Lunar Twins: We Choose to Go to the Moon with Large Language Models

Xin-Yu Xiao[1,5], **Yalei Liu**[1], **Xiangyu Liu**[2], **Zengrui Li**[3], **Erwei Yin**[4], **Qianchen Xia**[1,6*]
[1]The Future Laboratory, Tsinghua University [2]Sichuan University
[3]Beijing Institute of Technology [4]Tianjin Artificial Intelligence Innovation Center
[5]University of Chinese Academy of Sciences [6]National Key Laboratory of Human Factors Engineering
qianchenxia@tsinghua.edu.cn

Abstract

In recent years, the rapid advancement of large language models (LLMs) has significantly reshaped the landscape of scientific research. While LLMs have achieved notable success across various domains, their application in specialized fields such as lunar exploration remains underdeveloped, and their full potential in this domain has yet to be fully realized. To address this gap, we introduce **Lunar Twins**, the first LLMs designed specifically for lunar exploration, along with a collaborative framework that combines both large and small models. Additionally, we present **Lunar_GenData**, a multi-agent collaborative workflow for generating lunar instructions, and establish the first specialized lunar dataset, which integrates real data from the Chang'e lunar missions. Lastly, we developed **Lunar_Eval**, the first comprehensive evaluation suite for assessing the capabilities of LLMs in lunar exploration tasks. Experimental validation demonstrates that our approach not only enhances domain expertise in lunar exploration but also reveals preliminary indications of embodied intelligence potential.

1 Introduction

The Moon, Earth's closest celestial neighbor, has long been a focal point of human space exploration (Pei et al., 2020). Since 2018, numerous countries and organizations have updated their lunar exploration strategies, establishing long-term objectives such as the construction of large-scale lunar research facilities and the execution of extended lunar missions (Wang et al., 2024b; Li et al., 2019; Gaddis et al., 2023; Chai et al., 2024). The complexity of lunar exploration missions continues to rise, driven by increasing communication delays and limited prior knowledge of mission objectives. Consequently, there is a growing demand for autonomous capabilities in spacecraft operating within the Moon's remote, unknown, and uncertain environment (Zhang et al., 2024; Wang et al., 2024c). With the rapid advancement and practical implementation of artificial intelligence (AI) technologies, integrating AI into lunar exploration to enhance spacecraft autonomy has become both essential and feasible(Li et al., 2023).

The rise of LLMs, such as ChatGPT (OpenAI, 2022) and DeepSeek (Liu et al., 2024), has sparked successive waves of AI innovation. Various specialized domains have utilized knowledge distillation and supervised fine-tuning on open-source foundational models, including Qwen (Bai et al., 2023), LLaMA (Dubey et al., 2024), and ChatGLM (Zeng et al., 2022), with ongoing updates. Despite these advancements, current models remain far from fully meeting the specific needs of lunar exploration.

To address these challenges and support the research and applications of the International Lunar Research Station (ILRS) (Xu and Ou, 2023), we have developed the first Earth-Moon collaborative twin model, **Lunar Twins**. This system comprises two models: the "Chang'e" large model and the "Yutu" small model, named after the Chinese mythological figures of Chang'e, the Moon goddess, and Yutu, the Jade Rabbit. Our key contributions are summarized as follows:

- We propose an efficient human-machine interaction framework that leverages LLMs for resource-constrained lunar environments, as illustrated in Figure 1.

- We design a multi-agent cooperative network for generating lunar instructions and construct a dataset that integrates data from the Chang'e mission.

- We introduce twin models (Chang'e and Yutu-Text) and a multimodal model (Yutu-VL) for lunar exploration, incorporating a retrieval-augmented generation (RAG) mechanism to mitigate hallucinations.

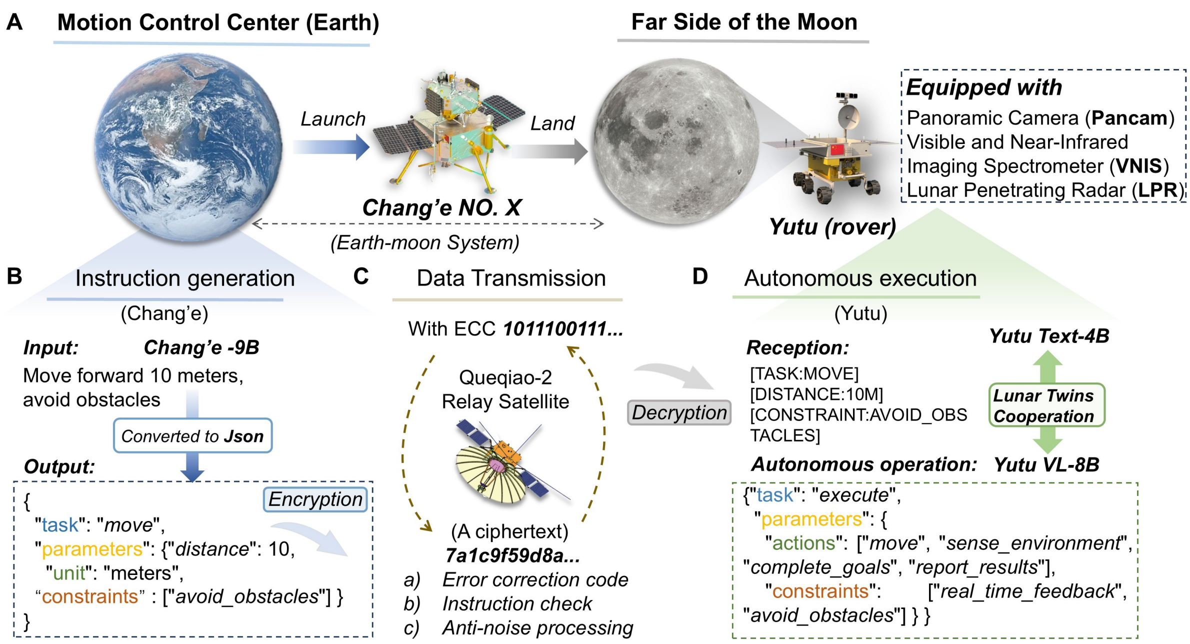


Figure 1: The Earth-moon synergy framework led by **Lunar Twins**. The Earth control center sends simple task instructions (e.g., *"move 10 meters, avoid obstacles"*) to the lunar rover "Yutu" via the Queqiao-2 relay satellite. After encryption and transmission, Yutu-Text optimizes the instructions for autonomous execution, including environmental sensing, goal completion, and real-time feedback. ECC and noise filtering ensure reliable data transmission over long distances.

2 Related Works

2.1 Frontiers of Space Exploration

Early lunar exploration rovers relied on remote control (Kalery et al., 2010), while later systems, such as Kirobo (Samani and Ceccarelli, 2021) and Robonaut2 (Pataranutaporn et al., 2021), integrated speech recognition and natural language processing technologies to enable intelligent interaction. In lunar and Mars exploration missions, including "Yutu-2" (Ding et al., 2022a), "Spirit" (Morris et al., 2010), and "Opportunity" (Squyres et al., 2006), visual cameras for terrain mapping and interactive path planning via ground control centers have played a crucial role. Operators have optimized the robots' paths using visual data. More advanced systems, such as "Curiosity" (Welch et al., 2013) and "Perseverance" (Mangold et al., 2021), have achieved highly autonomous navigation by integrating vision systems with path planning subsystems. "Zhurong" (Ding et al., 2022b) has significantly enhanced positioning accuracy through precise vision-based path planning.

2.2 LLMs and RAG in Science

In recent years, domain-specific LLMs have emerged rapidly, particularly in fields characterized by stable multimodal terms, formats, and data structures, such as healthcare (Wang et al., 2023), law (Nguyen, 2023), education (Liu et al., 2023), finance (Yang et al., 2023), and psychology (Yang et al., 2024). These advancements highlight the potential of LLMs in scientific research. Additionally, technologies such as Graph RAG (Edge et al., 2024) and Agentic RAG (Singh et al., 2025) have accelerated the integration of external modules. Through supervised fine-tuning on domain-specific datasets and enhanced retrieval from local knowledge bases, LLMs have demonstrated remarkable adaptability, solidifying their value in specialized fields and amplifying their scientific and practical impact (Ling et al., 2023).

3 Lunar Dataset

3.1 Raw Data Collection

To ensure the diversity of the lunar corpus, we collected extensive lunar-related text data from a variety of sources (Xu et al., 2020; Yuan et al., 2021; He et al., 2023), including lunar exploration textbooks, literatures and historical mission records. These datasets encompass a broad range of topics relevant to lunar exploration, and the statistics for the pre-training corpus are presented in Table 1.

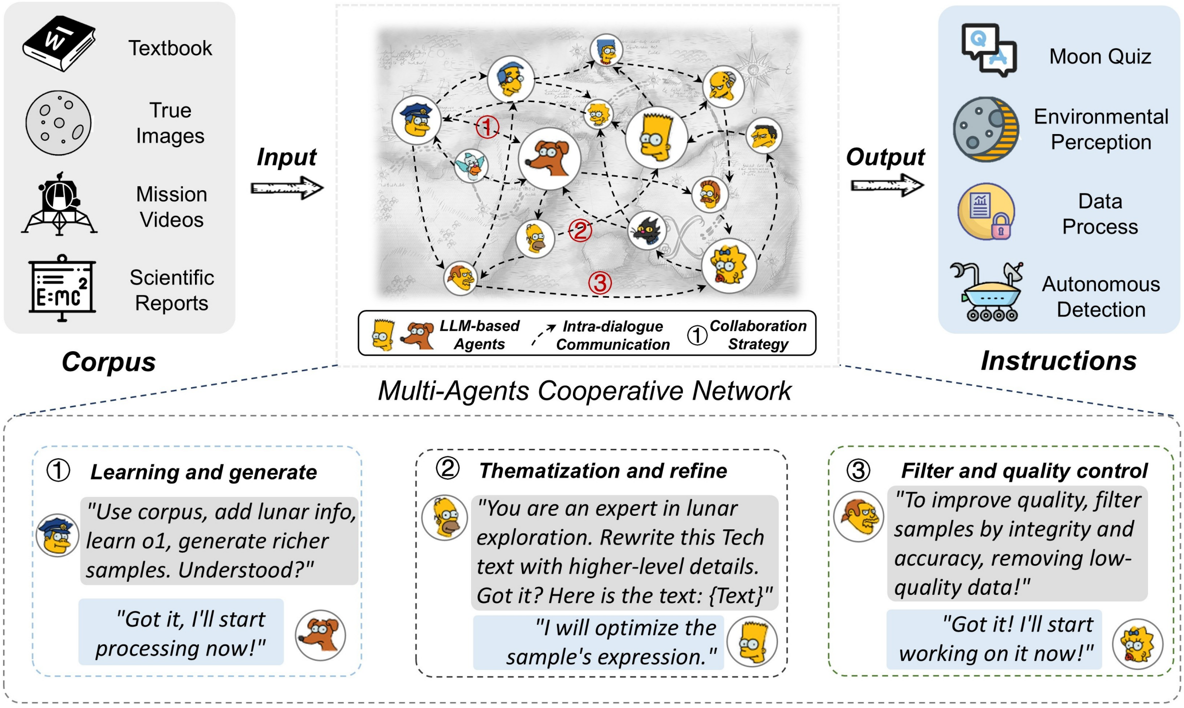


Figure 2: Illustration of the Multi-Agents Cooperative Network for lunar datasets generation and optimization. **Note: **The Simpsons family** images shown here are included solely for illustrative purposes and do not imply any endorsement, affiliation, or official authorization by the rights holders.*

Type	Task	Description	#Samples
Textbook	Lunar Wiki	Lunar-related knowledge points	61973
	Lunar News	A compilation of past lunar-related events	133049
	Lunar Instruct	Content designed to enhance instruction representation capabilities for tasks	71653
	Lunar Exam	Professional questions for evaluation	13625
	Lunar Poetry	Chinese poetry celebrating the Moon	30639
	Lunar Gen	QA dataset distilled from GPT-4	101343
Literatures	Lunar Paper	A collection of academic papers	12280
Multimodal	Lunar Mission	Data released from the Chang'e Mission	37324

Table 1: Pre-training data statistics and sources for Lunar Twins.

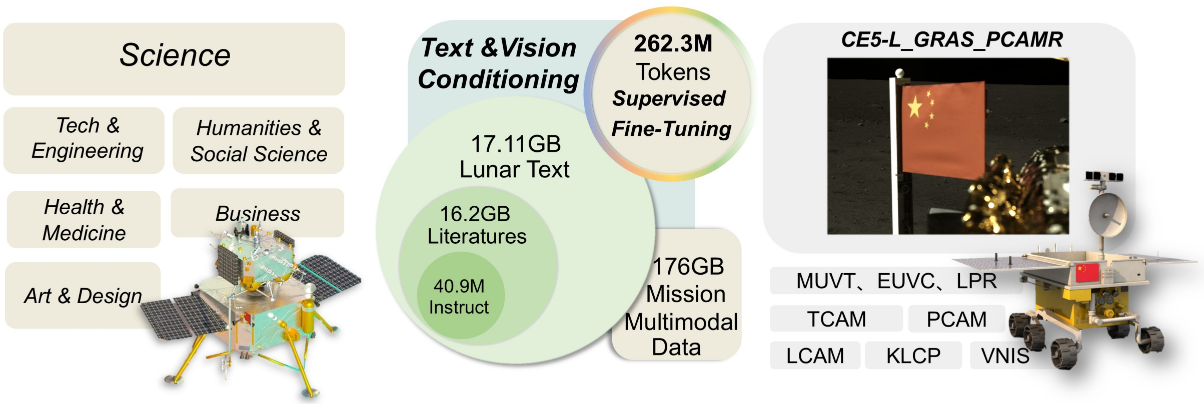


Figure 3: Statistics of final instruction Lunar dataset, showing the data volume (262.3M tokens) used for supervised fine-tuning (SFT), categorized by various themes such as *Science*, *Tech & Engineering*, and *Humanities & Social Science*. The dataset incorporates 17.11GB of Lunar Text, 16.2GB of literature, and 40.9M instruction tokens, with additional multimodal data (176GB) and specific mission payloads, including **PCAMR** (Panoramic Camera).

Relying solely on textbook knowledge is insufficient for developing a lunar large model capable of executing real-world tasks. Successful lunar missions require practical experience, expert insights, and even intuition. To address this, we also collected data from China's Chang'e missions[1]. Detailed mission information can be found in Appendix B, with data types listed in Appendix C.

3.2 Instruction Generation with Multi-Agents

Employing a cooperative multi-agent network architecture (see Figure 2), we introduce **Lunar_GenData**—a domain-specific instruction-generation framework that leverages two complementary fine-tuning strategies, namely task-adaptive parameter optimization and context-driven data augmentation, to produce high-fidelity instruction datasets precisely tailored to the operational profiles of the Chang'e and Yutu models.

Algorithm 1 Lunar Instruction Generation

```
Require:
Dataset T, format as (Inst, Input & Output);
Science Corpus C;
Pre-defined Rule sets R_1 and R_2 for Filtering
Ensure:
Chang'E dataset D_C and Yutu dataset D_Y
▷ Initialization:
Initialize datasets: D_C ← ∅, D_Y ← ∅
▷ Proliferation Generation:
    For each sample in T do:
      Inst, Input & Output ← sample
      enriched_sample ← Enrich(Instruction, C)
      refined_sample ← Refine(enriched_sample)
      D_Y ← D_Y ∪ refined_sample
    End for
▷ Rule-Based Filtering (R_1):
    For each sample in D_C do:
        Apply R_1, discard invalid samples
    End for
▷ Agent-Based Selection (R_2):
    For each sample in D_Y do:
        Apply R_2, retain high-quality samples
        D_C ← D_C ∪ Filter(D_Y, R_2)
    End for
    Data Optimization: Refine D_Y
Return D_C & D_Y
```

[1]https://moon.bao.ac.cn/

3.3 Multi-agents Cooperation Network

Reproductive Generation Strategy. We propose a reproductive generation strategy inspired by biological reproduction. Specifically, initial parent samples generated by ChatGPT (o1) are selected from a parent dataset. These samples undergo a reproductive process facilitated by collaboration among multiple agents. The reproductive process consists of two key stages:

- **Sample Depth Expansion**: Agents enhance the selected samples by incorporating lunar science-related background knowledge from the corpus, resulting in more detailed and informative child samples.
- **Self-Organizing Evolution**: Agents generate samples with higher-level details and logical coherence based on a specific theme or concept, thereby improving the expression and applicability of the child samples.

Dataset Distillation and Optimization. To enhance the quality of the dataset and optimize the training effectiveness of smaller models, we employ a knowledge distillation strategy. Through rigorous filtering by the agents, the reproductive samples are assessed based on criteria such as content completeness, background consistency, and semantic accuracy. Samples that fail to meet these quality standards are discarded, while only high-quality data are retained. The selected samples then undergo further refinement, including noise reduction, logical reinforcement, and linguistic precision enhancement. More details can be found in Appendix E.

The two datasets were used to train the Chang'e large model and the Yutu small model, with the specific process outlined in Algorithm 1. Dataset statistics are presented in Figure 3. Subsequently, 10% of the examples were extracted for testing the dataset quality. Final evaluations on factual accuracy, relevance, and data freshness indicated that 97.3% of the data met the research standards.

4 Training and evaluation

4.1 Training and Experimental Setup

The inaugural Chang'e, **Chang'e-9B**, is derived from ChatGLM-4-9B and trained using a multi-stage approach: first, continual domain-adaptive pre-training on a large-scale, diverse lunar-centric corpus D_C, followed by targeted fine-tuning via

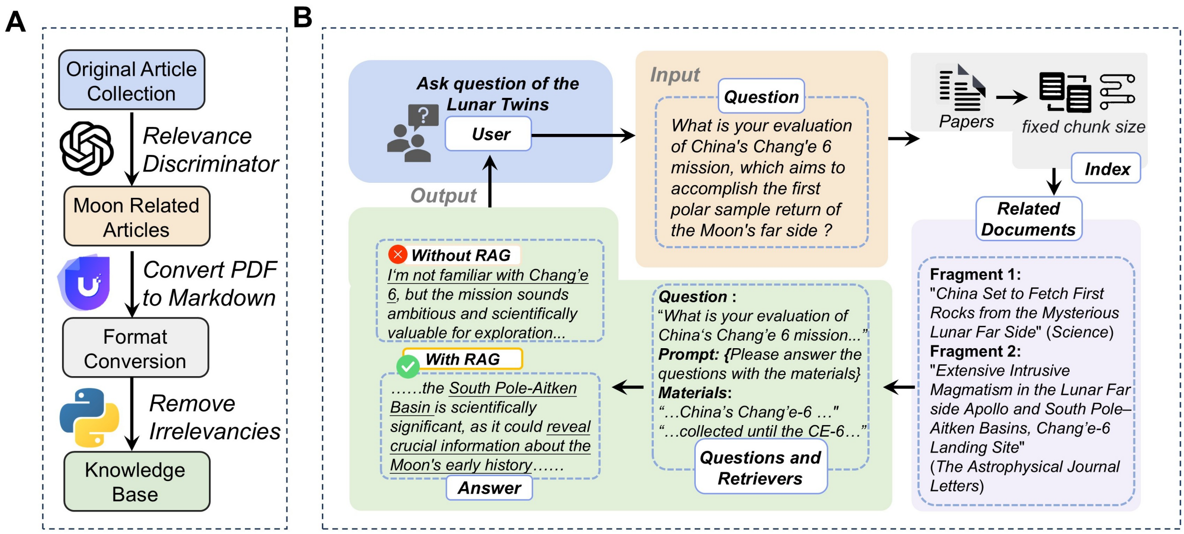


Figure 4: Workflow for processing lunar-related articles and utilizing RAG for enhanced question answering. Panel A illustrates the process of collecting and preparing original articles, including relevance discrimination, conversion to Markdown, and removal of irrelevant information. Panel B demonstrates an example of RAG applied to a question about Chang'e 6 mission, where the documents are split into chunks, relevant sections from scientific articles are retrieved, and the enhanced answer provides more detailed information compared to a response without RAG.

Low-Rank Adaptation (LoRA) (Hu et al., 2021). While primarily designed for lunar exploration, Chang'e-9B also retains broad applicability across general domains, including entertainment, encyclopedic knowledge, and question answering.

In parallel, the Yutu family (**Yutu-Text** and **Yutu-VL**) is built upon MiniCPM3 and MiniCPM-o 2.6 base models, which underwent continual pre-training and full-parameter fine-tuning using a knowledge-distilled, optimized dataset D_Y. This process is specifically tailored to enhance fine-grained reasoning and semantic expression capabilities on high-fidelity lunar tasks. Comprehensive experimental details are provided in Appendix D.

4.2 Retrieval-augmented Generation

As previously mentioned, we crawled a substantial number of academic articles from Web of Science[2] and CNKI[3] on the topic of "lunar exploration." These articles were then analyzed using LLMs for relevance, resulting in the selection of 10,481 complete PDF documents (4,770 in Chinese and 5,711 in English). Subsequently, we employed the open-source tool MinerU (Wang et al., 2024a) to convert these documents into Markdown format, creating a local database.

To ensure data quality and consistency, further processing of the collected dataset was required. Regular expressions were applied to filter out extraneous spaces, line breaks, and other non-text characters. The processed files encompass a range of lunar exploration topics, including astrophysics, geology, and aerospace.

In this work, we leverage RAG to enhance lunar exploration knowledge modeling, benefiting both the Chang'e and Yutu models by improving their precision on specialized queries and reducing hallucinations. For the RAG processing, we selected LightRAG (Guo et al., 2024), one of the most efficient solutions currently available. The approach for RAG is illustrated in Figure 4.

4.3 Benchmark Evaluation

Given the absence of a dedicated benchmark dataset for lunar exploration, we developed a set of multiple-choice and single-choice questions related to lunar exploration, for the text-based evaluation of the Chang'e large model and the Yutu small model. This set is referred to as **Lunar_Eval**, as illustrated in Figure 5. To ensure the fairness and reliability of the evaluation results, we conducted a manual assessment involving five annotators, all PhD candidates in astronomy. The annotators performed pairwise comparisons of each question and its corresponding answer choices. A total of 200 responses, generated by Lunar Twins and their base models, were presented for evaluation.

[2] https://www.webofscience.com/wos/
[3] https://www.cnki.net/

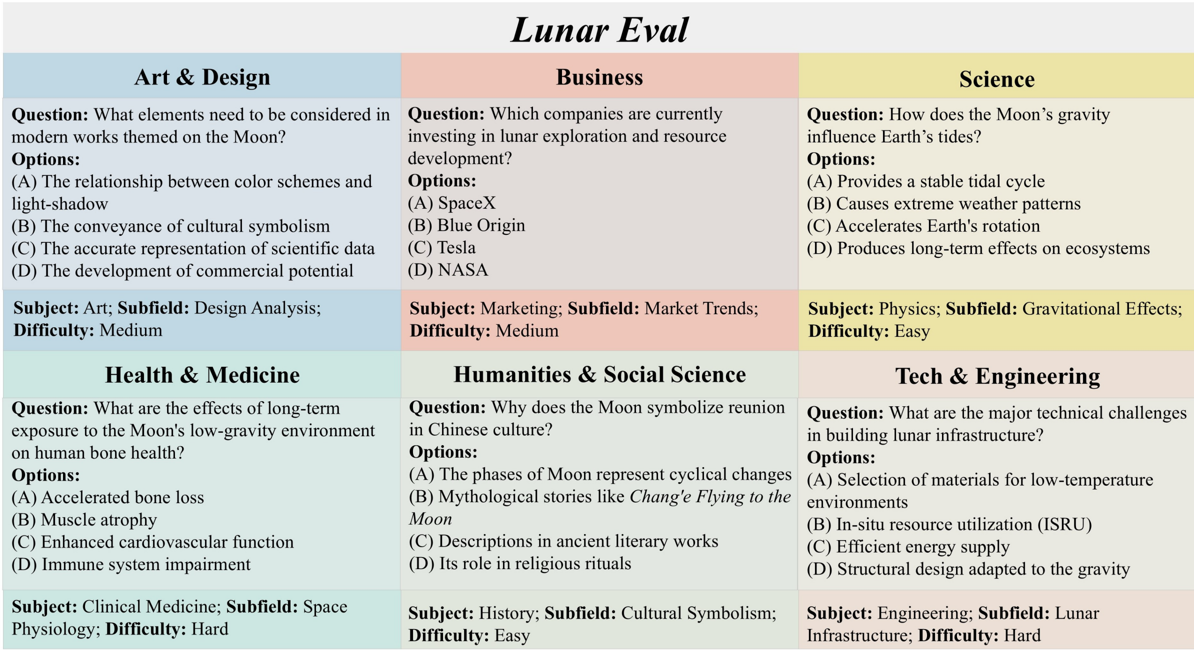


Figure 5: Lunar_Eval for evaluation.The eval includes both single-choice and multiple-choice questions across six major themes, totaling 15,000 questions. The difficulty distribution of the Lunar Eval follows an approximate 3:4:3 ratio. For detailed dataset definitions, refer to Appendix A and Figure 10.

Model	Avg. zh / en	SE zh / en	HS zh / en	TE zh / en	AD zh / en	BS zh / en	HM zh / en
Internlm2.5-7B-Chat	71.5/83.7	71.8/84.0	69.6/81.4	<u>73.1</u>/83.2	70.4/84.2	**74.3**/83.4	69.5/85.8
Yi-1.5-6B-Chat	62.2/62.2	61.0/60.1	60.7/60.3	64.8/60.3	62.2/61.0	62.9/61.0	61.7/70.8
Qwen2.5-7B-Instruct	71.0/84.7	68.5/84.9	68.4/81.8	70.9/82.8	70.0/83.4	72.0/85.3	**76.0**/<u>90.0</u>
LLaMA3.1-8B-Instruct	65.1/78.1	64.0/<u>85.9</u>	62.9/84.3	66.4/85.5	64.7/58.5	65.2/60.0	67.5/**94.2**
ChatGLM3-6B	67.5/82.3	66.5/82.3	68.2/80.6	66.7/80.0	67.3/81.9	67.1/80.6	69.5/88.3
Qwen2-7B-Instruct	65.3/84.3	62.8/83.4	63.3/83.0	65.2/82.1	63.9/84.5	63.2/85.1	70.8/82.5
Gemma-2-9B	70.8/82.5	69.9/81.5	69.5/81.4	70.5/83.1	70.6/82.9	73.2/83.9	70.8/82.5
ChatGLM-4-9B	68.0/81.7	66.7/81.5	67.7/79.8	71.0/81.3	67.5/85.8	67.7/80.2	67.5/88.3
MiniCPM-3-4B	46.8/62.6	46.5/59.9	47.4/59.9	47.9/59.9	46.5/63.5	50.0/62.1	42.2/70.0
Chang'e (Ours)	**73.6/88.0**	**74.4/88.1**	**72.5/87.2**	**73.5/88.2**	**75.3/88.3**	73.3/**88.2**	<u>75.3</u>/<u>90.0</u>
Yutu-Text (Ours)	<u>72.8</u>/<u>86.9</u>	<u>74.2</u>/85.4	<u>71.4</u>/<u>86.0</u>	72.1/<u>87.3</u>	<u>71.8</u>/<u>86.1</u>	<u>73.5</u>/<u>87.6</u>	74.0/89.2

Table 2: Single choice zh/en results of Lunar Eval. 'Avg." measures the micro-average accuracy. 'SE' stands for *Science*. 'HS' stands for *Humanities & Social Science*. 'TE' stands for *Tech & Engineering*. 'AD' stands for *Art & Design*. 'BS' stands for *Business*. 'HM' stands for *Health & Medicine*.

Model	Avg. zh / en	SE zh / en	HS zh / en	TE zh / en	AD zh / en	BS zh / en	HM zh / en
Internlm2.5-7B-Chat	64.1/76.0	70.2/78.0	65.6/79.5	62.4/**90.2**	72.1/69.2	52.8/56.8	61.3/82.0
Yi-1.5-6B-Chat	61.0/62.4	57.6/63.4	61.6/62.6	58.6/51.5	65.9/63.1	70.2/52.5	51.9/61.1
Qwen2.5-7B-Instruct	<u>75.0</u>/**86.7**	59.3/77.4	73.2/69.9	**83.2**/86.6	<u>80.8</u>/<u>81.1</u>	<u>80.1</u>/85.8	73.4/79.5
LLaMA3.1-8B-Instruct	74.6/<u>84.8</u>	67.7/**85.2**	76.2/75.2	71.9/<u>89.8</u>	**81.1**/76.8	76.2/**96.7**	74.2/**84.9**
ChatGLM3-6B	68.7/81.2	71.6/77.2	61.2/<u>84.8</u>	69.6/81.6	61.9/69.6	71.7/84.8	76.5/79.2
Qwen2-7B-Instruct	66.3/80.3	62.0/<u>84.1</u>	73.8/70.3	53.3/76.4	73.9/78.6	50.9/87.4	**83.7**/74.9
Gemma-2-9B	66.1/74.1	58.9/68.1	65.1/72.7	66.0/70.3	63.3/74.0	59.6/89.2	<u>83.4</u>/70.6
ChatGLM-4-9B	69.9/78.3	<u>74.1</u>/66.2	57.3/**89.9**	63.7/80.4	66.5/67.7	79.1/85.8	79.0/80.0
MiniCPM-3-4B	46.7/62.2	52.0/52.0	44.7/52.7	43.8/67.3	49.1/75.7	42.4/63.0	48.2/62.3
Chang'e (Ours)	74.0/72.3	74.0/78.7	**84.7**/81.7	65.7/74.0	76.6/**81.8**	65.5/70.7	77.7/77.1
Yutu-Text (Ours)	**79.3**/82.9	**79.1**/69.5	<u>78.6</u>/80.7	<u>75.4</u>/69.8	79.2/79.1	**80.5**/<u>95.1</u>	83.2/<u>83.2</u>

Table 3: Multiple choice zh/en results of Lunar Eval. The optimal value is **in-bold** and the suboptimal is <u>underlined</u>.

The evaluation was based on two criteria: **correctness** and **helpfulness**. Correctness measures whether the response provides accurate scientific knowledge to address the posed question, while helpfulness assesses whether the model can assist the user in a concise and effective manner, taking user intent into account. In practice, an answer may be correct but still not helpful if it is excessively verbose or lacks clarity. To determine the final result for each evaluation instance, we employed a majority voting approach. If at least two annotators agreed, their preference was considered the final answer; otherwise, the outputs from both models were regarded as a tie.

For the multi-modal evaluation of the Yutu model, we conducted a detailed comparison with ChatGPT-4 in Section 5.2.

5 Result

5.1 Performance Analysis Results

Lunar Twins showed superior performance compared to LLMs of similar scale. In Table 2 and Table 3, we compare the performance of the Chang'e large model and the Yutu small model against their respective base models (ChatGLM-4-9B and MiniCPM3) as well as against other open-source models with similar parameter scales. Using both automated and human evaluators, we calculated the win rates relative to these baselines. Across most tasks, Lunar Twins consistently outperforms models of comparable size, validating the effectiveness of the proposed approach. The distribution of difficulty-level evaluation results is presented in Table 4 and Table 5.

Lunar Twins Receives Greater Preference. Consistent with previous results, Lunar Twins outperforms its base models in manual evaluations. Figure 6 show that the Chang'e model generates more accurate answers compared to ChatGLM-4, with 48% of its responses rated as the top choice by domain experts. In addition to correctness, the Yutu model particularly excels in helpfulness. Compared to correctness, Lunar Twins shows more significant improvement in helpfulness. This aligns with expectations, as the core goal of supervised fine-tuning (SFT) is to enhance the instruction-following ability of LLMs (Zhang et al., 2023), enabling them to better meet the diverse needs of users, rather than merely expanding their knowledge domain. More experimental results are presented in Appendix F.

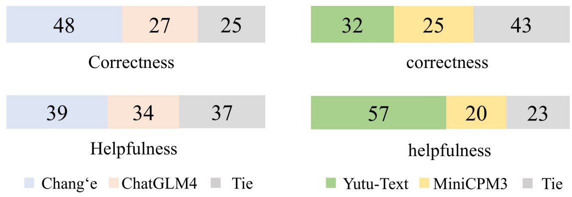


Figure 6: Results of manual evaluation of Lunar Twins and base models,with assessment based on correctness and helpfulness.

Model	Easy zh / en	Medium zh / en	Hard zh / en
Internlm2.5-7B-Chat	61.1/60.1	63.8/60.4	62.7/61.8
Yi-1.5-6B-Chat	47.0/59.6	48.5/62.2	46.4/61.2
Qwen2.5-7B-Instruct	71.6/85.8	73.7/87.5	71.2/86.2
LLaMA3.1-8B-Instruct	66.8/80.8	67.4/81.0	67.5/81.5
ChatGLM3-6B	63.3/84.1	64.6/83.5	64.8/82.6
Qwen2-7B-Instruct	71.0/83.1	72.3/83.7	71.5/82.6
Gemma-2-9B	69.9/81.5	69.1/80.9	70.3/83.1
ChatGLM-4-9B	69.4/82.1	70.7/84.1	70.3/83.6
MiniCPM-3-4B	63.4/84.8	65.7/86.5	65.5/86.0
Chang'e (Ours)	**72.1**/**87.6**	74.3/**88.5**	**72.8**/**87.2**
Yutu-Text (Ours)	67.5/81.6	**79.9**/80.6	67.4/81.3

Table 4: Single choice zh/en results of Lunar Eval. 'Easy' represents the easy questions, 'Medium' represents medium difficulty questions, and 'Hard' represents the hard questions.

Model	Easy zh / en	Medium zh / en	Hard zh / en
Internlm2.5-7B-Chat	66.0/60.6	68.8/52.7	66.3/60.3
Yi-1.5-6B-Chat	52.6/60.3	51.1/69.8	41.8/57.5
Qwen2.5-7B-Instruct	69.4/82.6	72.1/75.9	71.7/**89.2**
LLaMA3.1-8B-Instruct	73.6/72.5	63.0/88.0	80.9/88.0
ChatGLM3-6B	64.1/73.3	65.6/76.9	79.0/65.5
Qwen2-7B-Instruct	66.7/**90.8**	80.6/75.6	71.8/70.7
Gemma-2-9B	68.0/72.8	60.7/77.3	**83.5**/77.6
ChatGLM-4-9B	57.7/74.4	53.1/76.6	70.6/67.7
MiniCPM-3-4B	54.0/88.6	**91.7**/91.8	60.7/84.2
Chang'e (Ours)	74.5/77.8	69.8/76.3	79.3/67.6
Yutu-Text (Ours)	**77.2**/79.7	81.3/**92.1**	70.8/81.0

Table 5: Multiple choice zh/en results of Lunar Eval.

5.2 Case Study

Furthermore, we present a multi-modal case study in Figure 7, which demonstrates the advantages of the **Yutu** model over **GPT-4**[4]. The results indicate that Lunar Twins exhibits a superior level of expertise in describing lunar exploration research. Its outputs provide detailed, context-specific information closely aligned with the mission background—whereas GPT-4's outputs remain at a more general level, focusing primarily on surface morphology and physical characteristics.

[4] https://openai.com/index/gpt-4/

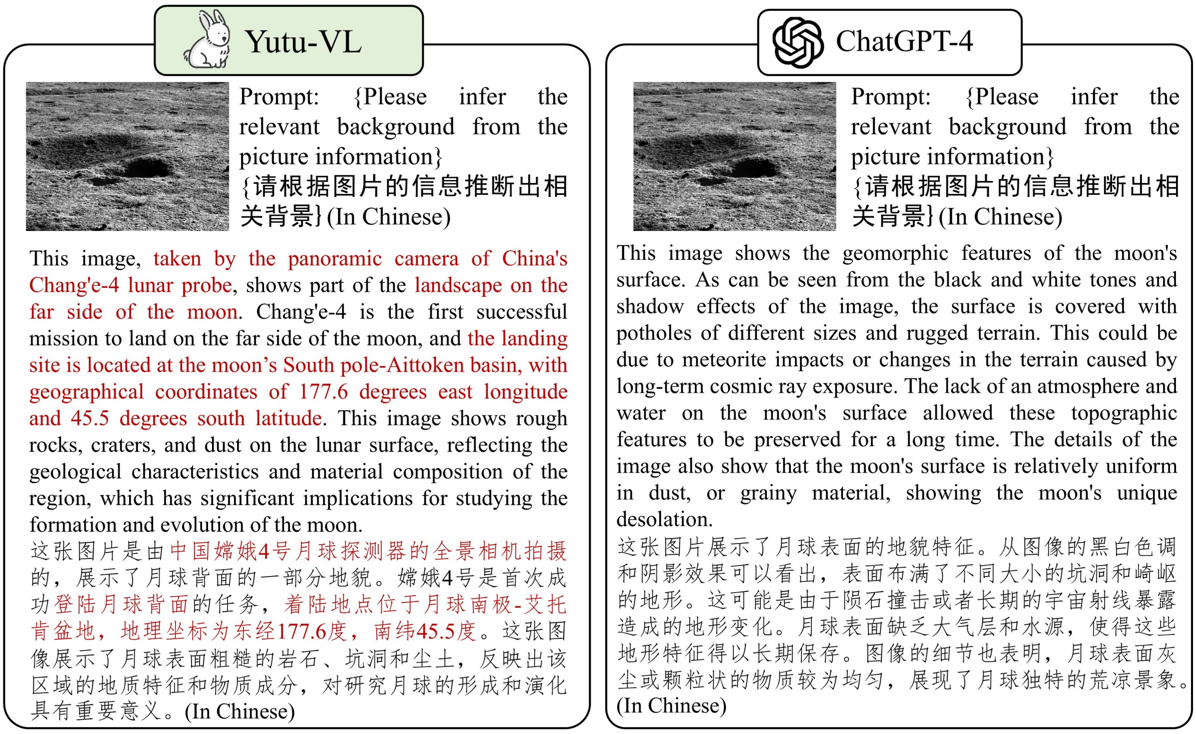


Figure 7: Comparison between **Yutu-VL** and **ChatGPT-4** outputs. The left panel shows the result of Yutu-VL, which provides more detailed background information in red text about the landscape on the moon's far side as captured by the Chang'e-4 probe. The right panel shows ChatGPT-4's output, which includes a more general description of the moon's surface features. Both outputs are presented in Chinese and English, highlighting the difference in detail and relevance of the background information.

5.3 On-Device Language Models

Due to the high adaptability of the MiniCPM-V base model on edge devices, we conducted performance tests on MiniCPM-V2.0 (2.8B) and MiniCPM-V2.5 (8B) on mobile devices, following the official documentation[5]. The testing environment consisted of a Snapdragon 8 Gen 3 platform with 16 GB of RAM. The results demonstrated that these two models achieved smooth inference speeds of 16.17 tokens/s and 8.62 tokens/s, respectively, highlighting their strong real-time inference capabilities on mobile devices.

Building on this, we successfully deployed the fine-tuned Yutu model on the RK3588 processor platform and achieved preliminary human–machine dialogue interaction. This marks an important step toward enabling the Yutu model to support complex environmental perception, robotic arm control, and planning tasks in offline lunar rover applications under constrained hardware conditions. We consider this deployment a solid technical foundation for advancing future on-device lunar mission capabilities.

[5] https://github.com/OpenBMB/MiniCPM-o

6 Conclusion

This study addresses the limitations of current LLMs in the domain of lunar exploration by proposing and developing the first twin system specifically tailored for this field, named Lunar Twins. The system consists of the "Chang'e" large model and the "Yutu" smaller model. We introduce a collaborative framework that harnesses the synergy between large and small language models, facilitating efficient human-machine interaction through natural language instructions, particularly in the resource-constrained environment of lunar exploration.

Additionally, we have constructed a specialized dataset that integrates real lunar exploration mission data and designed a multi-agent collaborative workflow to generate a lunar surface-specific corpus. This approach helps address the knowledge gaps in mainstream pre-trained models, which often overlook the unique aspects of lunar exploration. Experimental evaluations demonstrate that our models outperform similar-scale baseline models in a variety of lunar exploration tasks, underscoring their potential value in specialized domains.

Acknowledgement

This work was supported by the Beijing Natural Science Foundation of China (Grant No.3254040), the National Natural Science Foundation of China (Grant No. 62441219), and National Key Laboratory of Human Factors Engineering.

Limitations

Hallucination There is existing evidence indicating that the distribution of pre-training and instruction data may suffer from significant bias, which can negatively impact the quality of model outputs (Ji et al., 2023). A substantial portion of the data used in Lunar Twins is synthesized by LLMs. Although manual review was conducted, the overall quality still cannot match that of human-annotated data. While the introduction of RAG techniques has alleviated this issue to some extent, from a scientific perspective, this is not the optimal solution.

Lack of Evidence for Embodied Intelligence There is a lack of sufficient evidence regarding the embodied intelligence capabilities of the Yutu model. Although, theoretically, Yutu provides a potential solution for autonomous lunar exploration based on vision-language models (VLMs), we currently lack a quantitative understanding of its intelligence level, interaction logic, and stability.

Ethics Statement

In conducting our research, we prioritize the highest ethical standards to ensure integrity and make a positive contribution to the scientific community. We exclusively use open-source datasets, ensuring that our work is built upon accessible and transparent resources. Our methods employ models that are either open-source or widely recognized for their reliability and ethical use within the academic community. Additionally, we have carefully designed our methodology to prevent the generation of harmful or misleading information, thus safeguarding the integrity of our findings.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, and Fei Huang. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2023. Oceangpt: A large language model for ocean science tasks. *arXiv preprint arXiv:2310.02031*.

Jingxuan Chai, Youmin Gong, Jie Mei, Pengyu Wang, and Guangfu Ma. 2024. Modeling and trajectory tracking control for a novel multinode flexible small-body lander based on a port-hamilton framework. *Space: Science & Technology*, 4:0113.

L. Ding, R. Zhou, T. Yu, H. Gao, H. Yang, J. Li, Y. Yuan, C. Liu, J. Wang, and Y.-YS Zhao. 2022a. Surface characteristics of the zhurong mars rover traverse at utopia planitia. *Nature Geoscience*, 15(3):171–176. ISBN: 1752-0894 Publisher: Nature Publishing Group UK London.

Liang Ding, Ruyi Zhou, Ye Yuan, Huaiguang Yang, Jian Li, Tianyi Yu, C. Liu, Jian Wang, Shu Li, and Haibo Gao. 2022b. A 2-year locomotive exploration and scientific investigation of the lunar farside by the yutu-2 rover. *Science Robotics*, 7(62):eabj6660. ISBN: 2470-9476 Publisher: American Association for the Advancement of Science.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Lisa R. Gaddis, Katherine H. Joy, Ben J. Bussey, James D. Carpenter, Ian A. Crawford, Richard C. Elphic, Jasper S. Halekas, Samuel J. Lawrence, and Long Xiao. 2023. Recent exploration of the moon: Science from lunar missions since 2006. *Reviews in Mineralogy and Geochemistry*, 89(1):1–51. ISBN: 1529-6466 Publisher: Mineralogical Society of America.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. 2023. WanJuan: A comprehensive multimodal dataset for advancing English and Chinese large models. ArXiv:2308.10755 [cs].

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. ISBN: 0360-0300 Publisher: ACM New York, NY.

Alexander Yu Kalery, Igor V. Sorokin, and Mikhail V. Tyurin. 2010. Human space exploration beyond the international space station: Role of relations of human, machine and the “earth”. *Acta Astronautica*, 67(7-8):925–933. ISBN: 0094-5765 Publisher: Elsevier.

Chunlai Li, Chi Wang, Yong Wei, and Yangting Lin. 2019. China’s present and future lunar exploration program. *Science*, 365(6450):238–239. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

Danyi Li, Yinkang Li, Xu Liu, Bin Yang, Xuxing Huang, Yong Yang, Bingheng Wang, and Shuang Li. 2023. Pseudospectral convex programming for free-floating space manipulator path planning. *Space: Science & Technology*, 3:0030.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, and Chong Ruan. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Wentao Liu, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, and Aimin Zhou. 2023. Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*.

Nicolas Mangold, S. Gupta, O. Gasnault, G. Dromart, J. D. Tarnas, S. F. Sholes, B. Horgan, C. Quantin-Nataf, A. J. Brown, and S. Le Mouélic. 2021. Perseverance rover reveals an ancient delta-lake system and flood deposits at jezero crater, mars. *Science*, 374(6568):711–717. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

Richard V. Morris, Steven W. Ruff, Ralf Gellert, Douglas W. Ming, Raymond E. Arvidson, Benton C. Clark, D. C. Golden, Kirsten Siebach, Göstar Klingelhöfer, and Christian Schröder. 2010. Identification of carbonate-rich outcrops on mars by the spirit rover. *Science*, 329(5990):421–424. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*.

OpenAI. 2022. Introducing chatgpt.

Pat Pataranutaporn, Valentina Sumini, Ariel Ekblaw, Melodie Yashar, Sandra Häuplik-Meusburger, Susanna Testa, Marianna Obrist, Dorit Donoviel, Joseph Paradiso, and Pattie Maes. 2021. SpaceCHI: Designing human-computer interaction systems for space exploration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Zhaoyu Pei, Jizhong Liu, Qian Wang, Yan Kang, Yongliao Zou, He Zhang, Yuhua Zhang, Huaiyu He, Qiong Wang, Ruihong Yang, Wei Wang, and Jinan Ma. 2020. Overview of lunar exploration and international lunar research station. *Chinese Science Bulletin*, 65(24):2577–2586.

F. Samani and M. Ceccarelli. 2021. *An experimental characterization of TORVEastro, cable-driven astronaut robot. Robotics 2021, 10, 21.* s Note: MDPI stays neutral with regard to jurisdictional claims in published ....

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*.

S. W. Squyres, A. H. Knoll, R. E. Arvidson, B. C. Clark, J. P. Grotzinger, B. L. Jolliff, S. M. McLennan, N. Tosca, J. F. Bell Iii, and W. M. Calvin. 2006. Two years at meridiani planum: Results from the opportunity rover. *Science*, 313(5792):1403–1407. ISBN: 0036-8075 Publisher: American Association for the Advancement of Science.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, and Fukai Shang. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Chi Wang, Yingzhuo Jia, Changbin Xue, Yangting Lin, Jianzhong Liu, Xiaohui Fu, Lin Xu, Yun Huang, Yufen Zhao, Yigang Xu, Rui Gao, Yong Wei, Yuhua Tang, Dengyun Yu, and Yongliao Zou. 2024b. Scientific objectives and payload configuration of the chang’E-7 mission. *National Science Review*, 11(2):nwad329.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with Chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Jingqian Wang, Yong Zhao, Xiang Zhang, Zhenyu Li, and Wen Yao. 2024c. Model-free intelligent control for space soft robotic manipulators. *Space: Science & Technology*, 4:0120.

Richard Welch, Daniel Limonadi, and Robert Manning. 2013. Systems engineering the curiosity rover: A retrospective. In *2013 8th international conference on system of systems engineering*, pages 70–75. IEEE.

Fengna Xu and Jun Ou. 2023. Promoting international cooperation on the international lunar research station: Inspiration from the ITER. *Acta Astronautica*, 203:341–350. ISBN: 0094-5765 Publisher: Elsevier.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. CLUECorpus2020: A large-scale Chinese corpus for pre-training language model. ArXiv:2003.01355 [cs].

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. *AI Open*, 2:65–68.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Jiawei Zhang, Chengchao Bai, C Patrick Yue, and Jifeng Guo. 2024. Deep marl-based resilient motion planning for decentralized space manipulator. *Space: Science & Technology*, 4:0145.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, and Fei Wu. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

A Explanations of Lunar Issues

To provide researchers with a clear and organized resource, lunar science data has been categorized into six major topics following discussions with multiple experts in lunar exploration and planetary remote sensing. Below are the detailed explanations of each topic:

1. **Art and Design**.This theme explores the application of lunar-related concepts within the fields of art and design. It includes the relationship between color and light, the expression of cultural symbols, and how modern design can present scientific research findings.

2. **Business**. Focused on the commercialization potential of lunar exploration and development, this topic covers market trend analysis, corporate investment activities, and economic value assessments.

3. **Science**. Dedicated to foundational scientific questions related to the Moon, this theme includes research on gravity, astrophysics, astrometry, and lunar geological characteristics.

4. **Health and Medicine**.This topic investigates the short-term and long-term effects of the Moon's low gravity and radiation environment on human health, providing support and data for space medicine.

5. **Humanities and Social Science**. This theme explores the symbolic significance of the Moon in human culture, along with its interpretation and legacy across different cultural contexts.

6. **Tech and Engineering**. Focused on the technological and engineering aspects of lunar exploration, this topic includes infrastructure development, energy utilization, and equipment research and development.

B Introduction to the Chang 'e Missions

Timeline Figure 9 highlights the evolution of LLMs from 2013 to 2024, marking key advancements alongside lunar exploration milestones. Key lunar events, such as Chang'e-3 (2014) and Chang'e-6 (2024), are shown in black boxes for reference. [6]

C Example for the Yutu-VL dataset

It is important to note that during the fine-tuning process of Yutu-VL, we enhanced domain adaptation by incorporating real data from the Chang'e series of lunar exploration missions. For example, high-resolution remote sensing images (Figure 8) and geological spectral data serve as the foundation for studying lunar material composition, topographical features, and environmental changes.

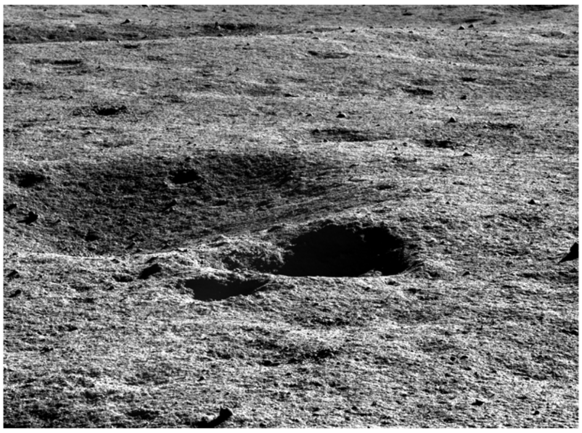
D The experimental setup

The detailed information of LLMs experimental settings are shown in Table 6

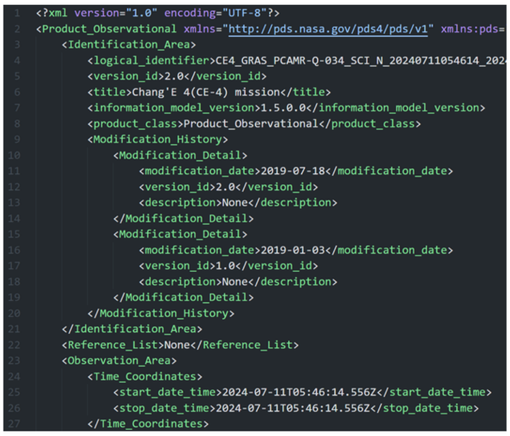
Hyperparameter	Setting
Fine-tuning method	LoRA
Batch Size	512
Device	Nvidia A100
GPU number	*8
Learning Rate (LR)	0.001
LoRA r	8
LoRA α	16
LoRA Dropout	0.05
Epoch	3

Table 6: Detailed experimental settings.

[6] *Data from Chang'e-6 is expected to be released in June 2025 and is therefore not included*



CE4_GRAS_PCAMR-Q-034_SCI_N_
20240711054614_20240711054614_0327_B.2B



```xml
<?xml version="1.0" encoding="UTF-8"?>
<Product_Observational xmlns="http://pds.nasa.gov/pds4/pds/v1" xmlns:pds=
    <Identification_Area>
        <logical_identifier>CE4_GRAS_PCAMR-Q-034_SCI_N_20240711054614_202
        <version_id>2.0</version_id>
        <title>Chang'E 4(CE-4) mission</title>
        <information_model_version>1.5.0.0</information_model_version>
        <product_class>Product_Observational</product_class>
        <Modification_History>
            <Modification_Detail>
                <modification_date>2019-07-18</modification_date>
                <version_id>2.0</version_id>
                <description>None</description>
            </Modification_Detail>
            <Modification_Detail>
                <modification_date>2019-01-03</modification_date>
                <version_id>1.0</version_id>
                <description>None</description>
            </Modification_Detail>
        </Modification_History>
    </Identification_Area>
    <Reference_List>None</Reference_List>
    <Observation_Area>
        <Time_Coordinates>
            <start_date_time>2024-07-11T05:46:14.556Z</start_date_time>
            <stop_date_time>2024-07-11T05:46:14.556Z</stop_date_time>
        </Time_Coordinates>
```


CE4_GRAS_PCAMR-Q-034_SCI_N_
20240711054614_20240711054614_0327_B.2BL

Figure 8: Sample data display from the Chang'e-series missions. On the left, an image of the lunar surface captured by the Chang'e-4 panoramic camera, processed for display and analysis. On the right, the corresponding scientific data and metadata in XML format, including details such as capture time, location, instrument settings, and other relevant information essential for interpreting the image.

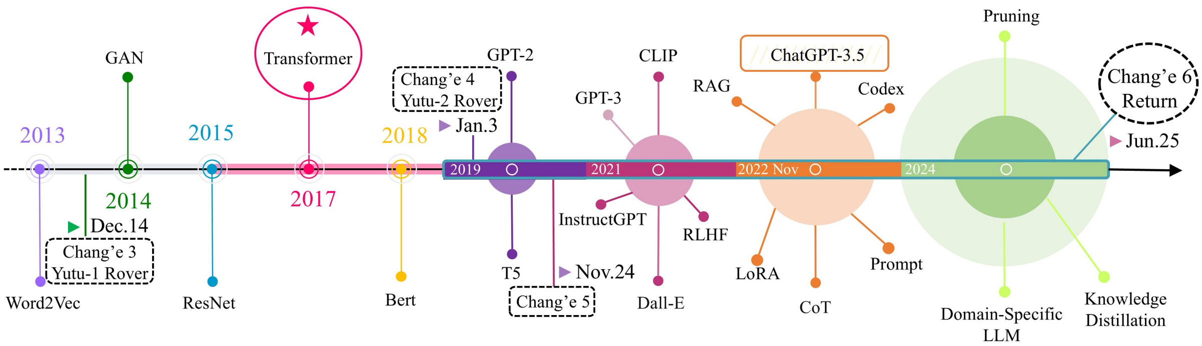


Figure 9: Timeline of LLM Evolution. This timeline illustrates the key milestones in the evolution of LLMs, highlighting landmark advances such as Word2Vec, GANs, Transformers, BERT, GPT series, and specialized techniques like RAG, LoRA, CoT, and domain-specific LLMs. Notably, it also aligns these developments with major milestones in China's lunar exploration program, including the Chang'e 3, 4, 5, and 6 missions, emphasizing the parallel progress of AI and space exploration.

Figure 10: Lunar_Eval's Word Cloud Map of Subject Topic Distribution

E Details of the Multi-agent Network

To systematically present the application of the multi-agent cooperation network in data construction, we provide a formal description covering the network structure, generation pipeline, and diversity control, ensuring clarity and reproducibility.

We formalize the cooperation process as a directed acyclic graph (*DAG*):

$$G = (\mathcal{V}, \mathcal{E})$$

where:

- $\mathcal{V} = \{v_i\}_{i=1}^n$ is the set of nodes, each representing an agent responsible for generation, transformation, or evaluation tasks;
- $\mathcal{E} = \{\langle v_i, v_j \rangle \mid v_i, v_j \in \mathcal{V}, i \neq j\}$ is the set of directed edges indicating information flow.

Each node v_i is associated with an agent $a_i = \rho(v_i)$, where ρ maps the node to a foundation model, domain-specific tools, and role-specific memory. For each edge $\langle v_i, v_j \rangle$, interaction as:

$$\tau(a_i, a_j) = \left(a_i \xrightarrow{\textit{request}} a_j,\ a_j \xrightarrow{\textit{response}} a_i \right)^m$$

where m is the number of reflection-refinement cycles, and only the final artifact $A_j^{(m)}$ is propagated.

The topological ordering satisfies:

$$\forall \langle v_i, v_j \rangle \in \mathcal{E}, \quad I(a_i) < I(a_j)$$

ensuring an ordered, non-circular workflow. Given the input task dataset:

$$\mathcal{T} = \{(I_i, X_i, Y_i)\}_{i=1}^N$$

where I_i is the instruction, X_i the input, and Y_i the expected output, we first sample:

$$A^{(0)} = \text{Sample}(I, X)$$

Agents then apply transformation functions $\phi_{v_j}(\cdot)$:

$$A^{(t+1)} = \phi_{v_j}(A^{(t)}), \quad \text{for } \langle v_i, v_j \rangle \in \mathcal{E}$$

with each transformation comprising:

- *Knowledge Enrichment*:

$$A^{(t)\prime} = \phi_{\text{enrich}}(A^{(t)} \mid \mathcal{C})$$

where $\mathcal{C}$ is external knowledge;

- *Refinement*:

$$A^{(t+1)} = \phi_{\text{refine}}(A^{(t)\prime})$$

We define an evaluation function ψ : Artifact $\rightarrow [0, 1]$ to score artifacts:

$$\sigma_k = \psi(A^{(k)})$$

Artifacts are retained if:

$$\mathcal{D}_{\text{final}} = \left\{ A^{(k)} \mid \sigma_k \geq \theta \right\}, \quad \theta \in (0, 1)$$

To encourage diversity, we apply:

- Sampling with a high temperature ($T > 0.8$);
- Introducing heterogeneous roles (“optimistic” and “critical”) to ensure varied perspectives.

We remove redundant samples by semantic similarity $\text{sim}(A_i, A_j)$, defining:

$$\mathcal{D}_{\text{pruned}} = \{A_i \in \mathcal{D}_{\text{final}} \mid \forall j < i,\ \text{sim}(A_i, A_j) < \delta\}$$

where δ is a predefined similarity threshold.

F Additional experimental results

Model Selection Rationale

In the early-stage experimental phase, we constructed a test set of 1,000 examples for each of three key evaluation dimensions:

1. Domain knowledge, assessed via accuracy;
2. Alignment with user preferences via ROUGE;
3. Collaboration efficacy, measured through Parse Rate and semantic similarity.

Table 7: Early-Stage Model Selection Results

Model	Acc	ROUGE	Parse Rate	Similarity
Qwen2.5-7B	71.3%	62.4	85.6%	78.9%
LLaMA3.1-8B	73.1%	64.0	87.2%	80.4%
ChatGLM-4-9B	**74.8%**	**68.3**	**91.5%**	**85.7%**

We benchmarked several leading open-source models under 14B parameters, including Qwen2.5-7B-Instruct, LLaMA3.1-8B, and ChatGLM-4-9B. As summarized in Table 7, ChatGLM-4-9B consistently outperformed its counterparts, achieving the highest scores across all evaluation metrics, namely 74.8% accuracy in domain knowledge tasks, a ROUGE score of 68.3 for preference alignment, a Parse Rate of 91.5%, and a similarity of 85.7%.

Comparison with State-of-the-Art LLMs

To further validate the performance of our proposed models, we conducted additional evaluations comparing them against several state-of-the-art (SOTA) LLMs, including DeepSeek-R1, DeepSeek-V3-0324, as well as ChatGPT-o1 and o4 mini high.

These SOTA models achieved superior scores, largely attributable to their larger parameter scales and advanced reasoning capabilities, such as Chain-of-Thought (CoT)-based inference mechanisms (see Table 8). For instance, the Chang'e model lags behind DeepSeek-V3 by approximately 4.3% on specific evaluation metrics, underscoring the inherent limitations of smaller models: even with fine-tuning, they cannot surpass cutting-edge systems when evaluated under identical benchmark.

Table 8: Comparison with State-of-the-Art LLMs

Model	Single ZH/EN	Multi ZH/EN
DeepSeek-R1	*80.2* / 94.3	**84.5** (↑5.2) / 92.5
DeepSeek-V3	78.9 / 91.6	81.3 / 90.7
ChatGPT-o1	79.6 / *95.0*	82.0 / *93.1*
o4 mini high	**82.4** (↑9.8) / **96.7** (↑8.7)	*83.1* / **94.2** (↑7.5)

Ablation Study on Generation Methods

To further highlight the contributions of our workflow design, we conducted an ablation study comparing our proposed data generation method, *Lunar_GenData*, against existing approaches, including DOINSTRUCT (Bi et al., 2023) and direct distillation from GPT-4. We reconstructed 5,000 data samples using each method and evaluated the outputs. The results, summarized in Table 9, clearly demonstrate that our *Lunar_GenData* method significantly outperforms both OceanGPT's DOINSTRUCT and direct GPT-4 distillation across all evaluation dimensions.

Table 9: Ablation Study on Generation Methods

Method	ROUGE	Mean Opinion Score	GPT-4-Score
Lunar_GenData	**89.5**	**4.4**	**8.1**
DOINSTRUCT	64.4	3.1	5.7
Distillation	38.0	2.7	3.5

Effect of Number of Agents in Network

We further investigated how the number of nodes in the collaborative agent network affects data quality. Specifically, we conducted experiments using 2, 3, 4, and 5-agent configurations, all based on the Qwen2.5-7B backbone, to evaluate the impact of increasing expert participation.

The results, summarized in Table 10, show that as the number of expert agents increases, the overall generation quality improves consistently across all evaluation metrics. This non-linear improvement highlights the emerging collaborative effect of the multi-agent mechanism in enhancing semantic understanding and multi-task integration.

Table 10: Impact of Number of Agents on Data Quality

Number of Agents	ROUGE ↑	MOS ↑	GPT-4-Score ↑
2 Experts	61.7	3.1	5.3
3 Experts	69.2	3.5	6.7
4 Experts	75.8	4.0	7.5
5 Experts	**83.9**	**4.4**	**8.9**

Cross-LLM Comparisons on the Same Dataset

We further evaluated the performance of fine-tuned models across different base LLMs using our lunar dataset. Specifically, we compared models fine-tuned on Qwen2.5-7B, LLaMA3.1-8B, and ChatGLM4-9B.

Table 11: Performance Comparison of Different LLMs Fine-Tuned with Lunar Data

SFT with Lunar Data	Single ZH/EN	Multi ZH/EN
Qwen2.5-7B	73.1 / 86.3	**77.4** / 88.3
LLaMA3.1-8B	69.9 / 81.6	74.6 / **86.5**
ChatGLM4-9B	**73.6** / **88.0**	74.0 / 72.3

As summarized in Table 11, while Qwen and LLaMA marginally outperform ChatGLM on certain tasks, ChatGLM remains the most suitable foundation for our system, primarily due to its superior compatibility and collaborative performance with small models.

G Detailed Analysis of RAG

Why RAG Remains Necessary After SFT

While supervised fine-tuning (SFT) can inject substantial domain knowledge into the base model, it cannot fully resolve issues such as hallucination or capture rapidly evolving knowledge beyond the training corpus.

As shown in Figure 4 of the main paper, when LunarTwins was tasked with evaluating the Chang'e-6 mission (the latest mission, highlighted in Figure 9), it failed to provide accurate answers without RAG, despite the fine-tuning dataset including 1,542 expert-curated samples related to

Chang'e-6's lunar far side sample return. This failure occurred because the training corpus did not explicitly cover this information. However, when equipped with RAG, Lunar Twins successfully retrieved correct details, including the precise landing site in the South Pole–Aitken Basin. These tests were controlled such that the only variable was the presence or absence of RAG.

Moreover, lunar exploration instructions are highly specialized and detailed, making it impractical for model weights alone to memorize all knowledge. RAG offers a practical "temporary memory" mechanism, essential in offline lunar mission environments, by enabling the model to look up and use relevant external knowledge as needed.

Contributions of SFT vs. RAG

Finally, we highlight two key findings:

***Q1:* Is fine-tuning sufficient, with RAG offering only marginal benefits?**

A: RAG plays a critical role in low-tolerance, high-stakes lunar scenarios where untrained or missing knowledge can lead to factual errors. It serves as an essential complement by dynamically retrieving unseen or evolving information.

***Q2:* Which contributes more to overall performance — SFT or RAG?**

A: Supervised fine-tuning (SFT) is the primary driver, significantly enhancing lunar domain reasoning and task alignment. For example, our Yutu model improves over its base MiniCPM by +26% on Chinese single-choice tasks, while RAG contributes an additional +9.5%.

In summary, SFT equips the model with foundational expertise, and RAG supplies dynamic, precision-driven knowledge retrieval, mitigating hallucinations and closing knowledge gaps.

H Future Work

Looking ahead, we identify several key directions to extend and deepen this research.

First, we will publicly release the developed dataset on Hugging face and GitHub following the review process, leveraging the visibility and influence of the NLP research community to promote its use and encourage further innovation in lunar-focused applications.

Second, while the current system is primarily designed for question-answering (QA) tasks, future work will advance toward reasoning-capable models that can perform multi-step inference, causal analysis, and complex decision-making aligned with the challenges of lunar exploration scenarios.

Third, we plan to integrate these models into onboard lunar rover systems, moving beyond software-only evaluations to embodied intelligence. This effort will require adapting models for resource-constrained environments and enabling real-time perception, robotic arm control, and autonomous planning in offline lunar conditions. We have already completed supplementary tests deploying the fine-tuned Yutu-VL and Yutu-Text models on the RK3588 development board, providing a promising technical foundation, although full system integration will demand further engineering efforts.

Fourth, we recognize that our current evaluation benchmarks, while insightful, have relatively low difficulty levels, enabling even small models to achieve comparably high scores. Future work will focus on developing more comprehensive, fair, and high-impact benchmark suites specifically tailored to the unique demands of lunar exploration tasks, ensuring robust and meaningful evaluations across models.

Finally, although we have demonstrated the benefits of RAG in the current system, we will continue refining its integration, including optimizing retrieval strategies, enhancing dynamic context selection, and testing RAG-enhanced models under more open-ended and complex task settings.

The Moon, a perennial nexus of human exploratory ambition, has historically seen missions characterized by substantial reliance on human expertise and direct manual intervention. The contemporary efflorescence of NLP and LLMs signifies a paradigm shift, offering transformative potential for enhancing the autonomy, safety, and scientific yield of future lunar missions, including ambitious multinational endeavors for the International Lunar Research Station (ILRS).

We fully acknowledge that bringing this profound vision to fruition will require a prolonged, intellectually challenging, and deeply collaborative research journey.

*"**We choose to go to the moon.** We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard."*

—On September 12, 1962, at Rice University, President John F. Kennedy delivered his historic speech on the nation's space effort.