Development of an Intelligent Design and Simulation Aid System for Heat Treatment Processes Based on LLM

Yixiao Sun [1], Chao Liu [2], Xiaohu Deng [3], Jiangang Wang [4], Zeyu Zhang [3], Tengyang Wen [4]
Tianyu Song [4], Dongying Ju [5,6,*]

[1] *School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, China.*
[2] *IdeaMap, Co., Ltd.,Tokyo 101-0022, Japan;*
[3] *School of Mechanical Engineering; Tianjin University of Technology and Education, Tianjin 300222, China*
[4] *College of Material science and Engineering, Hebei university of science and Technology, shijiazhuang 050018, China;*
[5] *Saitama Institute of Technology, Fukaya 3690293, Japan;dyju.sitec@gmail.com*
[6] Tokyo Green Power Electric Technology, Co., Ltd., Tokyo 111-0022, Japan; ishikawa@gpower.jp (K.I.)

In this paper, an intelligent learning system that combines LLM with heat treatment simulation software is developed. The system can provide users with knowledge learning Q&A function. The user can propose performance requirements for materials and components, and the system can give the corresponding heat treatment process conditions and data files during heat treatment simulation by means of Q&A. The system encodes the knowledge of heat treatment of steel in vector form, stores it in the vector database Chroma, and applies dialogue tasks to match the user's questions with the vectorized knowledge, find the relevance, and input it into the gpt-3.5-turbo-16k model in combination with hint engineering. This approach helps the large language model to refer more to the knowledge in the knowledge base rather than the pre-trained knowledge to complete the knowledge and answer the query. In addition, as an attempt to apply it, CHATGLM-6B can convert user-given or recommended process descriptions into process input files that can be used directly by the heat treatment simulation software COSMAP.

Building a vector knowledge base is an effective approach that addresses the challenge of mastering expertise in the heat treatment domain faced by large language models. Compared with the traditional method of fine-tuning large language models using knowledge to achieve expertise, the vectorized approach allows intelligent expansion of the knowledge base, and the complete task from knowledge Q&A to recommendations to simulations can achieve direct recommendation of process effects, greatly simplifying the workflow of heat treatment practitioners and improving R&D efficiency. In addition, process results validated by COSMAP can help optimize and improve the knowledge base of the recommendation module through self-iterations. The test results demonstrate that this approach leverages the capabilities of LLM to provide accurate and fast answers to queries related to the heat treatment of metallic materials. interoperability between LLM and CAE software has been achieved. This work provides a new avenue for the development of expert knowledge base systems in the field of heat treatment of metallic materials and is expected to have applications in other fields as well.

Keywords:, *metal material heat treatment, expert knowledge base large language models, ,knowledge embedding , CAE software*

1. Introduction

As modern manufacturing continues to evolve, the role of heat treatment technologies for metal materials has become increasingly prominent across diverse industries. Such treatments not only bolster the mechanical attributes of these materials but also fortify their resistance to corrosion and wear. Despite these benefits, the design and optimization of heat treatment processes present inherent challenges, often necessitating extensive expertise, a wealth of knowledge, and long validation periods.

Recent advancements have seen the rise of Large Language Models (LLMs) like ChatGPT [1] , which have demonstrated remarkable proficiency in knowledge comprehension and generation across an array of domains. These models, benefiting from exhaustive pre-training on vast textual datasets, exhibit capabilities that often rival or even surpass domain experts.

However, there remains a palpable gap: while LLMs possess comprehensive generalized knowledge, their proficiency in niche domains, such as heat treatment, is often limited by their pre-training data. The challenge lies in harnessing the robust capabilities of LLMs while integrating domain-specific expertise, especially when providing recommendations for heat treatment procedures. Traditional knowledge repositories and query systems, with their inherent limitations, further compound the issue, often failing to address the dynamic needs of today's engineers and researchers.

In response to these challenges, we present the Intelligent Design and Simulation Aid System for Heat Treatment Processes based on LLM (CHAT-IMHST). This novel approach endeavors to provide a holistic, end-to-end solution. Spanning the spectrum from steel material knowledge retrieval and process optimization recommendations to seamless integration with simulation software such as COSMAP [2] , our approach endeavors to streamline the workflow of heat treatment processes.

2. Theory

2.1 System Framework and Design

The framework proposed in this study is adept at offering heat treatment process recommendations based on specific user requirements, as illustrated in Figure 1. The red part represents the user operation, and the blue part represents the system work. Once the user validates the suggested process, it is readily translatable into a process parameter file (POS) compatible with the CAE software, COSMAP. Subsequently, COSMAP is employed to authenticate the practical implications of these recommended processes. Simultaneously, leveraging the knowledge of LLM provides users with a comprehensive and coherent heat treatment knowledge Q&A workflow. The system facilitates multi-turn dialogues on heat treatment knowledge through a specially designed Q&A interface.

For the task of process recommendations, the system predominantly bases its suggestions on user-provided information, such as the hardness of steel parts, case-hardening depth, part dimensions, and intended applications. The enhanced chatgpt-3.5-turbo-16k model, bolstered by the vectorized knowledge base [3] , is harnessed to propose potentially apt heat treatment processes. The Q&A system mainly answers the user's questions about heat treatment knowledge, such as the definition of "heat treatment" or the specific process of "quenching" To enable a more efficient process parameter conversion, this study has also integrated LLM with the heat treatment simulation software COSMAP. Through the pretrained chatglm2-6b model, the system adeptly transforms descriptive process natural language into the POS file format, ready for invocation by the COSMAP software.

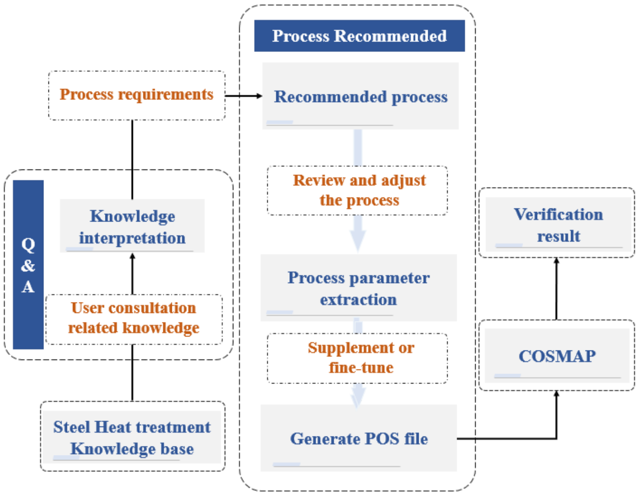

Figure.1 The framework of system and user interaction

2.2 Vector Knowledge Base and Few-Shot Query

Initially, we constructed a knowledge base tailored to steel material processing, as shown in Figure 2. The curated knowledge can be broadly categorized into two primary classifications. The first category consists of definitional entries, elucidating specific terminologies intrinsic to this domain, such as: "The exact procedure of quenching is...". The second category emphasizes process recommendations, where data from existing heat treatment manuals concerning steel material processing cases were collated and organized. This information was mapped into key-value pairs correlating steel materials, steel component hardness, carburization depth, component dimensions, use, and corresponding processes.

Subsequently, the knowledge base was vectorized using the "text-embedding-ada-002" method and stored in the vector database, Chroma. Whenever users pose queries, their questions are also vectorized and matched against the vectors in the knowledge base. Upon identifying the top N associated knowledge entries (where N is a user-defined parameter), both the knowledge and the query are relayed to the chatgpt-3.5-turbo-16k model. Harnessing the few-shot learning capabilities within Large Language Models (LLMs), by supplying the model with an array of pertinent knowledge, we ensured the comprehensive understanding of the specific nuances of heat treatment domain queries by the LLM, facilitating precise responses.

2.3 Description of affiliation(s)

In our methodology, the chatglm2-6b model functions as a pivotal information extractor, charged with extracting and encapsulating pivotal information from textual narratives.

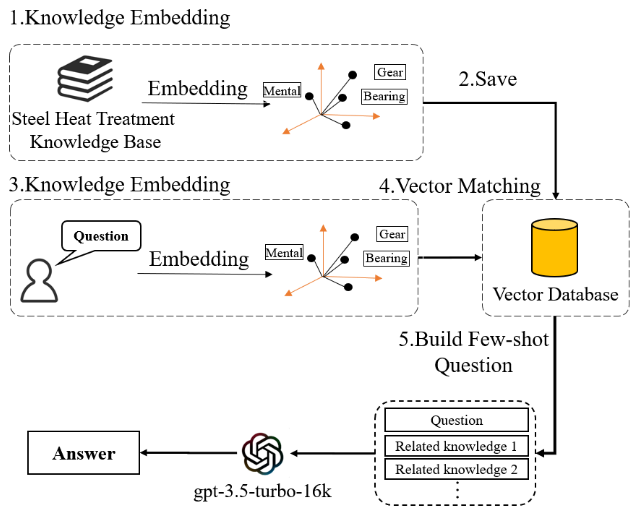

Figure.2 Vector knowledge base build and Few-shot Query

Whenever the model encounters incomplete information or requires more detailed data, it prompts the user to supplement the essential details. Once all the necessary data is gathered and verified, these critical details are input into a specially designed format conversion software.Through the aforementioned strategy, a seamless integration between the process recommendation task and COSMAP software validation was achieved, providing users with a coherent and streamlined workflow.

3. Experiment and Results

To validate the accuracy of our proposed system, we have devised an array of innovative techniques specifically tailored for evaluating the veracity of knowledge responses pertaining to steel heat treatment, as well as for assessing the aptness of process recommendations.

3.1 Knowledge Answering Task Testing

To more precisely assess the system's capability in steel heat treatment knowledge answering and verify the accuracy of the answers, we conducted two experiments, namely the Steel Heat Treatment Knowledge Understanding Test and the Expert Evaluation Test.

3.1.1 Steel Heat Treatment Knowledge Understanding

Drawing inspiration from MMLU [4] , we developed an evaluation method for steel heat treatment knowledge understanding (SHTKU). We first constructed 200 representative questions related to steel heat treatment knowledge, as shown in Figure 3. Questions were presented in either single-choice or multiple-choice format, with each question having four options. The tested language model was instructed to only provide the option number (a, b, c, d) as its answer. This test aims to determine if the model genuinely understands steel heat treatment. Additionally, an automated model evaluation system was developed.

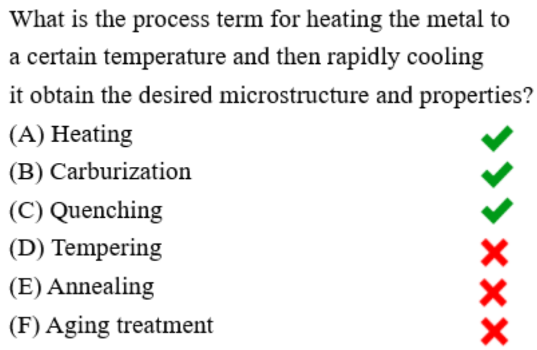

Figure.3 Question Example

3.1.2 Expert Evaluation Test

For this experiment, a set of questions and their corresponding answers were provided to five heat treatment professionals. These experts were tasked with manually evaluating the accuracy of the machine's responses. To ensure fairness in the evaluation, manually generated answers were mixed with the machine-generated answers to prevent bias against the machine-generated responses. Furthermore, the evaluators were asked to guess whether each answer was machine-generated or manually produced. This guess serves as a significant metric to evaluate whether the language model's generated answers gain human approval.

3.2 Process Recommendation Task Testing

We constructed heat treatment processes for five representative steel parts, processes that are not present in the heat treatment process recommendation knowledge base. During the testing process, the process requirements were first provided to the LLM. The language model then made its process recommendations. Subsequently, the transformed process parameter file, POS, was input into COSMAP for validation.

3.3 Result and Analysis

Based on the aforementioned experiments, we validated the performance of the proposed method in the steel heat treatment domain.

From a knowledge answering perspective, we first employed the SHTKU and expert evaluation methods to compare the performance of chatgpt-3.5-turbo-16k and chatglm2-6b models before and after vector enhancement of the knowledge base. This comparison validated the effectiveness of the vectorized knowledge base. Subsequently, we compared a chatgpt-3.5-turbo-16k model with part of its knowledge base removed against a version with a complete knowledge base. This comparison verified the importance of the completeness of the knowledge base to the quality of the model's responses. It also indicated that the proposed method possesses strong scalability; when there are updates in the knowledge domain, there's only a need to update the knowledge base, eliminating the necessity for retraining the model.

From a process recommendation perspective, the chatgpt-3.5-turbo-16k and chatglm2-6b models, before and after knowledge base enhancement, were compared. This comparison demonstrated that models enhanced with the process recommendation knowledge base in a vectorized format produced superior process recommendations.

4. Conclusions

With the rapid development of large language models, their potential applications in various fields are becoming increasingly apparent. This research introduces the "Intelligent Design and Simulation Aid System for Heat Treatment Processes based on LLM". Targeting the steel heat treatment industry, it showcases a system enhanced by knowledge vectors and powered by large language models. The system is not only capable of answering questions related to steel heat treatment but also recommends heat treatment processes for specific requirements. Furthermore, it seamlessly integrates the recommended results with the renowned CAE system, COSMAP, directly outputting the required POS files for the system.

Experimental results demonstrate that the vectorization of the knowledge base significantly enhances the model's answering quality and accuracy. Moreover, compared to traditional models, the knowledge-enhanced model exhibits superior performance in process recommendations. The success of this system validates that integrating expert knowledge with large language models is an effective approach in specific domains. Additionally, when knowledge in the steel heat treatment domain is updated, there's no need to retrain the entire model; updating the knowledge base suffices, enhancing the system's scalability and flexibility.

In summary, this study not only offers valuable insights academically into the application of large language models in specialized domains but also holds significant practical implications.

References

1.Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. and Schulman Training language models to follow instructions with human feedback. *Neural Information Processing Systems*.2022,35,27730-27744.

2.Liu CC, Ju DY, Inoue T. A numerical modeling of metallothermo-mechanical behavior in both carburized and carbonitrided quenching processes. *ISIJ International*.2002,42(10),1125-1134.

3.Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Neural Information Processing Systems*.2020,33,pp.9459-9474.

4.Jiang, Z., Xu, F.F., Araki, J. and Neubig, G. How can we know what language models know?. *Transactions of the Association for Computational Linguistics*.2020,8, pp.423-438.