Framework Design and implementation of the AI Aided Process Designing Platform for Shipbuilding Industry

Xiaoyang Liang 1,2
1.Shanghai Dongxin Software Engineering Co., Ltd
2.Hudong Zhonghua Shipbuilding (Group) Co., Ltd.
Shanghai, China
NO.2789, Pudong Avenue, Pudong New District, Shanghai
790751725@qq.com

Yanjun Ma[1,2]
1.*Shanghai Dongxin Software Engineering Co., Ltd*
2.*Hudong Zhonghua Shipbuilding (Group) Co., Ltd.*
Shanghai, China
NO.2789, Pudong Avenue, Pudong New District, Shanghai
kfmyj@163.com

Meng Xi
Hudong Zhonghua Shipbuilding (Group) Co., Ltd.
Shanghai, China
NO.2789, Pudong Avenue, Pudong New District, Shanghai
mengr8@126.com

Fei Li
School of Mechanical Engineering, Donghua University
Shanghai, China
No. 2999 Renmin North Road, Songjiang District, Shanghai
1229077@mail.dhu.edu.cn

***Abstract*—In order to achieve digital upgrading of traditional shipbuilding industry， framework design and program development of the artificial intelligence (AI) Aided Process Designing Platform have been completed. The platform is based on B/S architecture. By utilizing retrieval-augmented generation （RAG） technology and Large Language Model (LLM), four major function functional modules was implemented, including process knowledge integration, Knowledge Engineering Management, intelligent question and answer (Q&A) for process knowledge, and process plan generation. The retrieval accuracy of the AI Aided Process Designing Platform was proved to reach over 95%, while the accuracy of Q&A reached over 74% by manual testing, which can help technicians save query time and Improve design efficiency. Besides, process plans can be automatically composited rapidly through the templates. Therefore, the AI Aided Process Designing Platform can help shipbuilding enterprises improve the economic benefits and enhance the competitiveness.**

Keywords—AI application, LLM, RAG, shipbuilding, Framework Design

I. INTRODUCTION

With the rapid development of LLM technology, an important way for industry growth and innovation turns out to be more and more practical when combining the newest AI technology with the traditional domain. Based on the 3 core capabilities of LLM, i.e, deep learning, interactive assistance, and inference decision-making, it is accessible to empower the productivity of PLM-oriented ship product process, including design, manufacture, management,operation and maintenance service [1,2]. At present, LLM is widely applied in multiple industrial fields, which can improve the efficiency and effectiveness.Yaz Dinejad et al. [3] applied LLMs to drug supply chain management (DSCM). They proposed a technical proposal for user data to make an adaptive data governance and conversational data interaction, which was based on the capacities of LLM and development paradigms of the intelligent agent. Chen et al. [4] utilized LLM to extract information from locomotive maintenance data and developed a efficient pre-processing tool that encapsulates a basic LLM to standardize data. São Paulo State University (UNESP) proposed a LLM called PetroBERT [5] for the oil and gas field based on the Bidirectional Encoder Representations from Transformers (BERT) model. Fine tuning of PetroBERT was performed by named entity recognition and sentence classification on a private dataset in the vertical domain , which improved the accuracy of knowledge question answering. The researchers in ExxonMobil proposed the customLLM system [6], which introduced equipment manuals, work orders, and maintenance data to pre-train the model base. And the performance of the system in professional tasks was optimized by domain markers. In addition, LLM technology has also been explored and applied in fields such as nuclear power, chemical engineering, aviation, and automobiles[7-10], achieving fine results in knowledge graph creation, fault diagnosis, equipment operation and maintenance etc.. But generally speaking, the integration of LLM in industry field is still in its early stages of development. The overall technological level is not mature enough, and accuracy achieved needs to be improved.

Compared with developed countries, the exploration and application of LLM technology in the shipbuilding industry has not yet truly begun in China. New technologies are still disconnected from traditional shipbuilding enterprises, for it is difficult for these enterprises to find a focus on integrating LLMs. Currently, there is a technical gap for LLMs converge on the vertical field of shipbuilding in China. Therefore, investigation on LLM to meet the real business needs of shipbuilding enterprises is carried out in this research. The framework of the AI Aided Process Designing Platform for Shipbuilding Industry (APDPSI) has been designed, and program implementation has been basically accomplished. By manual testing, it is proved that APDPSI have fine

2024 IEEE First International Conference on Data Intelligence and Innovative Application (DIIA) | 979-8-3315-0443-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/DIIA62678.2024.10871768

979-8-3315-0443-4/24/$31.00 ©2024 IEEE

understanding and generating capabilities in the vertical field, which can innovate the development process of ship products and promote the digital upgrading of traditional shipbuilding industry.

II. FRAMEWORK DESIGN

A. Technical Framework

The technical framework of APDPSI system is shown in Fig. 1. By Combining with LLM fine-tuning and RAG technology, APDPSI system was established to meet the efficient knowledge sharing needs of shipbuilding enterprises. LLM was fine tuned by feeding professional knowledge in the vertical domain of ships. Different types of local knowledge databases were constructed based on the large number of structured and unstructured documents. Multiple heterogeneous documents, including Word, PDF, CSV, etc., are embedded to structure, graph and vectorize external retrieval data. When users input questions to APDPSI after selecting special knowledge base , the system would conduct question embedding, storing, and retrieving to obtain the context of the relevant answers in the knowledge databases. By setting Prompt templates, the context is transmitted to LLM as a prompt to achieve question answering with multi-modal document and multi document retrieval. During the Q&A process of APDPSI, users can rate the accuracy of the answers. A memory module was set up to store historical Q&A records , scoring results, search results, etc. These stored information could be taken as domain knowledge introducing into the ship professional corpus for further fine-tuning of LLM.

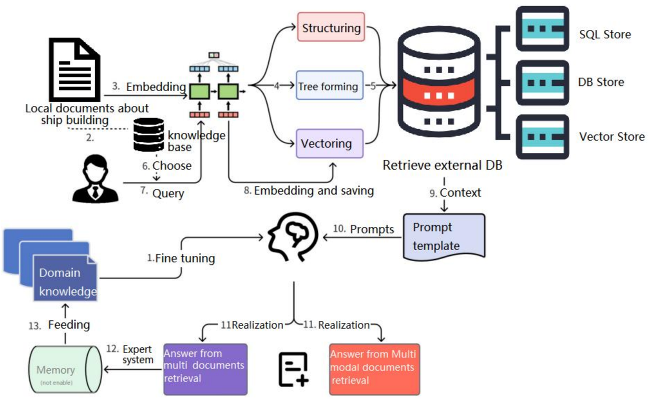

Fig. 1. Platform architecture combining LLM fine-tuning and RAG technology.

B. Three-tier Architecture Design

The business flow of the APDPSI is shown in Fig. 2. Two major tasks would be completed by the platform, of which one is performing intelligent Q&A in the vertical field of the shipbuilding industry, and the other is build a knowledge base for users. Users could submit questions to the platform, and the platform can quickly and accurately provide professional answers. Multimodal source documents, such as ship design manuals, guides, instructions, reports, etc., could be uploaded to the platform, where they would be processed into the structured local knowledge base.

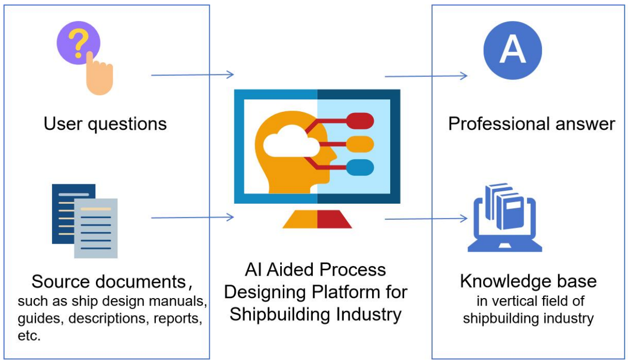

Fig. 2. Traffic analysis of three-tier architecture design.

Based on the technical framework construction and business flow analysis above , the three-tier architecture pattern [11] was adopted to design the platform in the B/S system. The three-tier architecture design pattern is a widely used architectural pattern in software development, particularly for building web applications and user interfaces. As shown in Figure 3, the Data Access Layer (DAL) stores and manages data from knowledge files, vector libraries and other databases with platforms such as OSS, MySQL, MongoDB, FAISS, etc. The Business Logic Layer (BLL) is mainly divided into AI Q&A module, processing tools module, knowledge management module and other general modules that support the common operation of the platform. By calling relevant functions and interfaces, intelligent Q&A, knowledge space construction, and user permission control can be normally performed with LLM. The User Show Layer (USL) mainly uses a web framework to create an interactive and friendly web interface for users to upload source files, input questions, search information and check professional results.

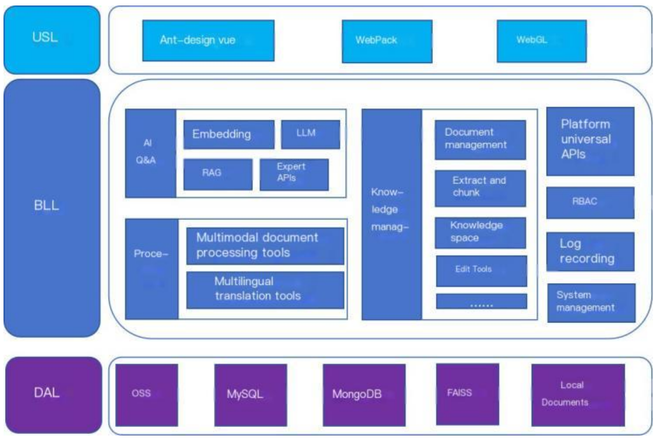

Fig. 3. Framework based on B/S and three tier mode.

III. LLMS SELECTION

As a crucial part of the entire platform design, the selection of the Basic LLMs directly affects the efficiency, performance, and user experience in handling specialized issues in the vertical domain of ships. Regarding the technical requirements above, the selection of the basic LLMs should follow the principles below.

- The basic LLMs should be open-source models with high trainability and strong convergence. Model parameters can be effectively learned and optimized through training and tuning.

- The basic LLMs has good comprehension and generation ability for Chinese. Besides, they can handle flexible and complex Chinese grammar, as well as lengthy paragraphs.
- The basic LLMs should show good performance in knowledge Q&A in the industrial and mathematical fields. Therefore, training time and resources can be saved by choosing this type of basic LLMs as a starting point.
- The basic LLMs should have moderate complexity and generalization ability, and their performance, efficiency, and accuracy can meet practical business needs.

Some recently published ranking lists of performance evaluation of the LLMs were refered to select the base models, which are authoritative and comprehensive. Related evaluation tools and systems about these ranking lists involve AlignBench [13], MTBench [14], MMLU [15], GSM8K [16], Math [17], HumanEval [18], C-Eval [19]. And the evaluation indicators not only include matching degree, similarity, accuracy, precision, the ability of long text generation, etc., but also include understanding and reasoning performance in humanities, social sciences, engineering, and other professional fields. The average score for evaluation can characterize the comprehensive performance of LLMs in all aspects. Combining the selection principle above, four open-source models called LLama-3-8b-instruction, Qwen2-72B-instruction, Qwen2-7B-instruction, and Chatglm-4-9B-chat have entered the selection range. Their average scores in the evaluation ranking lists above are shown in Table I and Fig. 4.

TABLE I. BASE LARGE MODEL PUBLIC TEST SCORE

Items	**Align-bench**	**MT-bench**	**MMLU**	**GSM8K**	**Math**	**Human-Eval**	**C-Eval**
LLama-3-8b-instruction	6.2	8.05	68.4	79.6	30	62.2	45.9
Qwen2-72B-instruction	8.27	9.12	82.3	91.1	59.7	86	82.8
Qwen2-7B-instruction	7.21	8.41	70.5	82.3	49.6	79.9	77.2
Chatglm-4-9B-chat	7.01	8.35	72.4	79.6	50.6	71.8	75.6

Funding agency: Shanghai Changxing Island Development and Construction Management Committee.

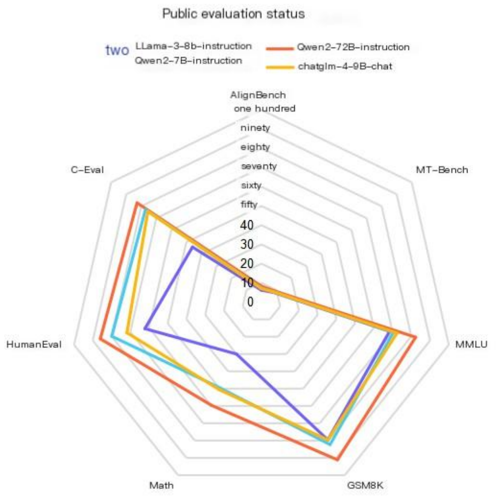

Fig. 4. Radar chart of the public beta performance of the large basic LLMs.

Besides, in order to deeply evaluate the quality of Q&A and interactive experience of these four open-source models, a local knowledge base was created based on one Chinese technological report in ship industry filed (with a word count of 7230 words, equivalent to approximately 4100 tokens). The inference speed and concurrency efficiency indicators were manually measured, specifically the output speed of the decoding stage and the maximum concurrency during program execution. The software and hardware environment and testing conditions for testing are shown in Table II.

TABLE II. TESTING SOFTWARE AND HARDWARE ENVIRONMENT AND TESTING CONDITIONS

Classification		Items	Information
Testing hardware environment	Server	Type	Power Edge R750
		CPU	Intel Xeon Gold 5317 3G, 12C/24T, 11.2GT/s, 18M cache, Turbo, HT (150W) DDR4-2933
		Additional CPU	Intel Xeon Gold 5317 3G, 12C/24T, 11.2GT/s, 18M cache, Turbo, HT (150W) DDR4-2933
		GPU	A800×2
Testing software environment	Server/Test end	System	Ubuntu20.04
		Software	Anaconda, LangChain, Pytest, Cucumber, etc
Network			Gigabit Internet
Testing conditions	Inference speed		Under the condition of single concurrency, the system inputs 500-600 tokens and outputs ≥ 200 tokens. The inference parameter *temp* was set to 0, and all other parameters were default values.
	Concurrent efficiency		The system inputs 2000~2300 tokens and outputs ≥ 400 tokens, with all parameters set to default values.

The performance of the four basic LLMs in framework of APDPSI is shown in Fig.5. The performance of Qwen2-72B-instruction, Qwen2-7B-instruction and Chatglm-4-9B-chat were better, with inference speeds exceeding 60 tokens/s and a maximum concurrency of over 100 during testing. Among them, Qwen2-72B-instruction has the fastest paragraph generation speed, reaching 97.01 tokens/s and a maximum concurrency of 106. Meanwhile, Chatglm-4-9B chat achieved a maximum concurrency of 123, with a relatively slow inference speed of 69.80 tokens/s. LLama-3-8b instruction performed poorly in Chinese environment testing, with an inference speed of only 40.12 tokens/s and a maximum concurrency of 92. And a preliminary evaluation of the output answers was conducted. It is proved that compared with the other three models, the answers generated by LLama-3-8b-instruction contained more redundant sentences with no relationship to the original text. Moreover, there were more errors in the numerical part of the response, resulting in poor overall performance. Therefore, Qwen2-72B-instruction, Qwen2-7B instruction, and Chatglm-4-9B-chat were ultimately selected as the basic LLMs for APDPSI.

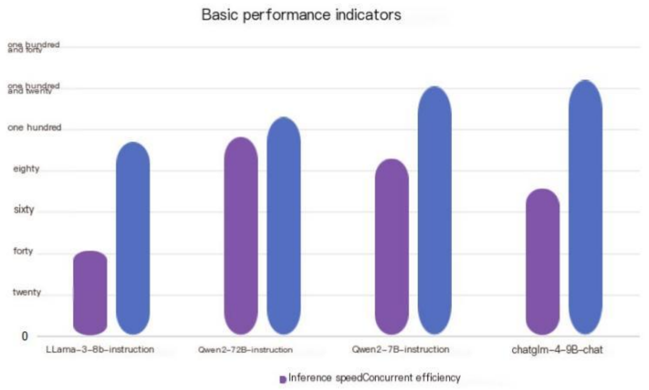

Fig. 5. Test results of inference speed and concurrency efficiency of the large basic LLMs.

IV. FINE-TUNING TECHNOLOGY

Framework of fine tuning training for LLMs is shown in Fig.6. Fine-tuning training for LLMs in the shipbuilding industry vertical domain was achieved in the following steps. Firstly, fine grained segmentation, professional word extraction and data organization of question and answer sets were carried out on the existing knowledge documents. Secondly, corpus alignment and verification were made to support the training process of cross domain corpus embedding. Then semantic classification of ship profession were built in LLMs. Thirdly, fast training system for Qwen2-72B-instruction, Qwen2-7B-instruction and Chatglm-4-9B-chat were designed with Prompt Tuning method, P-Unity v2 method, QLoRA method, UniPELT method, Adapter Tuning method, etc. Fine tuning training is completed based on hyperparameters, loss functions, position encoding techniques and other ways. Finally, PPO method was adopted in feedback training process of LLMs.Continuously fine-tuning and optimizing would further enhance the understanding and reasoning ability of APDPSI in ship vertical domain.

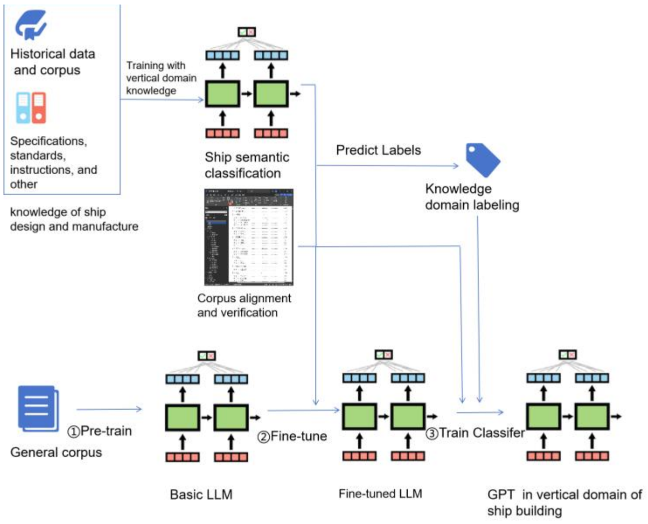

Fig. 6. System Framework of Fine-Tuning Training for LLMs.

V. RAG TECHNOLOGY

RAG enables LLMs to leverage domain-specific knowledge bases to improve the accuracy and richness of text generation. When responding to user queries, RAG first utilizes a retrieval system to extract relevant content from a domain-specific knowledge base, and then inputs the retrieved content along with the original query to the LLM. This allows the LLM to access the latest information, addressing issues such as knowledge hallucination and lack of domain-specific knowledge, and producing reliable outputs. As shown in Fig.7, this study leverages RAG technology to analyze the heterogeneous and multimodal data structures in ship design field. By integrating vector store technologies such as Milvus, FAISS, Chroma, and Qdrant, we establish a technical knowledge base that supports various indexing methods, including composite indexing and compound indexing, to improve the efficiency and effectiveness of RAG's information retrieval. Considering the diverse representational forms of multimodal information, we explored multimodal understanding and associations within the dataset. Then, we employed a primarily automated alignment approach, supplemented by manual alignment, to enable the intelligent generation of prompts for multimodal information, including text, images, and tables, during the question-answering process. Furthermore, by recording, scoring, inferring, and predicting the context of ship-related professional knowledge, we addressed the issues of data reuse and completion in the question-answering process, achieving automatic data completion and error correction, and enhancing the completeness and accuracy of the answers.

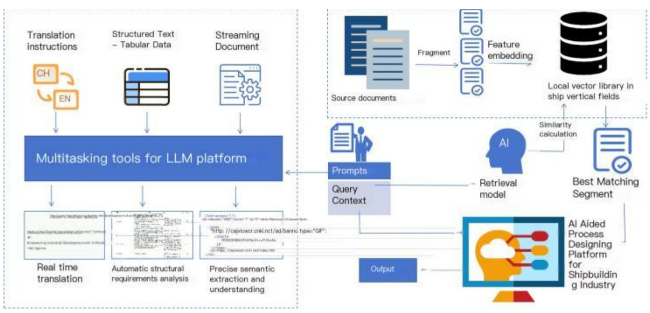

Fig. 7. Framework of RAG.

VI. EXPERIMENTS

A. GUI of APDPSI

The functional interface of APDPSI mainly includes login interface, knowledge management page, AI Q&A page, as is shown in Fig. 8. The Knowledge Management Page of APDPSI (shown in Fig. 8-b) provides functions for users to build new knowledge space or choose existing one, where they can perform document uploading, list display, vector embedding, database storage, deletion, document content viewing, and document selection. The AI Q&A Page (shown in Fig. 8-c,) allows users to create new chats and select the special scope of knowledge resource. This module implements the domain-specific knowledge question-answering functionality by constructing the knowledge sources from the entire knowledge base. Based on the user's ship-related professional questions, APDPSI can retrieve relevant question-answering text blocks, multi-level titles, images, tables, and related document sources from the knowledge sources. The Content box of checking knowledge sources, as shown in Fig. 8-d, provides functionalities such as document source viewing, image browsing, and original text copying, etc.

a) The login interface of APDPSI.

b) Knowledge Management Page of APDPSI.

c) AI Q&A Page of APDPSI.

d) Content box of checking knowledge sources Page of APDPSI.

Fig. 8. The Interactive Interface of APDPSI

B. Manual testing results

1) Testing procedures

To evaluate the domain-specific knowledge question-answering performance of APDPSI for the shipbuilding industry, a human evaluation method was employed. The evaluation process is as follows:

a) A "manual" Knowledge Space was established by uploading ship design manual documents, here we uploaded 30 word-formatted documents, with a total size of 256M.

b) In the evaluation process, 100 question-answer sets were designed by test personnel based on the source files. The same question was put into different LLMs by manual switch, the answers generated by each model, and the corresponding reference text segments from the source documents (i.e., the knowledge sources for the answers), as well as any exceptional situations such as errors, crashes, or lags that occurred during the answer generation process were recorded. The recording template is shown in Fig. 9.

c) The test personnel evaluated the correctness of the answers and the relevance of the retrieved knowledge sources. Then they gave their scores of the tested LLMs based on the quality of the generated answers on a scale of 100 points, with higher scores indicating more correct and useful information.

Fig. 9. The test record of APDPSI.

2) Evaluation System

After conducting the testing steps above, the Q&A results for the shipbuilding industry from APDPSI were recorded. The original results were analyzed and summarized. Therefore, the answers were divided into the following five categories based on their accuracy and quality. And the scoring rules for each category are shown in Table III.

TABLE III. TABLE TYPE STYLES

Grade	Answer situation	Score
	① No answer generated.	

I	② Completely incorrect answer. ③ The accuracy rate of useful information in the answer ≤ 90%. ④ Score by user ≤ 30.	0
II	① The answer has no professional errors but redundant text, useful information ratio ≤ 20% , and score by user ≤ 30. ② The answer shows high generality, useful information ratio ≥ 80% , and non-principled error ratio ≤ 5% . ③ Score by user ≥ 80.	0.2
III	① The answer has no professional errors, but the text is redundant and the useful information involved is incomplete. ② Score by user ≥ 60.	0.5
IV	① The answer has no professional errors and shows high generality, with useful information ratio exceeding 60% . ② Score by user ≥ 80.	0.8
V	① The answer has no professional errors and shows high generality, with useful information ratio exceeding 90% . ② Score by user ≥ 90.	1

Although authoritative platforms have published test reports and guidelines on the industrial applications of LLMs [20-23], the evaluation systems used in these studies tend to be overly generalized. The discussions on accuracy rates are often vague, and the assessments of model performance on domain-specific knowledge-based question answering are not comprehensive enough. Furthermore, the professional evaluation capabilities of these studies are limited. To address these gaps, the present research develops two parameters, retrieval accuracy (RA) and accuracy of Q&A (AA), to serve as specialized metrics for evaluating the performance of large language models in the context of the shipbuilding industry. The formulas for these two metrics are as follows：

$$RA = \frac{M}{N} \times 100\% \tag{1}$$

$$AA = \frac{\sum_{0}^{i=N} S_i}{N} \times 100\% \tag{2}$$

where, N represents the total number of test questions, M denotes the number of questions for which the model can accurately locate the highly relevant reference passages, and *Si* is the score for each individual response.

3) *Experimental Results and Analysis*

Based on the preliminary experimental results, the following conclusions can be drawn:

a) The big language model has 100% completed the understanding, analysis, and structuring of all source documents.

b) Due to the relative independence of the RAG framework and prompt engineering from the LLM modules, the three LLMs showed a high degree of consistency in their knowledge source retrieval performance. Out of the 100 question-answer pairs generated, 97 were able to accurately locate the highly relevant reference passages, while 3 were directed to the same incorrect passage, resulting in a RA of 97%.

c) By switching among Qwen2-72B-instruction, Qwen2-7B-instruction, and Chatglm-4-9B-chat, and testing the platform with the same question-answer set, the generated answers were classified according to the metrics in Table III, and the AA of Q&A values were calculated using Equation (2). The results are shown in Fig. 10. The calculated AA values for Qwen2-72B-instruction, Qwen2-7B-instruction, and Chatglm-4-9B-chat were 79.5%, 78.8%, and 74.9%, respectively. Among them, Qwen2-72B-instruction showed the best overall performance, with 86 answers scoring above 80. Chatglm-4-9B-chat performed the poorest, with 22% of the low-score (≤ 0.5) answers and one instance of an unanswerable question. All the LLMs exhibited varying degrees of hallucination during the answer generation process. More than 20% of the low-score (≤ 0.5) answers from the three LLMs occurred in question-answer pairs involving mathematical operators and Greek letters, indicating that the platform's capabilities in handling physics and mathematics-related domains require further optimization. It is worth noting that the samples that failed to successfully retrieve the knowledge source were not assigned a score of 0, suggesting that the LLM training process included relevant domain-specific data, providing a certain level of generalization and robustness.

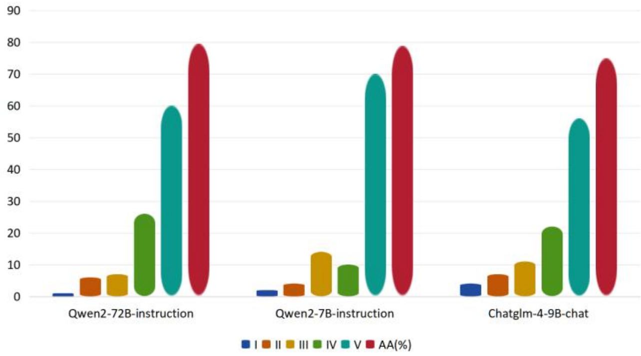

Fig. 10. Q&A test results and AA values.

VII. CONCLUSIONS

Framework design and implementation of the AI aided process designing platform for shipbuilding industry have been successfully completed to meet the business needs of knowledge management in vertical field. The proper basic LLMs have been selected according to the scores in public ranking list online and actual performance during runtime. By establishing multiple rapid training systems, fine tuning has been conducted on the basic LLMs. Meanwhile, the performance of the AI platform in specific professional knowledge domain tasks has been improved by utilizing RAG technology . Based on the technical requirements of professional Q&A in vertical knowledge domain, the manual evaluation system of APDPSI has been designed. And the Q&A performance of APDPSI in the vertical domain of ships was tested under this system. The test results showed that the retrieval accuracy reached 97% and the accuracy of Q&A exceeded 74%. In summary, APDPSI can well handle knowledge managing and sharing in the vertical field of ship industry. It has strong knowledge retrieval capabilities for understanding and generating professional and meaningful answers, which can provide intelligent assistance for ship design and construction.

ACKNOWLEDGMENT

This work is supported by Shanghai municipal government project fund (NO. JMRH-2024-1001). Besides, the authors are grateful to Qi Wen for his help with the translation and review.

References

[1] Chang Y , Wang X U , Wang J ,et al.A Survey on Evaluation of Large Language Models.ACM transactions on intelligent systems and technology, 2024(3):15.

[2] Zhang, Yaqi , et al. "MotionGPT: Finetuned LLMs are General-Purpose Motion Generators." ArXiv abs/2306.10900(2023).

[3] YU Tao, WANG Yipeng, LUO Qingquan, et al, "Preliminary Exploration for a Large Language Model-based User Characteristics tereoscopic Application Paradigm in New Power System",High Voltage Engineering,Vol.50, No.7: 2833-2848,July 31, 2024

[4] CHEN Ao, LI Chen, YAN Jiayun, et al. Specialized Large Language Model for Standardization of Locomotive Maintenance Data.Control and Information Technology, 2024(3): 72-79.

[5] RODRIGUES R B M, PRIVATTO P I M, DE SOUSA G J, et al. PetroBERT: A domain adaptation language model for oil and gas applications in Portuguese//PINHEIRO V, GAMALLO P, AMARO R, et al. Computational processing of the Portuguese language. Cham: Springer, 2022: 101-109.

[6] ABIJITH P Y, PATIDAR P, NAIR G, et al. Large language models trained on equipment maintenance text. SPE 216336-MS, 2023.

[7] ZHANG Caike, LI Xiaolong, ZHENG Sheng, et al. Research on the Construction and Application of Knowledge Graph Based on Large Language Model . Journal of Frontiers of Computer Science and Technology. https:// link.cnki.net/urlid/ 11.5602.TP.20240729.1002.002

[8] LIU He, REN Yili, LI Xin, et al. Research status and application of artificial intelligence large models in the oil and gas industry[J]. Petroleum Exploration and Development, 2024, 51(4): 910-923.

[9] Zeng Kang,Ye Jianyuan,Cun Xueling, et al."The Challenges and Exploration in the Application of Generative LLM Key Technologies for the Civil Aviation Maintenance".AVIATION MAINTENANCE & ENGINEERING,2024,(01):20-24.DOI:10.19302/j.cnki.1672-0989.2024.01.027.

[10] [Li Yong."LLM development strategy research". Xi'an University of Architecture and Technology, 2014.

[11] Beijing, et al. "Application Study of B/S based on MVC Design Pattern." (2006).

[12] Chang Y , Wang X U , Wang J ,et al.A Survey on Evaluation of Large Language Models.ACM transactions on intelligent systems and technology, 2024(3):15.

[13] Liu,X, Lei,X Y, Wang,S Y,et al . AlignBench(online). https://llmbench.ai/align

[14] Bai, G and Liu, J , Bu X Y ,et al. MT-Bench-101(online). https: //github. Com /mtbench101 /mt- bench -101.

[15] Hendrycks D , Burns C , Basart S ,et al. Massive Multitask Language Understanding (MMLU) (online). https://nlp.stanford.edu/helm/v2-lite-finch/?groups=1

[16] OpenAI. GSM8K(online). https://github.com/openai/grade-school-math

[17] Hendrycks D. The MATH Dataset(online). https://github.com/ hendrycks /math .git

[18] OpenAI. HumanEval(online). https://github.com/openai/human-eval

[19] [19]Chang Y , Wang X U , Wang J ,et al. Cevalbenchmark(online). https:// github.com /hkust-nlp/ ceval