close to 4.6 million dollars [2]).

Transfer learning[3]. Reuse the elements of a previously trained machine learning model in a new model. This technique is being widely used especially in machine vision systems [4].

Retrieval-Augmented-Generation (RAG)[5]. Optimizing the output of a model making use of a new knowledge base outside of the training data used for the creation for the model. This avoids the generation of false or unrealistic answers as well as obsolete.

Against this background, the solution proposed in this paper will focus on the generation of a multimodal system based on an LLM extended through the RAG technique with the expert knowledge of a research center such as Azterlan. The system presented introduces bibliographic knowledge through books, scientific knowledge through articles, technical knowledge through user manuals and, finally, visual knowledge about the real foundry analyses.

2 Experimental procedure

Once the desired style of our multimodal system has been determined, a working methodology was designed. More accurately, several models were built and integrated as we describe below.

Firstly, we selected an existing LLM to use it as a base. In this case, Meta's Llama3 was employed, but configuring its personality. The model must work as a researcher from our center, giving answers as it was an academic. The system will always answer any topic, but if it deviates from the defined field, it will indicate that the model is not made to answer those type of questions.

Subsequently, we generated 4 different models. These incorporate the knowledge of a book of our own production explaining different aspects of the sand, scientific articles on digitalization solutions developed by Azterlan and the user's manual of its thermal analysis system. Each one has been processed, extracted and stored in a vector database for RAG using the *Nomic Embed* model.

Then, we have generated a model for classifying metallographies by similarity. This is based on the weighted average of the 10 most similar ones extracted from the comparison of its vectorial representation created with *image2vec* model.

Finally, we have generated a routing agent. This system is responsible for determining to which model(s) the job should be sent. In this way, we have succeeded in creating the generalization of an agentic RAG multimodal system for assists in foundry related topics.

3 Result and discussion

The result of this research has been the generation of a multimodal AI assistant. This system is able to solve doubts related to sand (book added), digital twins and use of Deep Learning in the foundry (some papers included), the use of Thermolan® (manual of the thermal analysis system developed by Azterlan) and make classifications of nodular iron metallographies. Specifically, Figure 1 shows an example of a query related to sand problems.

For evaluating our development, several experts prepared relevant questions about the provided knowledge. These questions have been solved by the tool and they have

evaluated the answers. In summary, everyone thinks that results are good, although they can be improved adding more knowledge to the system.

The obtained AI assistant is a proof of concept. This is not the definitive supporting tool. To achieve it, it would need to face other problems:

Multi-Language capability, something that can be solved using another AI like NLLB from Meta.

Ability to manage speech questions. This problem can be solved, also, using Whisper (OpenAI *Speech-2-Text* free model).

Adding more knowledge to cover more tasks.

In addition, this proof of concept can be the basis for, later, increase the ecosystem with new models that our aforementioned router agent can integrate and redirect jobs to them.

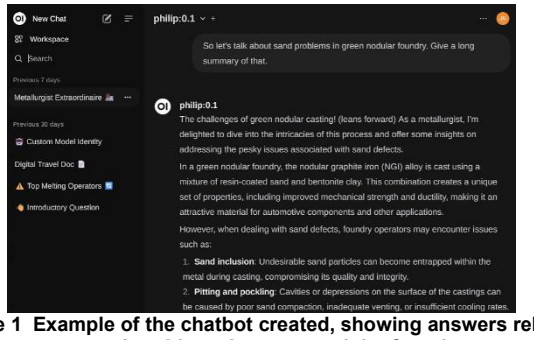

Figure 1 Example of the chatbot created, showing answers related to sand problems in green nodular foundry

4 Conclusion

Industry 5.0 has arrived, and we must facilitate the use of new technologies. Therefore, in this paper we present a multimodal system specialized in foundry that, through a simple communication, gives support to different tasks. The experts who analyzed it indicate that its quality is high and, by providing more data and more models, it will be very useful in the near future.

References

[1] ChangY P, WangX, WangJ D, WuY Y, Yang L Y, ZhuK J, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[2] https://fortune.com/2024/04/04/ai-training-costs-how-much-is-too-much-openai-gpt-anthropic-microsoft/.(29 June 2024).

[3] TorreyL, and ShavlikJ. Transfer learning. In handbook of research on machine learning applications and trends: Algorithms, methods, and techniques. IGI global, 2010: 242-264.

[4] SarrionandiaX, Nieves J, Bravo B, Pastor-López I, and BringasPG. An objective metallographic analysis approach based on advanced image processing techniques. Journal of Manufacturing and Materials Processing, 2023, 7(1): 17.

[5] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al.Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.