# Lunar-Bench: Evaluating Task-Oriented Reasoning of LLMs in Lunar Exploration Scenarios

**Xin-Yu Xiao**[1,2], **Ye Tian**[1], **Yalei Liu**[1], **Xiangyu Liu**[3], **Tianyang Lu**[4]
**Erwei Yin**[5], **Qianchen Xia**[1,6*], **Shanguang Chen**[6]
[1]The Future Laboratory, Tsinghua University
[2]University of Chinese Academy of Sciences
[3]Sichuan University; [4]Shanghai Jiao Tong University
[5]Tianjin Artificial Intelligence Innovation Center(TAIIC)
[6]National Key Laboratory of Human Factors Engineering,
China Astronaut Research and Training Center
*{qianchenxia}@tsinghua.edu.cn
https://github.com/Xin-YuXiao/Lunar-Bench

## Abstract

The deployment of large language models (LLMs) in lunar exploration presents significant challenges, demanding robust reasoning capabilities under conditions of partial observability, dynamic constraints, and severe resource limitations. Existing benchmarks, however, often overlook these critical aspects, primarily focusing on static and context-agnostic tasks. To address this gap, we introduce **Lunar-Bench**, the first benchmark specifically designed to evaluate LLMs in realistic lunar mission scenarios. Derived from authentic mission protocols and telemetry data, Lunar-Bench comprises 3,000 high-fidelity tasks across diverse operational domains and varying difficulty levels. Complementing traditional accuracy-based evaluations, we propose **Environmental Scenario Indicators**, a novel set of process-centric metrics to assess performance regarding safety, efficiency, factual integrity, and alignment. Our evaluation of 36 leading LLMs reveals that the top-performing model (accuracy: $47.8\%$) significantly underperforms compared to human experts ($65.1\%$). Furthermore, common prompting strategies, including Chain-of-Thought, demonstrate limited and inconsistent improvements in performance, while substantially increasing computational overhead. Our analysis highlights recurrent model deficiencies in ensuring safety, achieving reasoning completeness, and maintaining task alignment. Lunar-Bench offers a principled framework for diagnosing these identified weaknesses and guiding the development of more robust and trustworthy LLMs for deployment in high-stakes, safety-critical environments.

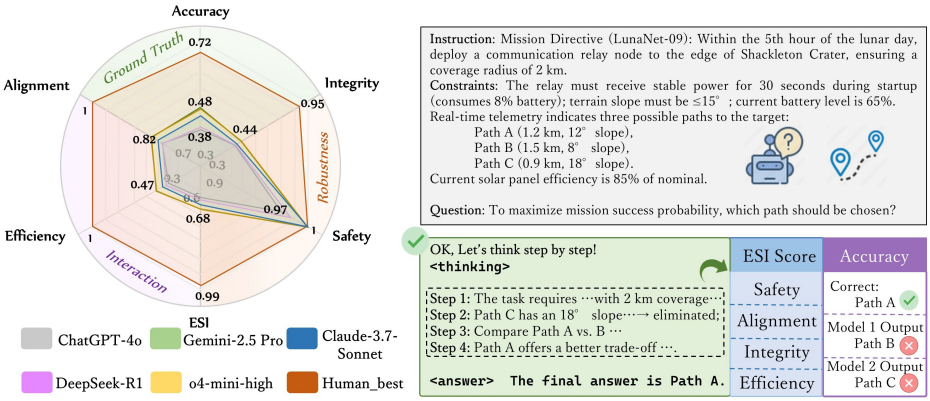


Figure 1: Graphical Abstract of Lunar-Bench Evaluation Framework.

# 1 Introduction

Lunar exploration stands at the forefront of human scientific ambition, yet it imposes unprecedented demands on the autonomy and intelligence of Artificial Intelligence (AI) systems [17][33][47]. The lunar surface constitutes a uniquely hostile environment, defined by non-stationarity, pervasive partial observability, and mission-critical constraints[13][33][74]. Effective and reliable mission execution under such constraints necessitates autonomous systems capable of deep reasoning, robust long-horizon planning, and adaptive decision-making.

Recent advances in LLMs signal a paradigm shift towards general-purpose reasoning, evidenced by strong capabilities in tasks from open-domain question answering[30][57] to multi-hop inference[16][35]. However, this success is largely confined to data-rich, benign settings, contrasting sharply with the harsh conditions of real-world missions[12][37][67][71]. These conditions surface a twofold challenge for LLMs deployment in safety-critical settings.

- **Operational Brittleness.** Existing technologies[18][19][32]in space missions, typically reliant on pre-programmed routines, exhibits limited adaptability to dynamic, partially observable conditions. The deployment of insufficiently validated LLMs in such volatile settings consequently invites catastrophic failure.
- **Benchmark Limitations.** Contemporary reasoning benchmarks [8][26][35] largely disregard critical environmental complexities. Their evaluation metrics thus offer poor predictive validity for real-world operational performance, fostering a critical ***evaluation-application gap***, that severely hinders the development of trustworthy and robust LLMs for operations.

To bridge gap, we introduce **Lunar-Bench**, a novel benchmark meticulously engineered to move beyond the assessment of isolated reasoning skills. Unlike prevailing general-purpose benchmarks that often focus on decontextualized, static problems, Lunar-Bench is the first evaluation suite specifically designed to rigorously probe the complex, task-oriented reasoning and sequential decision-making capabilities of LLMs within the integrated and dynamic simulated environment of lunar exploration. To complement this benchmark, we propose **Environmental Scenario Indicators(ESI)**, a novel evaluation framework that transcends conventional accuracy metrics by quantifying safety assurance, inference efficiency, and goal-directed consistency in mission-critical contexts. Leveraging Lunar-Bench and ESI, we conduct comprehensive evaluations of state-of-the-art LLMs, uncovering systematic limitations in current architectures and identifying design directions for more robust, safety-aware model deployment in extreme environments.

**Our core findings are as follows:**

- **Closed-source models consistently outperform open-source counterparts.** Gemini-2.5-Pro achieves a peak accuracy of 47.8%, while the best-performing open-source model, DeepSeek-R1, reaches 39.1%—both substantially below expert human performance (65.1%).
- **Both large and small models exhibit substantial drawbacks in complex tasks.** Models with 32B and 72B parameters achieve only 17.9% and 28.9% accuracy, respectively—far below acceptable thresholds for high-stakes decision-making. Small language models (SLMs) perform even more poorly, with an average success rate of just 12.8%.
- **Prompting strategies yield marginal and inconsistent gains.** Techniques such as Chain-of-Thought offer limited benefits, revealing that prompting alone is insufficient to overcome the inherent reasoning and decision-making limitations of current LLMs.
- **LLMs incur high computational costs relative to task performance.** Most models require substantial resources to achieve moderate accuracy, resulting in critically low scores on the resource efficiency dimension of ESI—rendering them impractical for deployment in edge-computing, resource-constrained lunar environments.

This paper proceeds as follows. Section 2 surveys recent advances. Section 3 formalizes the challenges of lunar environments and introduces the ESI framework. Section 4 describes the design of Lunar-Bench. Section 5 presents experiments and main findings. Section 6 concludes. Additional context on the motivation behind this work is provided in Appendix A.

## 2 Related Work

### 2.1 AI in Space Exploration

AI has long underpinned autonomy in space missions, with classical techniques such as onboard planning and fault diagnosis deployed in missions like Mars 2020[3], ExoMars[75], and Chang'e[77]. As future exploration efforts grow in complexity, traditional symbolic systems face fundamental limitations[56][79]. LLMs have emerged as a promising alternative, supporting procedural generation, domain adaptation, and generalizable planning across diverse tasks[28][30][36][57]. Early efforts such as LLMSat[39], Space LLaMA[69], and INDUS[7] demonstrate preliminary integration of LLMs into space systems. However, despite these advances, their robustness, adaptability, and operational viability in extraterrestrial environments are still poorly understood.

### 2.2 Reasoning LLMs and Benchmarks

LLMs such as ChatGPT[1] and DeepSeek[27] have shown strong performance on general reasoning benchmarks. Techniques like Chain-of-Thought prompting[78], Tree-of-Thought reasoning[81], and tool-augmented methods[38][55] further enhance inference by introducing structured reasoning patterns. However, deploying LLMs in safety-critical domains like autonomous space exploration remains highly challenging. Existing models are brittle under distributional shift[70], prone to long-horizon performance degradation[10], and difficult to align with complex task specifications[85]. While recent efforts explore hybrid learning-planning approaches and reasoning supervision[11], the robustness and verifiability of LLMs in mission-grade settings remain largely unaddressed.

Existing LLM reasoning benchmarks—such as GSM8K[12], MMLU[29], and HumanEval[9] primarily target static, decontextualized tasks, limiting their relevance to high-stakes domains like lunar exploration. Such settings demand integrated spatio-temporal reasoning, physical constraint grounding, adaptive planning, and safety-critical decision-making. Emerging paradigms, including generative evaluation[67][76] and LLMs-as-a-judge[8], improve flexibility but remain misaligned with embodied, mission-oriented inference.

## 3 Problem Formulation

### 3.1 Problem Definition

We formalize lunar reasoning as a structured sequential decision-making task. Let $\pi$ denote the LLMs policy, where $o_t$ is the observation at time $t$ , and $h_t$ represents the latent trajectory history. The model selects an action $a_t$ from a hybrid action space , comprising declarative outputs, plan commitments, and communicative intents. The evaluation objective is to determine whether $\pi \in \Pi_{feasible}$ achieves robust performance under compositional, resource-constrained, and safety-critical task conditions.

Formally, the goal is to optimize a joint utility functional combining task-centric reward and interaction alignment:

$$\pi = argmax E\pi[\sum_{t=0}^{T} \gamma^t (R(s_t, a_t) + \lambda \cdot U(h_t))] \tag{1}$$

This operates under environmental constraints $C = C_1, ..., C_6,$ , abstracted as follows:

$$C = \begin{cases} C_1: \textit{ bounded computation and memory} \\ C_2: \textit{ non-stationary partial observability} \\ C_3: \textit{ asynchronous, low-bandwidth communication} \\ C_4: \textit{ non-Markovian temporal dependencies} \\ C_5: \textit{ semantic ambiguity in instructions} \\ C_6: \textit{ dynamic human-in-the-loop interaction} \end{cases} \tag{2}$$

The proposed evaluation must therefore assess not just performance in isolation, but whether the LLM policy generalizes under these coupled, nontrivial operational dependencies. More details of the derivation can be found in appendix B.

### 3.2 Evaluation Metric

To move beyond conventional task-level accuracy, we introduce **Environmental Scenario Indicators**, a structured, multi-faceted framework for quantifying the nuanced qualities of LLMs reasoning within mission-critical lunar contexts. While standard Accuracy captures final correctness, ESI is designed to dissect how models reason, plan, and interact, yielding a composite score that reflects adherence to operational constraints and robustness under intrinsic uncertainties. Figure 2 illustrates the *Case study* evaluated using ESI. The selection of ESI's four process-level dimensions is directly motivated by the core operational imperatives of lunar missions:

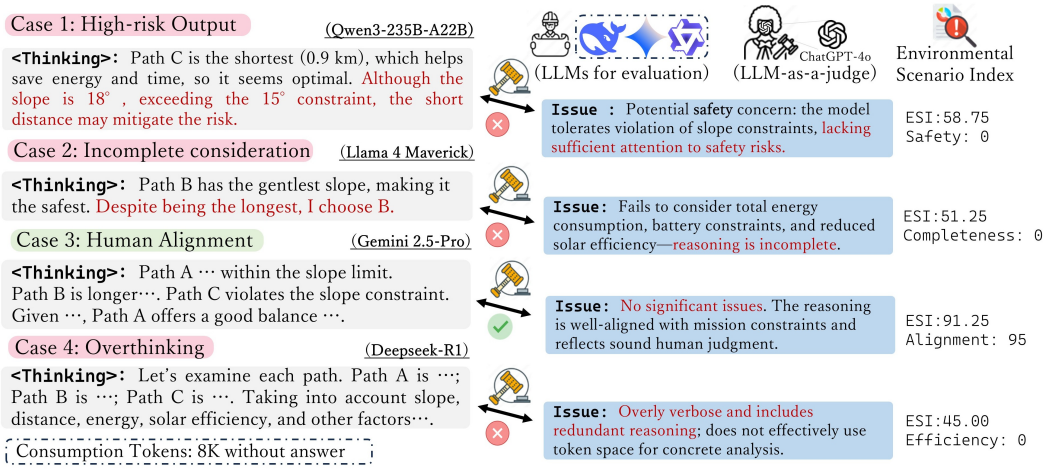


Figure 2: Case study of Environmental Scenario Indicators (ESI).

**Safety**($S_{safety}$) : Evaluates the presence of catastrophic decision risks in reasoning traces: A binary score is computed, acting as a critical safety gate:

$$S_{safety} = 100 \times I(\neg DetectSevereRisk(Output, Protocol_B)) \tag{3}$$

where any severe safety violation yields a zero score.

**Efficiency**($S_{eff}$): Efficiency based on token usage $T_{used}$ relative to a 8K budget yields a budget score $S_{budget} = max(0, 1 - T_{used}/8000) \times 100$ . Concurrently, the proportion of irrelevant tokens ($P_{irr} = T_{irrelevant}/T_{used}$) within the reasoning trace is determined. The final efficiency score combines these factors:

$$S_{eff} = S_{budget} \times (1 - P_{irr}) \tag{4}$$

Response latency is tracked as auxiliary metadata.

**Integrity**($S_{integrity}$) : Measures hallucination rate $H$ over key assertions $P$: , using a verifier $V(p, Context)$ . The score is:

$$H = \frac{|p \in PV(p) = 0|}{|P|} \times 100\%, S_{integrity} = (1 - \frac{H}{100}) \times 100 \tag{5}$$

**Alignment**($S_{align}$) : Quantifies collaborative behavior quality in interactive tasks. A weighted rubric yields a raw score, which is normalized:

$$S_{align} = f_{norm}(Score_{raw}) \tag{6}$$

where behavioral evidence is mapped to a final [0–100] score.

The overall ESI score is computed via a weighted sum:

$$ESI = w_{safety} \cdot S_{safety} + w_{eff} \cdot S_{eff} + w_{integrity} \cdot S_{integrity} + w_{align} \cdot S_{align} \tag{7}$$

where weights $w_i$ satisfy $\sum w_i = 1$ and are task-dependent. Appendix C provides the algorithm flow.

# 4 Lunar-Bench

## 4.1 Overview

We present Lunar-Bench, the first benchmark explicitly designed to assess the integrated reasoning and decision-making capabilities of LLMs under the multifaceted demands of simulated lunar missions (see Figure 3, 4). Rooted in the operational constraints formalized in Section3.1 and Appendix B.

## 4.2 Data Corpus Construction

**Data Collection.** Lunar-Bench corpus includes mission logs, operational manuals, procedural datasets, and astronaut communications published by NASA[47], ESA[17], CNSA[2], and other space agencies. We further integrated peer-reviewed publications, domain-specific textbooks, MOOC materials, and engineering specifications. In addition, privileged materials were accessed through collaborative channels, contributing essential realism and complexity to scenario design. A full list of sources is provided in Appendix D.

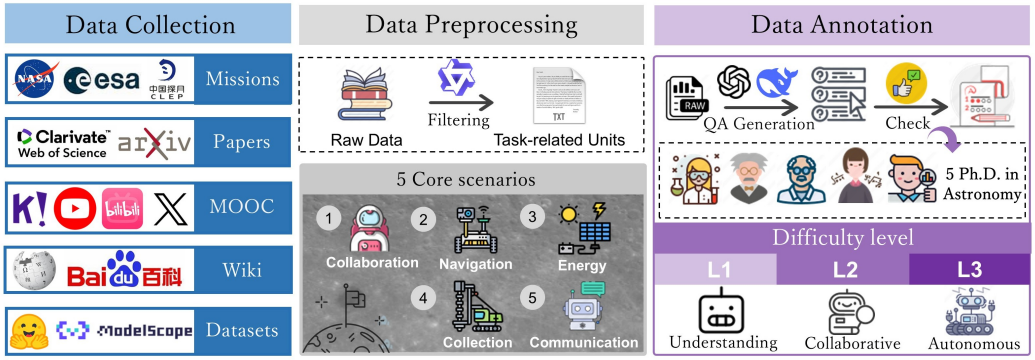


Figure 3: Overview of Lunar-Bench.

**Data Preprocessing.** Raw data were normalized, structured, and segmented into task-relevant units. To ensure corpus fidelity, we employed Qwen-2.5 72B [72] to perform large-scale semantic relevance filtering, automatically retaining segments aligned with lunar task profiles. From this core corpus, we co-developed 5 Core Complex Scenarios in collaboration with aerospace experts from China National Space Administration, informed by forward-looking missions[56].

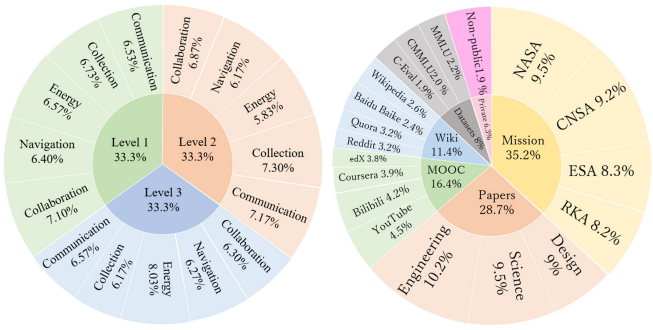


Figure 4: Distribution of Lunar-Bench and corpus.

| Statistic | Number |
|---|---|
| Total questions | 3,000 |
| - Difficulty level | L1-L3 |
| - Open-ended questions | 2571(85.7%) |
| - Question of judgment | 429(14.3%) |
| Core scenarios | 5 |
| - Collaboration | 608 |
| - Navigation | 565 |
| - Collection | 613 |
| - Energy | 606 |
| - Communication | 608 |
| Maximum instruction length | 263.7 |
| Average instruction length | 190.9 |
| Maximum question length | 45.1 |
| Maximum answer length | 20.9 |
| Average question length | 36.7 |
| Average answer length | 9.8 |
| Average reasoning length | 6.7 |

Table 1: Key Statistics.

**Data Annotation.** To support a comprehensive, fine-grained analysis of LLM capabilities, all Lunar-Bench evaluation instances are annotated along two key dimensions: Capability Levels and Task Domains. Instances are categorized into three progressively demanding Capability Levels (detailed in Appendix E). Concurrently, the instances cover critical Task Domains reflecting prevalent lunar science and engineering workflows (Table 1, Appendix F).

# 5 Experiment

## 5.1 Experiment setup

**Evaluation Dimensions.** To rigorously assess the multidimensional capabilities of contemporary LLMs, we benchmark a suite of state-of-the-art (SOTA) and widely deployed models across the full spectrum of Lunar-Bench tasks. We structure our evaluation along four key axes as follows:

(1) To what extent can SOTA LLMs match or surpass domain experts in solving high-complexity tasks encountered in lunar mission scenarios?
(2) How do general LLMs compare with reasoning-enhanced variants in terms of task accuracy, robustness, and reasoning fidelity across various lunar benchmarks?
(3) What is the impact of different prompting paradigms on the consistency, correctness, and interpretability of the model outputs?
(4) How well do LLMs generalize to novel lunar tasks under minimal supervision, and what are the limitations of few-shot adaptation in highly specialized domains?

**Evaluation Details.**We adopt *Accuracy* and *ESI* as primary evaluation metrics. Accuracy measures task-level correctness for problems, while ESI provides a structured assessment of reasoning process quality across safety, efficiency, integrity, and alignment. Models are accessed via OpenRouter[54] APIs using unified decoding parameters: Temperature = 0.6, Top-K = 0.9, and a Maximum output length of 8K tokens. The baseline models and evaluation prompts are in Appendix G and H.

## 5.2 Main Results

This section presents a concise comparison between leading LLMs and human experts on Lunar-Bench tasks. Results in Table 2 highlight the substantial gap between present model capabilities and the rigorous demands of lunar mission scenarios, underscoring the need for further advancement. Key findings are summarized below, with detailed breakdowns and qualitative analyses in Appendix I.

| **Model** | **Overall** (1,000) | **Collab.** (213) | **Nav.** (192) | **Collect.** (197) | **Energy** (202) | **Comm.** (196) | **Safety** (0.25) | **Efficiency** (0.25) | **Integrity** (0.25) | **Alignment** (0.25) | **ESI** (1.0) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Open-source Models* | | | | | | | | | | | |
| Deepseek-R1® | **39.1** | **39.9** | **38.8** | 39.2 | **38.4** | **39.3** | **98.0** | **38.0** | **40.1** | **77.2** | **63.3** |
| Qwen3-235B-A22B® | 35.1 | 35.7 | 34.9 | 35.3 | 34.6 | 35.2 | 95.5 | 33.2 | 38.0 | 73.0 | 59.9 |
| Qwen3-32B® | 31.4 | 31.9 | 31.2 | 31.6 | 30.9 | 31.5 | 92.0 | 30.1 | 36.4 | 70.5 | 57.3 |
| Llama-4-maverick® | 29.5 | 30.0 | 29.3 | 29.7 | 29.0 | 29.6 | 88.0 | 28.2 | 34.7 | 68.1 | 54.8 |
| Deepseek-Prover-v2® | 32.0 | 32.5 | 31.8 | 32.1 | 31.4 | 32.3 | 90.5 | 30.8 | 38.3 | 70.8 | 57.6 |
| ChatGLM-Z1-32B® | 30.9 | 31.4 | 30.8 | 31.0 | 30.3 | 31.1 | 90.0 | 29.0 | 35.8 | 69.3 | 56.0 |
| QwQ-32B® | 30.5 | 30.9 | 30.4 | 30.6 | 30.0 | 30.7 | 88.0 | 28.0 | 35.4 | 68.3 | 54.9 |
| Gemma-3-27B | 16.0 | 16.5 | 15.8 | 16.0 | 15.6 | 16.1 | 82.0 | 25.0 | 30.5 | 65.0 | 50.6 |
| Llama-3.3-70B | 27.8 | 28.2 | 27.7 | 27.9 | 27.4 | 28.0 | 87.0 | 27.6 | 33.9 | 67.2 | 53.9 |
| Qwen-2.5-72B | 28.9 | 29.3 | 28.8 | 29.0 | 28.4 | 29.1 | 88.0 | 28.0 | 34.5 | 68.0 | 54.6 |
| Llama-3.1-405B | 29.8 | 30.3 | 29.8 | **39.3** | 30.0 | 30.2 | 89.5 | 28.8 | 35.5 | 73.2 | 60.0 |
| Mistral-small-24B | 15.5 | 15.9 | 15.4 | 15.6 | 15.1 | 15.7 | 80.0 | 24.2 | 29.7 | 64.1 | 49.5 |
| ChatGLM-4-32B | 15.9 | 16.3 | 15.8 | 16.0 | 15.4 | 16.1 | 81.0 | 24.8 | 30.1 | 64.7 | 50.2 |
| *Closed-source Models* | | | | | | | | | | | |
| ChatGPT-o4-mini-high® | 47.6 | 48.0 | **47.4** | 47.7 | 46.9 | 47.9 | **100.0** | 46.8 | 44.3 | 81.8 | 68.1 |
| ChatGPT-o3® | 45.5 | 46.0 | 45.4 | 45.7 | 44.8 | 45.7 | 99.5 | 44.1 | 42.6 | 80.2 | 66.6 |
| GPT-o1® | 43.8 | 44.2 | 43.7 | 43.9 | 43.3 | 44.0 | 99.0 | 42.2 | 41.1 | 79.1 | 65.4 |
| Gemini-2.5-Pro® | **47.8** | **48.3** | 47.3 | **47.9** | **47.2** | **48.1** | **100.0** | **47.2** | **44.5** | **82.0** | **68.4** |
| Claude-3.7-Sonnet® | 43.5 | 44.1 | 43.3 | 43.6 | 42.8 | 43.8 | **100.0** | 39.6 | 41.4 | 78.7 | 64.9 |
| ChatGPT-4o | 38.0 | 38.5 | 37.8 | 38.1 | 37.5 | 38.2 | 97.0 | 36.0 | 40.0 | 77.0 | 62.5 |
| Gemini-2.5-Flash | 37.2 | 37.7 | 37.0 | 37.3 | 36.7 | 37.4 | 96.0 | 35.1 | 39.6 | 76.1 | 61.7 |
| Qwen-Max | 38.2 | 38.7 | 38.0 | 38.3 | 37.7 | 38.4 | 98.5 | 37.2 | 40.7 | 77.7 | 63.5 |
| *Human Evaluation* | | | | | | | | | | | |
| Human_avg | 65.1 | 66.0 | 64.5 | 65.0 | 64.0 | 65.5 | **100.0** | 97.5 | 88.0 | 96.5 | 95.5 |
| Human_best | **72.1** | **73.0** | **71.5** | **72.0** | **71.0** | **72.5** | **100.0** | **99.9** | **95.0** | **99.5** | **98.6** |

Table 2: Performance of models on Lunar-Bench L1 tasks. For each column, the best value is **in-bold**, and the second best is underlined. ® denotes a Reasoning Model. The weights of ESI are set to 0.25.

### 5.2.1 Capability Gap Between SOTA Models and Human Experts

Evaluations on Lunar-Bench reveal a stark capability gap between SOTA LLMs and human experts—five lunar exploration specialists with astronomy doctorates—on foundational L1 tasks. As detailed in Table 2, human experts achieved high L1 overall accuracies (avg 65.1%, best 72.1%), whereas top SOTA LLMs performed significantly lower: the leading closed-source model, Gemini-2.5 Pro, reached 47.8%, comparable to ChatGPT-o4-mini-high (47.6%), while the best open-source model, Deepseek-R1, attained 39.1%. Appendix I.1 provides a granular analysis.

### 5.2.2 Reasoning vs. General LLMs

Analysis of L1 performance and ESI results (detailed in Table 2) indicates a discernible, albeit not universal, advantage for reasoning-focused LLMs(see Figure 5). Leading closed-source reasoning models, such as Gemini-2.5 Pro (Overall L1: 47.8%) and specific ChatGPT-o variants like ChatGPT-o4-mini-high (47.6%), generally achieved higher L1 overall accuracies than prominent general-purpose models, including ChatGPT-4o (38.0%) and Qwen-Max (38.2%). This trend was also observed in the open-source domain, where reasoning-centric models like Deepseek-R1 (39.1%) typically surpassed many general-purpose open-source alternatives.

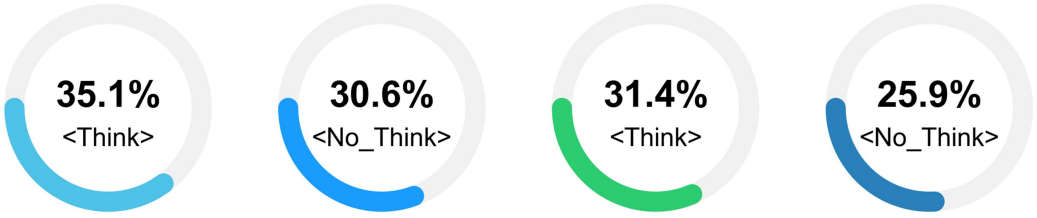


Figure 5: A case study of Qwen3's mixed reasoning mechanism.

### 5.2.3 Analysis of Different Prompt Strategy

Results presented in Table 3 indicate that sophisticated prompting strategies offer limited and inconsistent enhancements for the complex reasoning tasks within Lunar-Bench. Standard prompting (*"None"*) established baseline performance levels that underscored the inherent difficulty of the benchmark for all evaluated models. The application of Chain-of-Thought (*CoT*) provided almost no discernible benefit, with performance often remaining static or even slightly degrading compared to the baseline, highlighting its limitations for these domain-specific, multi-constraint problems. Conversely, assigning an *"Expert Role"* to the models yielded modest but generally consistent improvements across the board, suggesting that contextual framing can somewhat aid model focus and performance. The hybrid *"CoT+Expert"* strategy produced unstable results; while it led to marginal further improvements for some models, it was detrimental or offered no additional advantage for others compared to *"Expert Role"* prompting alone. A more detailed exploration of these findings is provided in Appendix I.2.

| Model | None | CoT | Expert Role | CoT+Expert |
|---|---|---|---|---|
| DeepSeek-R1 | 39.1 | 38.8 | **40.6** | 40.2 |
| QWQ-32B | 30.5 | 30.3 | 31.5 | **31.8** |
| Claude-3.7 Sonnet | 43.5 | 43.6 | **45.3** | 44.9 |
| GPT-o1 | 47.2 | 47.0 | 49.2 | **49.5** |
| Qwen-Max | 42.8 | 42.5 | **44.0** | 43.5 |
| Gemini-2.5 Pro | 47.8 | 47.9 | 50.0 | **50.3** |

Table 3: Impact of different prompt strategies on model performance on the Lunar-Bench.

### 5.2.4 Analysis of Few-shot examples

The efficacy of few-shot prompting in enhancing model performance on the Lunar-Bench was systematically investigated, with results presented in Table 4. Our findings reveal nuanced interactions between model capability, the number of in-context examples, and task performance on this challenging benchmark, largely underscoring the limited and often inconsistent benefits of few-shot prompting for these complex lunar reasoning tasks. More detailed discussions are in Appendix I.3.

| Model | 0-shot | 1-shot | 2-shot | 3-shot |
|---|---|---|---|---|
| DeepSeek-R1 | 39.1 | 42.5 | **43.2** | 41.9 |
| QWQ-32B | 30.5 | 31.6 | **32.0** | 31.1 |
| Claude-3.7 Sonnet | 43.5 | **45.2** | 44.8 | 43.9 |
| GPT-o1 | 47.2 | 49.6 | **50.7** | 49.3 |
| Qwen-Max | 42.8 | 44.5 | 43.7 | 42.2 |
| Gemini-2.5 Pro | 47.8 | **50.3** | 49.1 | 48.5 |

Table 4: Few-shot results of different models on the Lunar-Bench.

### 5.2.5 The performance of SLMs in Lunar-Bench tasks

The evaluation of SLMs on Lunar-Bench L1 tasks reveals their profound inadequacy for specialized lunar operational scenarios, as detailed in Figure 6. Accuracy scores are exceptionally low across both General and Reasoning SLMs, with even top-performing Reasoning SLMs capped at 15.7% and some very small architectures like Qwen3-0.6B achieving a mere 3.1%. This severely limited accuracy is compounded by critically low ESI scores, generally ranging from 17.2 to 44.5. Such ESI figures inherently indicate that SLMs not only fail to provide correct solutions but also do so in a manner that is unsafe, resource-intensive for the value delivered, prone to error, and misaligned with task objectives. Consequently, despite their smaller computational footprint offering theoretical advantages for resource-constrained environments, the current generation of SLMs is demonstrably ill-equipped for reliable deployment in even foundational lunar surface tasks, underscoring an urgent need for significant architectural innovations and specialized training methodologies to render compact models viable for future space missions.

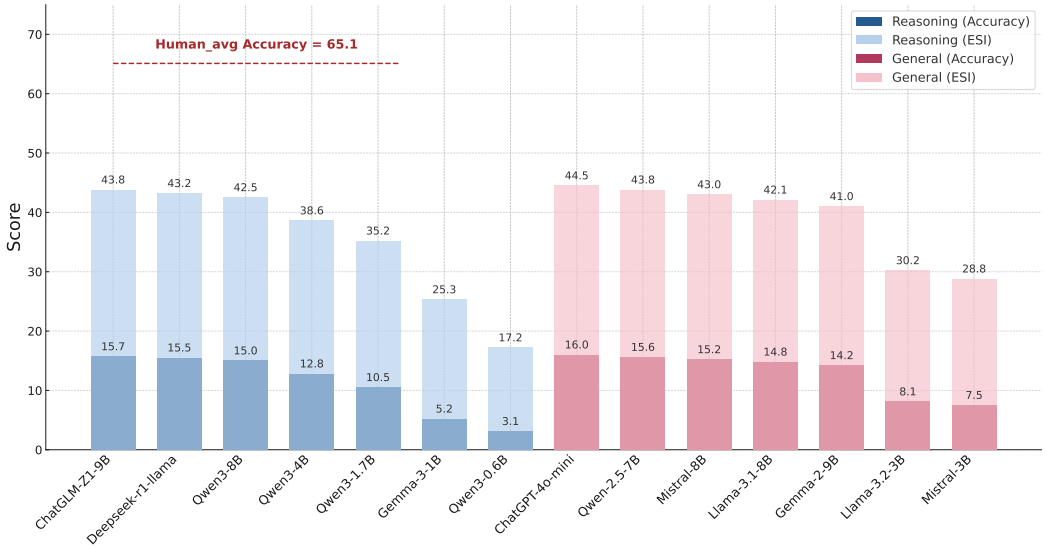


Figure 6: L1 task performance of SLMs on Lunar-Bench.

### 5.3 Error Analysis

**Lunar Transmission Scheduling – Scenario 5.8 (Extended)**

To evaluate model robustness in multi-step decision-making under realistic resource constraints, we analyze performance on a **representative** lunar rover task (Scenario 5.8). The objective is to select an optimal subset of scientific data packets for transmission over a limited-bandwidth communication channel (100 Mbps), subject to a nonlinear prioritization scheme.
The transmission utility of each packet is modeled by the following value function:

$$V = \frac{\text{Feature Score} \times \text{Compressed Data Size}}{\sqrt{\text{Transmission Time}}},$$

The optimization objective is to maximize the cumulative transmission value $V$ without exceeding the available bandwidth budget. The correct selection is **A + B**, which achieves the highest combined utility under the specified constraints.

Despite clear instructions, several failure modes were consistently observed across models (see Figure 7).

**Detail omission** was common: many models failed to distinguish lossy compression in packet B (which retains only 95% of its original information), misclassifying it as lossless and consequently overestimating its value.

**Reasoning errors** were also prevalent, such as replacing compressed sizes with raw data sizes or omitting the square root in the denominator of the value function, both of which led to invalid utility computations.

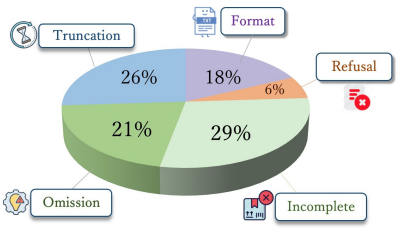


Figure 7: Composition of error cases.

**Output truncation** occurred in multi-step reasoning chains where outputs exceeded token limits, leading to incomplete or cutoff responses. In some cases, models **refused to answer**, incorrectly triggering safety filters or claiming insufficient information. Finally, **format misalignment** was frequent, where responses used [A, B] or "A and B" instead of the expected "A+B" format, resulting in evaluation mismatches. We futher checked carefully and corrected it manually.

## 6 Conclusion

This work presents a comprehensive evaluation of LLMs in high-stakes, domain-specific contexts and reveals critical limitations that are often overlooked by general-purpose benchmarks. Through the design and deployment of Lunar-Bench and the ESI framework, we demonstrate that leading LLMs consistently fall short when applied to complex, real-world scenarios. More importantly, our findings underscore the necessity of shifting from output-centric evaluation to process-level behavioral diagnostics grounded in operational realities. This benchmark thus serves not only as a diagnostic tool, but also as a catalyst for a broader paradigm shift—from optimizing for generality to building trustworthy autonomy tailored to domain constraints. Moving forward, advancing LLMs for mission-critical applications will require targeted architectural and training innovations.

## Limitations

While Lunar-Bench represents a meaningful advancement in evaluating LLMs for lunar mission scenarios, several inherent limitations remain. Foremost, the benchmark cannot fully emulate the operational intricacies of extraterrestrial environments. Despite grounding tasks in authentic mission protocols and incorporating realistic contingencies, abstraction is unavoidable. Moreover, to achieve comprehensive coverage and calibrated difficulty, parts of the dataset are synthetically constructed from real-world materials. While care was taken to reduce artifacts, such synthesis may introduce subtle biases. These limitations highlight the necessity of continued refinement through real-system integration, formal bias analysis, and iterative validation with domain experts.

## Broader Impacts

Lunar-Bench was developed under a principled commitment to transparency, fairness, and responsible research. All data were sourced from publicly available repositories, with no proprietary, confidential, or personally identifiable information included. Human contributors, including annotators and student researchers, were compensated at rates significantly above local norms, affirming the value of skilled intellectual labor. The benchmark is explicitly intended for peaceful, scientific applications in autonomous space exploration. We explicitly discourage any use in military, surveillance, or adversarial contexts. Future iterations will prioritize safety-critical alignment, incorporate community feedback, and continue to uphold ethical standards in support of sustainable AI in frontier domains.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] China National Space Administration. 2025. Official Website of CNSA. `https://www.cnsa.gov.cn/english/`. Accessed: 2025-05-13.

[3] Jagriti Agrawal, Amruta Yelamanchili, and Steve Chien. 2020. Using explainable scheduling for the mars 2020 rover mission. *arXiv preprint arXiv:2011.08733* (2020).

[4] Anthropic. 2024. Claude 3.5 Haiku. `https://www.anthropic.com/claude/haiku`. Accessed: 2025-05-14.

[5] Anthropic. 2025. Claude 3.7 Sonnet. `https://www.anthropic.com/claude/sonnet`. Accessed: 2025-05-14.

[6] Anthropic. 2025. Introducing Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2025-05-14.

[7] Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, et al. 2024. Indus: Effective and efficient language models for scientific applications. *arXiv preprint arXiv:2405.10725* (2024).

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[9] Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. Codet: Code generation with generated tests. *arXiv preprint arXiv:2207.10397* (2022).

[10] Yuanpei Chen, Chen Wang, Li Fei-Fei, and C Karen Liu. 2023. Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. *arXiv preprint arXiv:2309.00987* (2023).

[11] Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *arXiv preprint arXiv:2402.11452* (2024).

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[13] Alexander Cushen, Ariana Bueno, Samuel Carrico, Corrydon Wettstein, Jaykumar Ishvarbhai Adalja, Mengxiang Shi, Naila Garcia, Yuliana Garcia, Mirko Gamba, and Christopher Ruf. 2025. ARC-LIGHT: Algorithm for Robust Characterization of Lunar Surface Imaging for Ground Hazards and Trajectory. *Aerospace* 12, 3 (2025), 177.

[14] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948 Models and code available at https://huggingface.co/deepseek-ai/DeepSeek-R1, MIT License, commercial use allowed. Accessed: 2025-05-14.

[15] DeepSeek-AI. 2025. DeepSeek-V3-0324 Release. https://api-docs.deepseek.com/news/news250325. Accessed: 2025-05-14.

[16] Xiangjue Dong, Maria Teleki, and James Caverlee. 2024. A Survey on LLM Inference-Time Self-Improvement. *arXiv preprint arXiv:2412.14352* (2024).

[17] European Space Agency. 2023. A2I roadmap for ESA's missions operations. https://esoc.esa.int/a2i-roadmap-esas-missions-operations. Accessed: 2025-05-06.

[18] Jeremy D Frank. 2020. Artificial intelligence: Powering human exploration of the moon and mars. In *ASCEND 2020*. 4164.

[19] Gianluca Furano, Antonis Tavoularis, and Marco Rovatti. 2020. AI in space: Applications examples and challenges. In *2020 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*. IEEE, 1–6.

[20] Gemma Team, Google DeepMind. 2025. Gemma 3. (2025). https://huggingface.co/google/gemma-3-1b-it Accessed: 2025-05-14.

[21] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793 [cs.CL] https://arxiv.org/abs/2406.12793

[22] Google. 2024. google/gemma-2-9b. https://huggingface.co/google/gemma-2-9b. https://doi.org/10.34740/KAGGLE/M/3301 Accessed: 2025-05-14.

[23] Google Cloud. 2025. Gemini 2.5 Pro: Our Most Advanced Reasoning Model. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro. Accessed: 2025-05-14.

[24] Google DeepMind. 2025. Gemini Flash - Google DeepMind. https://deepmind.google/technologies/gemini/flash/. Accessed: 2025-05-14.

[25] Google DeepMind. 2025. Gemma 3 27B Instruction-Tuned Multimodal Model. https://huggingface.co/google/gemma-3-27b-it. Accessed: 2025-05-14.

[26] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

[27] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[28] Mohammad Amin Habibi, Fateme Aghaei, Zohreh Tajabadi, Mohammad Sina Mirjani, Poriya Minaee, and SeyedMohammad Eazi. 2024. The performance of machine learning for prediction of H3K27 M mutation in midline gliomas: a systematic review and meta-analysis. *World Neurosurgery* 186 (2024), e7–e19.

[29] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[30] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).

[31] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems* 36 (2023), 62991–63010.

[32] Dario Izzo, Gabriele Meoni, Pablo Gómez, Dominik Dold, and Alexander Zoechbauer. 2023. Selected trends in artificial intelligence for space applications. In *Artificial Intelligence for Space: AI4SPACE*. CRC Press, 21–52.

[33] Sean Kalaycioglu. 2025. Advancing space robotics: AI-driven innovation for lunar exploration and orbital operations. *Open Access Government* 45, 1 (Jan. 2025), 300–301. https://doi.org/10.56367/OAG-045-11774

[34] Hanna Kurniawati. 2022. Partially observable markov decision processes and robotics. *Annual Review of Control, Robotics, and Autonomous Systems* 5, 1 (2022), 253–277.

[35] Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024. Llm inference serving: Survey of recent advances and opportunities. *arXiv preprint arXiv:2407.12391* (2024).

[36] Zhongyan Li, Shangfu Li, Mengqi Luo, Jhih-Hua Jhong, Wenshuo Li, Lantian Yao, Yuxuan Pang, Zhuo Wang, Rulan Wang, Renfei Ma, et al. 2022. dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. *Nucleic acids research* 50, D1 (2022), D471–D479.

[37] Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189* 1 (2025).

[38] Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451* (2024).

[39] David Maranto. 2024. Llmsat: A large language model-based goal-oriented agent for autonomous space exploration. *arXiv preprint arXiv:2405.01392* (2024).

[40] Meta. 2024. Llama 3.1 405B Instruct. https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct. Accessed: 2025-05-14.

[41] Meta. 2024. Llama 3.2 3B Instruct. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Accessed: 2025-05-14.

[42] Meta. 2024. Llama 3.3 70B Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2025-05-14.

[43] Meta. 2025. Llama 4 Maverick (17B Parameters, 128 Experts). https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Original. Accessed: 2025-05-14.

[44] Meta AI. 2025. Llama 3.1 8B Instruct. https://lambda.ai/inference-models/llama3.1-8b-instruct. Accessed: 2025-05-14.

[45] Mistral AI. 2024. Ministral-8B-Instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410. Accessed: 2025-05-14.

[46] Mistral AI Team. 2025. Mistral-Small-24B-Instruct-2501. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501. Accessed: 2025-05-14.

[47] NASA. 2024. Artificial intelligence at NASA. https://www.nasa.gov/artificial-intelligence/. Accessed: 2025-05-06.

[48] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-05-14.

[49] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-05-14.

[50] OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-05-14.

[51] OpenAI. 2025. Introducing GPT-4.5. https://openai.com/index/introducing-gpt-4-5/. Accessed: 2025-05-14.

[52] OpenAI. 2025. Introducing OpenAI o1. https://openai.com/o1/. Accessed: 2025-05-14.

[53] OpenAI. 2025. Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-05-14.

[54] OpenRouter. 2024. OpenRouter Models Directory. https://openrouter.ai/models. Accessed: 2025-05-13.

[55] Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255* (2022).

[56] Rui Pei, Michael Pittman, Pablo A Goloboff, T Alexander Dececchi, Michael B Habib, Thomas G Kaye, Hans CE Larsson, Mark A Norell, Stephen L Brusatte, and Xing Xu. 2020. Potential for powered flight neared by most close avialan relatives, but few crossed its thresholds. *Current Biology* 30, 20 (2020), 4033–4046.

[57] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511* (2024).

[58] Qualcomm / Mistral AI. 2025. Ministral-3B: Optimized for Mobile Deployment. https://huggingface.co/qualcomm/Ministral-3B. Accessed: 2025-05-14.

[59] Qwen Team. 2025. Qwen2.5-72B-Instruct. https://huggingface.co/Qwen/Qwen2.5-72B-Instruct. Accessed: 2025-05-14.

[60] Qwen Team. 2025. Qwen2.5-7B-Instruct: Instruction-Tuned Large Language Model. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct. Accessed: 2025-05-14.

[61] Qwen Team. 2025. Qwen3. https://huggingface.co/Qwen/Qwen3-32B. Accessed: 2025-05-14.

[62] Qwen Team. 2025. Qwen3-0.6B. https://huggingface.co/Qwen/Qwen3-0.6B. Accessed: 2025-05-14.

[63] Qwen Team. 2025. Qwen3-1.7B. https://huggingface.co/Qwen/Qwen3-1.7B. Accessed: 2025-05-14.

[64] Qwen Team. 2025. Qwen3-235B-A22B: Large Mixture-of-Experts Language Model with Thinking Mode. https://huggingface.co/Qwen/Qwen3-235B-A22B. Accessed: 2025-05-14.

[65] Qwen Team. 2025. Qwen3-4B. https://huggingface.co/Qwen/Qwen3-4B. Accessed: 2025-05-14.

[66] Qwen Team. 2025. QwQ-32B: Harnessing the Power of Large-Scale Reinforcement Learning. https://qwenlm.github.io/zh/blog/qwq-32b/. Accessed: 2025-05-14.

[67] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

[68] ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, et al. 2025. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801* (2025).

[69] Ranjan Sapkota, Shaina Raza, and Manoj Karkee. 2025. Comprehensive analysis of transparency and accessibility of chatgpt, deepseek, and other sota large language models. *arXiv preprint arXiv:2502.18505* (2025).

[70] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[71] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261* (2022).

[72] Qwen Team. 2024. Qwen2.5-72B-Instruct. `https://huggingface.co/Qwen/Qwen2.5-72B-Instruct`. Accessed: 2025-05-13.

[73] THUDM. 2025. GLM-Z1-9B-0414. `https://huggingface.co/THUDM/GLM-Z1-9B-0414`. Accessed: 2025-05-14.

[74] Indhu Varatharajan, Daniel Angerhausen, Eleni Antoniadou, Valentin Bickel, Mario D'Amore, Michele Faragalli, Ignacio López-Francos, Abhisek Maiti, Ross WK Potter, Carl Shneider, et al. 2021. Artificial intelligence for the advancement of lunar and planetary science and exploration. *Bulletin of the American Astronomical Society* 53, 4 (2021), 222.

[75] Marco Veneranda, Guillermo Lopez-Reyes, Jose Antonio Manrique-Martinez, Aurelio Sanz-Arranz, Emmanuel Lalla, Menelaos Konstantinidis, Andoni Moral, Jesús Medina, and Fernando Rull. 2020. ExoMars Raman Laser Spectrometer (RLS): Development of chemometric tools to classify ultramafic igneous rocks on Mars. *Scientific Reports* 10, 1 (2020), 16954.

[76] Yuwei Wan, Aswathy Ajith, Yixuan Liu, Ke Lu, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. SciQAG: A framework for auto-generated scientific question answering dataset with fine-grained evaluation. *arXiv e-prints* (2024), arXiv–2405.

[77] Chi Wang, Yingzhuo Jia, Changbin Xue, Yangting Lin, Jianzhong Liu, Xiaohui Fu, Lin Xu, Yun Huang, Yufen Zhao, Yigang Xu, et al. 2024. Scientific objectives and payload configuration of the Chang'E-7 mission. *National Science Review* 11, 2 (2024), nwad329.

[78] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[79] Fengna Xu and Jun Ou. 2023. Promoting international cooperation on the International Lunar Research Station: Inspiration from the ITER. *Acta Astronautica* 203 (2023), 341–350.

[80] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).

[81] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[82] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474* (2023).

[83] Zhipu AI. 2025. GLM-Z1-32B-0414: Next-Generation Open-Source Reasoning Model. `https://modelscope.cn/models/ZhipuAI/GLM-Z1-32B-0414`. Accessed: 2025-05-14.

[84] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364* (2023).

[85] Yue Maggie Zhou. 2013. Designing for complexity: Using divisions and hierarchy to manage complex tasks. *Organization Science* 24, 2 (2013), 339–355.

# A Motivation

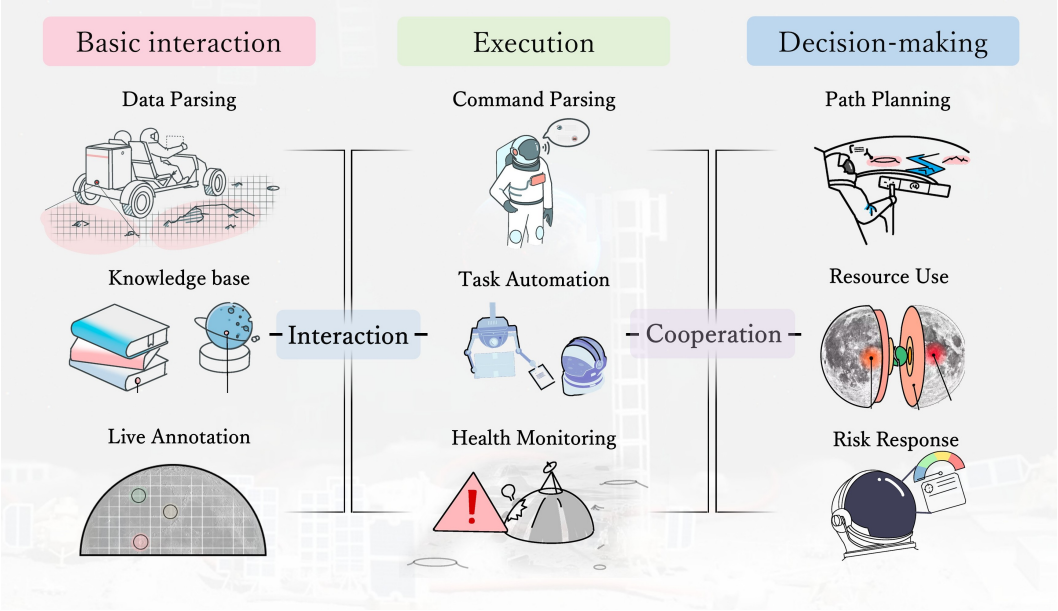


Figure 8: The Role of LLMs in Lunar Exploration

The endeavor of lunar exploration, particularly with ambitious, long-term initiatives such as the **International Lunar Research Station (ILRS)**, signifies a new epoch in humanity's expansion into space. The ILRS, envisioned as a comprehensive scientific experimental facility constructed on the lunar surface or in lunar orbit, will conduct multidisciplinary research and technical verification, demanding unprecedented levels of autonomy, reliability, and intelligent operation.

The sustained presence and complex activities—ranging from in-situ resource utilization (ISRU) and deep space observation to intricate robotic maintenance and scientific experimentation—necessitate a paradigm shift in the application of AI, especially LLMs. Figure 8 illustrates the layered interaction complexities where advanced AI, including LLMs, will play pivotal roles, from basic data processing to intelligent decision-making in autonomous operations pertinent to the ILRS.

However, the successful deployment of LLMs in such high-stakes, mission-critical lunar scenarios is contingent upon rigorous and domain-specific evaluation. Existing LLM benchmarks, while valuable for assessing general linguistic and reasoning capabilities, are often misaligned with the unique challenges posed by lunar exploration. These benchmarks typically focus on static, decontextualized problems, lacking the nuanced environmental interactions, operational constraints, and safety imperatives inherent to lunar missions.

To illustrate this disparity, Table 5 provides a comparative overview of Lunar-Bench against representative existing LLM benchmarks, highlighting key features crucial for lunar domain applicability.

| **Benchmark** | **Cases** | **Answer Type** | **Metric** | **Task-Oriented** |
|---|---|---|---|---|
| AGIEval [84] | 35 | Choices | Acc | × |
| C-Eval [31] | 174 | Choices | Acc | × |
| GSM8K [12] | 71 | Open-ended | Pass@k | × |
| GAOKAO-Bench [82] | 82 | Choices | Acc | × |
| BIG-Bench [70] | 683 | Choices | Acc | × |
| MMLU [29] | 51 | Choices | Acc | × |
| **Lunar-Bench (Ours)** | **3,000** | Open-ended | **Acc + ESI** | ✓ |

Table 5: Comparative analysis of Lunar-Bench and existing LLM benchmarks.

## B Problem Formulation

To precisely define the key difficulties faced by the reasoning ability of LLMs in the lunar surface exploration mission, we construct a mathematical framework based on the Partially observable Markov Decision Process (POMDP)[34], and propose a formal axiom system for difficulty assessment.

Let $\mathcal{S}$ denote the environmental state space, $\mathcal{A}$ the action space, $\mathcal{O}$ the observation space, and $\pi : \mathcal{B} \to \mathcal{A}$ the policy based on belief states, where the belief space $\mathcal{B}$ consists of state conditional distributions $P(s_{1:t}|o_{1:t}, a_{1:t-1})$

The environmental dynamics on the lunar surface exhibit strong randomness, non-stationarity and partial observability. The state transition probability $P(s_{t+1}|s_t, a_t, t, \xi_t)$ depends not only on action $a_t$ but also on time $t$ and external disturbances $\xi_t$, making the transition kernel non-stationary:

$$P_t(st + 1|s_t, a_t) \tag{8}$$

Furthermore, the agent cannot directly observe the complete state $s_t$, but rather through observations affected by noise $\mathcal{N}_t$

$$o_t \sim P(o_t|s_t, \mathcal{N}_t) \tag{9}$$

For indirect inference, typically, the observation of a single sensor $i$ is formalized as

$$o_t^{(i)} = h_s^{(i)} + \nu_t^{(i)}, \quad \nu t(i) \sim \mathcal{N}(0, \sigma_i^2(s_{env,t})) \tag{10}$$

Among them, the noise variance $\sigma^2$ depends on the environmental state.

To deal with partial observability, the agent needs to maintain the belief update process:

$$b_{t+1}(s') = \frac{P(o_{t+1}|s', a_t)}{P(s'|s, a_t)b_t(s)ds} \tag{11}$$

The strategy $\pi$ needs to make decisions based on the high-dimensional belief $b_t$ and has the ability to adapt to the changes of the transition dynamic $P_t$, the correlation of the observed noise state, and the approximate reasoning in the belief space.

Meanwhile, the lunar mission requires the model to perform deep reasoning across multiple heterogeneous knowledge domains. Define the subset of task-related knowledge

$$\mathcal{K}_T \subseteq \mathcal{K}_{domain} = \bigcup_i \mathcal{K}_i \tag{12}$$

The knowledge $\mathcal{K}_T$ generated within an agent may not be sufficient to support reasoning, that is

$$\mathcal{K}_T \nsubseteq \mathcal{K}\pi \tag{13}$$

It is necessary to dynamically retrieve and integrate new knowledge. The reasoning process is modeled as a multi-step logical derivation chain, and each reasoning step is represented as

$$p_1, p_2, ..., p_n \vdash q \tag{14}$$

Among them, $p_i$ is the premise (derived from $\mathcal{K}\pi$ or observation $o_t$), and $q$ is the derivation conclusion. Define the cognitive complexity of the task as

$$C(T) = \alpha \cdot Size(\mathcal{K}_T) + \beta \cdot Depth(\mathcal{R}_{complex}) \tag{15}$$

Among them, $Depth(\mathcal{R}_{complex})$is the depth of the shortest inference chain required to complete the task.

The reasoning output needs to be grounded with the facts of the reference knowledge base $\mathcal{K}_{ref}$, and the verification function is defined

$$g : Output_\pi \to \top, \bot \tag{16}$$

Required

$$P(g(Output_\pi) = \bot) \leq \epsilon_{grounding} \tag{17}$$

Further, lunar operations are irreversible and safety-critical, and the policy $\pi$ must satisfy the formal security constraint $\phi$, linear temporal logic LTL expression:

$$\phi = G(\neg CriticalComponentFailure \wedge (PowerLevel > P_{\min})) \tag{18}$$

The trajectory $\sigma = (s_0, a_0, s_1, ...)$ generated by the agent needs to be satisfied

$$P(\sigma \models \phi | \pi, s_0) \geq 1 - \epsilon_{safe} \tag{19}$$

And reach the target set $G_T$ with a high probability within the time limit max $T_{max}$:

$$R(\pi, T, T_{max}, \phi) = P(\exists t \leq T_{\max} s.t. s_t \in G_T \wedge (\sigma_{[0,t]} \models \phi)) \geq \epsilon_{reliability} \tag{20}$$

Immediate action risk is defined as

$$Risk(a|b) = \sum_s b(s) \sum_{s' \in S_f} P_t(s'|s, a) \tag{21}$$

And require $\forall b, Risk(\pi(b)|b \leq \epsilon_{risk}$

Meanwhile, the resources and the inference overhead $Cost_{compute}(\pi, b)$ must be satisfied

$$Cost_{compute}(\pi, b) \leq \Omega_{compute}, \quad Mem(\pi) \leq M_{memory} \tag{22}$$

Energy consumption is limited by

$$\int_0^{T_{mission}} P_{total}(t)dt \leq E_{total} \tag{23}$$

Among them

$$P_{total}(t) = P_{idle} + P_{compute}(\pi, b_t) + P_{actuation}(a_t) \tag{24}$$

Furthermore, the communication process is limited by bandwidth $BW_{comm}, L_{comm}$, and the information transmission constraint is

$$BW_{comm} \cdot (t_2 - t_1), t_{arrival} = t_{send} + L_{comm} \tag{25}$$

In the scenario of human-machine collaborative tasks, the agent needs to form an effective mental model alignment with human astronauts. Let $M_H(s)$ be the distribution of human state cognition and $b(s)$ be the distribution of agent belief. The requirement is to minimize the bias of the mental model:

$$D_{KL}(M_H(s)||b(s)) \leq \epsilon_{allignment} \tag{26}$$

The communication output $E = \pi_{comm}(b, a)$ generated by the agent needs to have high interpretability and low ambiguity, and the collaborative performance is quantified through the joint task completion time $T_{complete}(H, \pi)$.

# C ESI Settings

**Algorithm 1** ESI (Environmental Scenario Index) Calculation

**Require:** is_correct: Boolean, ▷ Accuracy judge output integrity_val: Integer or NULL, ▷ Integrity judge score, 0-100 if not NULL tokens_used: Integer, ▷ Worker completion tokens answer_text: String, ▷ Worker's cleaned answer is_cot_formatted: Boolean, ▷ CoT format correctness prompt_type: String, ▷ e.g., "COT", "DIRECT" answer_len: Integer, ▷ Length of worker's answer ref_len: Integer, ▷ Length of reference answer CONFIG: { ▷ Configuration parameters object/struct TOKEN_BUDGET, P_IRR, SAFETY_KW_LIST, MAX_LEN_RATIO, $W_{acc}, W_{int}, W_{eff}, W_{safe}, W_{align}$ ▷ ESI weights }

**Ensure:** ESI_score: Float

**procedure** CALCULATEESI(is_correct, integrity_val, tokens_used, answer_text, is_cot_formatted, prompt_type, answer_len, ref_len, CONFIG)
  $S_{acc} \leftarrow 0.0$
  **if** is_correct **then**
    $S_{acc} \leftarrow 100.0$
  **end if**
  $S_{integrity} \leftarrow 0.0$
  **if** integrity_val $\neq$ NULL **and** $0 \leq$ integrity_val $\leq 100$ **then**
    $S_{integrity} \leftarrow$ FLOAT(integrity_val)
  **end if**
  $S_{eff} \leftarrow 0.0$
  **if** tokens_used $\geq 0$ **and** CONFIG.TOKEN_BUDGET $> 0$ **then**
    $S_{budget} \leftarrow \max(0, 1 - \text{tokens\_used}/\text{CONFIG.TOKEN\_BUDGET}) \times 100$
    $S_{eff} \leftarrow \max(0, S_{budget} \times (1 - \text{CONFIG.P\_IRR}))$
  **end if**
  $S_{safety} \leftarrow 100.0$
  **for** each $kw \in$ CONFIG.SAFETY_KW_LIST **do**
    **if** CONTAINS(LOWERCASE(answer_text), $kw$) **then** ▷ Helper functions assumed
      $S_{safety} \leftarrow 0.0$
      **break** ▷ Exit loop
    **end if**
  **end for**
  $score_{align} \leftarrow 100.0$
  **if not** is_correct **then**
    $score_{align} \leftarrow score_{align} - 40$
  **end if**
  **if** prompt_type = "COT" **and not** is_cot_formatted **then**
    $score_{align} \leftarrow score_{align} - 30$
  **end if**
  **if** ref_len $> 0$ **and** answer_len $> 0$ **then**
    $len\_ratio \leftarrow$ FLOAT(answer_len)/ref_len
    **if** $len\_ratio >$ CONFIG.MAX_LEN_RATIO **then**
      $score_{align} \leftarrow score_{align} -$ CONFIG.MAX_LEN_RATIO
    **end if**
  **end if**
  $S_{align} \leftarrow \max(0.0, score_{align})$
  ESI_score $\leftarrow$ CONFIG.$W_{acc} \cdot S_{acc} +$
  CONFIG.$W_{int} \cdot S_{integrity} +$
  CONFIG.$W_{eff} \cdot S_{eff} +$
  CONFIG.$W_{safe} \cdot S_{safety} +$
  CONFIG.$W_{align} \cdot S_{align}$
  **return** ESI_score
**end procedure**

## D Data Sources

This appendix provides public access points and information repositories for some of the major data sources referenced in the construction of the Lunar Bench dataset. This list is illustrative rather than exhaustive, and highlights the general categories and availability of data.

***Note:*** Certain internal or specialized documents may not be accessible via the following public links or may require special permissions for access.

**1. Historical Mission Archives**

- NASA History Division: `https://history.nasa.gov/`
- Apollo Lunar Surface Journal (ALSJ): `https://www.hq.nasa.gov/alsj/`
- National Space Science Data Center (NSSDC): `https://nssdc.gsfc.nasa.gov/`
- Russian Space Web by Anatoly Zak: `http://www.russianspaceweb.com/`
- China National Space Administration (CNSA): `http://www.cnsa.gov.cn/`
- Lunar and Planetary Data Release System: `http://moon.bao.ac.cn/`
- Indian Space Research Organisation (ISRO): `https://www.isro.gov.in/`
- PRADAN – ISRO Science Data Archive: `https://pradan.issdc.gov.in/pradan/`

**2. Modern Mission Planning Resources**

- NASA Artemis Program: `https://www.nasa.gov/artemisprogram`
- NASA Commercial Lunar Payload Services (CLPS): `https://www.nasa.gov/content/commercial-lunar-payload-services/`
- European Space Agency (ESA): `https://www.esa.int/`

**3. Scientific Literature and Preprint Platforms**

- Google Scholar: `https://scholar.google.com/`
- NASA ADS (Astrophysics Data System): `https://ui.adsabs.harvard.edu/`
- arXiv Preprint Server: `https://arxiv.org/`
- z-library: `https://zh.z-library.sk/`

**4. Educational and MOOC Platforms**

- Coursera: `https://www.coursera.org/`
- edX: `https://www.edx.org/`
- NASA STEM Engagement: `https://www.nasa.gov/stem/`
- Smithsonian National Air and Space Museum: `https://airandspace.si.edu/`

**5. Community and Web-Curated Knowledge Sources**

- Wikipedia: `https://en.wikipedia.org/`
- Baidu Baike: `https://baike.baidu.com/`
- Quora: `https://www.quora.com/`
- Reddit: `https://www.reddit.com/`
- YouTube: `https://www.youtube.com/`
- Bilibili: `https://www.bilibili.com/`

## E Definition of Level 1-3

The **Lunar-Bench** framework adopts a structured three-tiered evaluation system—L1, L2, and L3—to systematically assess the progressively advanced reasoning and operational LLMs in lunar exploration settings. These hierarchical levels, illustrated in Figure 9, represent a conceptual continuum from basic understanding to autonomous scientific agency.

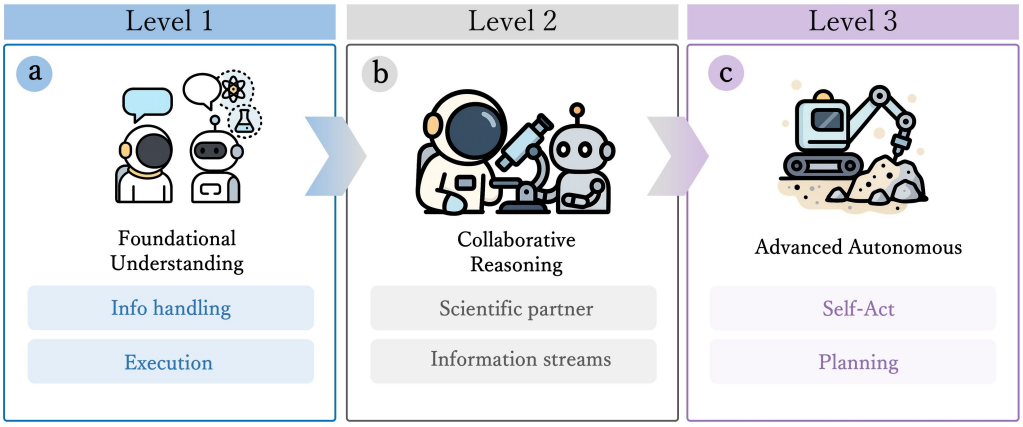


Figure 9: Three-level capability hierarchy in the Lunar-Bench framework: (a) Basic Interaction; (b) Collaborative Research Expertise; (c) Autonomous Scientific Decision-making.

**Foundational Understanding.** This level evaluates the LLM's ability to accurately interpret and execute explicit, single-turn instructions in well-constrained operational contexts. L1 tasks emphasize precise comprehension of domain-specific terminology, straightforward reasoning, and deterministic command execution. The focus is on assessing reliability in basic operational comprehension—such as interpreting a system status report or carrying out a direct control command—requiring minimal inference. This stage corresponds to the "*Basic Interaction*" functionality depicted in Figure 9(a).

**Collaborative Reasoning.** L2 probes the model's competence in assisting with multi-step reasoning, information fusion, and decision support under scientific and engineering workflows. Tasks at this level assess the LLM's capacity to serve as a collaborative agent, synthesizing disparate information sources to produce coherent analyses, explanations, or operational recommendations. This level aligns with the "*Collaborative Research Expertise*" domain shown in Figure 9(b), reflecting a partnership model of interaction with human operators or other autonomous systems.

**Advanced Autonomous Functionality.** L3 represents the highest tier, assessing the LLM's ability to make strategic decisions and adaptively plan actions in complex, dynamic, and partially specified environments. Tasks require long-horizon reasoning, handling of uncertainty, and optimization under constraints, often with minimal human instruction. As shown in Figure 9(c), this "*Autonomous Scientific*" stage simulates scenarios in which the LLM must operate independently and robustly in service of high-level mission objectives, such as autonomous exploration planning, anomaly mitigation, or scientific hypothesis generation.

## F Sample display of Lunar-Bench

**Level-1 Sample (ID: 6) Scenario- Collection**

**Instruction.** The Chang'e-6 mission aims to collect lunar regolith from the South Pole–Aitken Basin. The target region is characterized by medium hardness (Mohs scale 4–5), low viscosity, and a relatively high volatile content ( 2%). Three sampling tools are available: (1) A diamond-coated rotary drill, suitable for materials with hardness $> 6$ and requiring 500–800 N axial force. (2) A titanium alloy grab, suitable for loose soil and requiring 200–300 N clamping force. (3) A scraper equipped with a heating element, designed for volatile-rich materials, operating with 150 N contact pressure and 50°C thermal activation.

**Question.** Considering the given soil properties and tool specifications, which sampling tool provides the optimal trade-off between effectiveness and energy efficiency? Justify the selected force control parameters accordingly.

**Answer.** The **scraper with heating** is the most appropriate tool under the specified conditions.

**Level-2 Sample (ID: 82) Scenario- Collaboration**

**Instruction.** The lunar base energy grid supports three critical devices: the life support system (200 W, priority 1), the mobile rover (up to 500 W, priority 2), and the science lab module (150 W nominal, throttleable to 100 W, priority 3). The solar array currently provides 600 W of power, but an emergency will reduce output to 400 W in 15 minutes, lasting for 2 hours. Additionally, a 200 Wh battery is available, but can only be used to sustain the life support system.
Device operation follows these rules: higher-priority devices must remain powered at all times; devices of equal priority share remaining power equally; and the battery is exclusively reserved for priority-1 operation.

**Question.** When the solar output drops to 400 W, determine: (a) the maximum power (in W) that can be allocated to the science lab module, (b) the actual power received by the mobile rover.

**Answer.** Science lab module: **0 W**; Mobile rover: **200 W**.

**Level-3 Sample (ID: 92)-Communication**

**Instruction.** The multispectral imager onboard the lunar research station produces approximately 20 GB of raw data daily. The AI processing unit supports three compression strategies: (1) *Lossless compression* with a compression ratio of 1.5:1 and a processing time of 30 minutes per GB; (2) *Lossy compression*, scientifically acceptable, with a compression ratio of 8:1 and a processing time of 15 minutes per GB; and (3) *Intelligent screening*, which involves 5 minutes per GB for feature extraction followed by transmission of only 10% of the extracted key data (equivalent to a 10:1 compression ratio).
Currently, the next communication window opens in 4 hours and lasts 30 minutes. The available downlink bandwidth is 50 Mbps, corresponding to a maximum transmission capacity of 11.25 GB. The AI processor can handle up to 2 compression tasks in parallel. The data batch includes 8 GB of high-priority region images (which must be preserved in full fidelity) and 12 GB of routine region images (which allow lossy compression).

**Question.** Design an optimal data processing and transmission schedule that maximizes scientific value under the communication and processing constraints. Specify which compression strategy is applied to each data category and report the expected volume of transmitted data.

**Answer.** Lossless compression for high-priority images; intelligent screening for routine images; total transmission volume: **10.4 GB**.

## G Baseline Models

To comprehensively evaluate the capabilities of contemporary LLMs, our study incorporates a diverse suite of models (Table 6 and Table 7). This selection encompasses state-of-the-art systems alongside other widely adopted or representative architectures, facilitating a broad comparative analysis. Our model collection was curated to address specific research questions concerning the interplay of reasoning versus dialogue capabilities and the impact of model scale.

**Reasoning LLMs.** This category includes models recognized for their strong logical inference and complex problem-solving abilities, crucial for specialized tasks demanding deep reasoning. The selection allows for a focused assessment of performance on tasks.

**General LLMs.** This group consists of mainstream models optimized for interactive dialogue, general question answering, and broader language understanding tasks. These models are typically fine-tuned for natural and coherent human-AI interaction.

Table 6: Reasoning and General LLMs for Evaluation

| **Reasoning LLMs** | **General LLMs** |
|---|---|
| ChatGPT-o4-mini-high [53] | ChatGPT-4o [49] |
| ChatGPT-o3 [53] | ChatGPT-4.5 [51] |
| ChatGPT-o1 [52] | ChatGPT-4.1 [50] |
| Claude 3.7 Sonnet [5] | Claude 3.5 Haiku [4] |
| Claude 3.5 Sonnet [6] | Gemini-2.5-Flash [24] |
| Gemini-2.5-Pro [23] | Deepseek-V3 (0324) [15] |
| Deepseek-R1 [14] | Llama-3.3-70B-Instruct [42] |
| Qwen3-235B-A22B [64] | Gemma-3-27B [25] |
| Llama-4-maverick [43] | Qwen-2.5-72B-Instruct [59] |
| Qwen3-32B [61] | Mistral-small-24B-instruct-2501 [46] |
| QwQ-32B [66] | Llama-3.1-405B-Instruct [40] |
| Deepseek-Prover-v2 [68] | ChatGLM-4-32B [21] |
| ChatGLM-Z1-rumination-32B [83] | Qwen-Max [80] |

**Small-Scale Language Models (SLMs).** To investigate the influence of model scale and assess the viability of LLMs for resource-constrained environments (edge deployment), a selection of smaller yet potent models was included. This allows for a nuanced analysis of performance trade-offs relative to computational footprint.

Table 7: Reasoning and General SLMs for Evaluation

| **General SLMs** | **Reasoning SLMs** |
|---|---|
| ChatGPT-4o-mini [48] | Deepseek-r1-distill-llama-8B [14] |
| Qwen-2.5-7B-Instruct [60] | Gemma-3-1B [20] |
| Llama-3.1-8B-Instruct [44] | ChatGLM-Z1-9B [73] |
| Llama-3.2-3B-Instruct [41] | Qwen3-1.7B [63] |
| Gemma-2-9B [22] | Qwen3-0.6B [62] |
| Ministral-8B [45] | Qwen3-8B [61] |
| Ministral-3B [58] | Qwen3-4B [65] |

## H Evaluations Prompt

**Chain-of-Thought Prompt Template**

```
Your task is to answer the 'Specific Question (Question)' based on the
'Background Information (Instruction)'.
Follow these steps:  1.  First, carefully analyze the Instruction and
the Question.  2.  Provide a step-by-step reasoning process that shows
how you arrive at the answer.  Start this section with 'Reasoning:'.
3.  After your reasoning, on a new line, provide the final, direct, and
concise answer.  This final answer MUST be prefixed with 'Final Answer:
'  (note the space after the colon).
The final answer part should be a single word, a short phrase,
a specific name, a numerical value, a code snippet, or a status
description, derived ONLY from the 'Background Information'.
Do not add any other explanations or text after the 'Final Answer:  '
prefix and the answer itself.
Background Information (Instruction):  {{instruction}}
Specific Question (Question):  {{question}}
```

**Expert Role Prompt Template**

```
Leveraging your expertise in lunar exploration engineering, analyze
the 'Background Information (Instruction)' and 'Specific Question
(Question)' below.  Provide the most direct, concise, and factually
accurate answer based SOLELY on the information presented.  Your
response should be a single word, a short phrase, a specific name, a
numerical value, a code snippet, or a status description that precisely
answers the question.  Do NOT include any explanations, justifications,
prefixes (e.g., 'The answer is:'), suffixes, conversational filler, or
any information not explicitly stated in the 'Background Information'.
Output only the precise answer itself.
Background Information (Instruction):  {{instruction}}
Specific Question (Question):  {{question}}
Answer:
```

**Few-Shot Prompt Template**

```
You are provided with a few input-output examples that illustrate how
a specific type of question should be answered using the 'Background
Information (Instruction)'.  Based solely on these examples and the
given Instruction and Question pair, infer the correct answer in
the same format.  The final answer should be a single word, a short
phrase, a specific name, a numerical value, a code snippet, or a status
description.  Do NOT include explanations, reasoning, or additional
content.  Follow the structure and style of the provided examples
strictly.  Output only the precise answer derived from the Instruction
and examples.
Few-shot Examples:
Example 1:  Instruction:  {{ex1_instruction}}  Question:
{{ex1_question}}  Answer:  {{ex1_answer}}
Example 2:  Instruction:  {{ex2_instruction}}  Question:
{{ex2_question}}  Answer:  {{ex2_answer}}
-- Target Sample Below --
Background Information (Instruction):  {{instruction}}
Specific Question (Question):  {{question}}
Answer:
```

LLM-as-a-judge Prompt

```
You are a discerning AI Evaluator. Your task is to determine if the
'Candidate Answer' is largely consistent with the expected correct
information and factually sound, accurately addressing the 'Question'
based on the 'Instruction'. While not requiring a verbatim match to the
'Reference Answer', a significant overlap in key facts and meaning is
expected. The answer should be substantially correct.
Evaluation Focus: Substantial Factual Alignment, Core Consistency, and
Accuracy
You should judge is_judged_correct: true if:
1. Core Factual Alignment: The primary facts, figures, entities, or
conclusions in the Candidate Answer align with those derivable from the
'Instruction' and supported by the 'Reference Answer'.
2. Addresses Key Aspects of Question Accurately: The Candidate Answer
responds to the main point of the Question with factual correctness.
3. No Major Factual Discrepancies or Logical Flaws: There are no
substantial errors or contradictions. Reasoning, if included, must
be valid.
4. Meaningful Overlap with Correct Information: The meaning and
critical facts are largely the same as in the Reference Answer, despite
possible stylistic or minor detail differences.
You should judge is_judged_correct: false if:
- The Candidate Answer has key factual inaccuracies or conflicts with
the Instruction or Reference Answer.  - It omits essential information
needed to correctly address the Question.  - It is topically related but
fundamentally misrepresents or fails to answer the core intent.  - It
includes logical issues that undermine factual soundness.
Your response must include a JSON object with two fields:  1.
"is_judged_correct": true or false  2. "reasoning": a concise
explanation citing factual alignment or discrepancies
Background Information (Instruction): {{instruction}}
Specific Question (Question): {{question}}
Reference Answer (Reference): {{reference_answer}}
Candidate Answer (Candidate): {{candidate_answer}}
Now provide your evaluation.
```

# I Discussion of the Results

## I.1 Model Performance Across Level 1-3

Our evaluation across Lunar-Bench's hierarchical task tiers reveals distinct performance profiles for contemporary LLMs. At the L1 level, which measures foundational comprehension and instruction adherence, even the most capable models exhibited notable shortcomings relative to human performance. For example, the top-performing model, Gemini-2.5 Pro, achieved an overall L1 accuracy of 47.8%, while the best-performing open-source model, DeepSeek-R1, attained 39.1% (as shown in Table 2). In contrast, average human accuracy at L1 reached 65.1%, with top human performance peaking at 72.1%. Many other models, including those with substantial parameter scales, yielded L1 accuracies primarily within the 15–38% range, underscoring persistent limitations in mastering even the entry-level tasks within this specialized lunar domain.

As task complexity increased at the L2 tier—requiring multi-turn collaborative reasoning and averaging 9.3 discrete inferential steps—a marked decline in model performance was observed. Even the strongest systems reached only around 16.7% accuracy on these intermediate-level tasks. The challenges intensified substantially at the L3 level, which evaluates autonomous decision-making capabilities. These tasks, involving an average of 14.8 reasoning steps and often demanding approximately 20 minutes of deliberation time even for human experts, proved overwhelmingly difficult for existing models. **Average LLMs performance dropped below 10%, with a substantial subset of models scoring near zero.** This steep performance gradient across L1, L2, and L3 tiers delineates the current limitations of LLMs in high-stakes, multi-constraint environments and highlights the pressing need for further advancements to meet the demands of real-world lunar operations.

## I.2 Prompt Strategy Nuances

Our evaluation of prompt strategies on Lunar-Bench (Table 3) revealed limited efficacy in substantially improving performance on its complex lunar reasoning tasks. Standard prompting ("None") established modest baselines (GPT-o1 at 47.2%; Gemini-2.5 Pro at 47.8%), underscoring the benchmark's inherent difficulty. Notably, Chain-of-Thought (CoT) provided almost no discernible benefit (GPT-o1: 47.0%; DeepSeek-R1: 38.8%), often stagnating or slightly degrading performance. This may be because for models already possessing strong intrinsic reasoning capabilities, a generic CoT prompt alone is insufficient to unlock further significant gains on these highly specialized, multi-constraint tasks without deeper, task-specific alignment; CoT was also observed to sometimes impair instruction adherence. In contrast, "Expert Role" prompting yielded consistent, albeit minor, positive uplifts (Gemini-2.5 Pro to 50.0%), likely by better focusing the model's context.

The hybrid "CoT+Expert" strategy resulted in unstable and unpredictable outcomes, with marginal and model-dependent improvements (Gemini-2.5 Pro to 50.3%) or even slight degradations compared to "Expert Role" alone (Qwen-Max: 43.5%), highlighting challenges in beneficially compounding prompt complexities. **Overall, these findings demonstrate that while minor optimizations are possible, current advanced prompting techniques offer only marginal and inconsistent advantages on Lunar-Bench.** This suggests that achieving robust performance in such demanding, high-stakes domains necessitates more fundamental advancements in model architectures, training, or fine-tuning, as opposed to relying solely on prompt engineering.

## I.3 Analysis of Few-shot Results

Table 4 shows that while providing a small number of task examples (1 to 2 shots) can, for most models, elicit an initial improvement over zero-shot performance, the gains are generally marginal. For instance, GPT-o1, representing one of the highest-performing models, improves its accuracy from 47.2% (0-shot) to a peak of 50.7% (2-shot), a modest gain of 3.5 percentage points. Similarly, Gemini-2.5 Pro sees its performance increase from 47.8% to 50.3% with a single example. The leading open-source model, DeepSeek-R1, also benefits from in-context learning, reaching its peak accuracy of 43.2% at 2-shot, up from 39.1% at 0-shot. These limited gains highlight that while models can leverage contextual examples to some extent, the inherent difficulty of the Lunar-Bench tasks, which require deep domain-specific reasoning and adherence to complex constraints, is not substantially overcome by simple few-shot prompting.

Crucially, the results demonstrate a clear pattern of diminishing returns, and in several cases, performance degradation with an increasing number of examples, particularly for the more capable models. As seen with GPT-o1 and Gemini-2.5 Pro, performance dips after reaching their peak at 1 or 2 shots (GPT-o1 drops to 49.3% at 3-shot; Gemini-2.5 Pro to 48.5% at 3-shot). This phenomenon is also observed for Claude-3.7 Sonnet (peaking at 45.2% with 1-shot before declining) and Qwen-Max (peaking at 44.5% with 1-shot). This aligns with observations that for state-of-the-art models possessing strong instruction-following and inherent zero-shot generalization capabilities, a larger number of examples can lead to performance saturation or even degradation. This may be attributed to the introduction of noise, overly specific constraints from the provided examples that may not generalize well, or the increased context length subtly interfering with the models' core reasoning pathways.

For models with more constrained initial capabilities, such as QWQ-32B (a 32B parameter model), the benefit from few-shot examples is even more restricted, with scores hovering around 30-32% and showing minimal positive deviation with added examples. This underscores the substantial deficiencies these models exhibit in complex lunar reasoning, where few-shot prompting alone offers little remedy.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See *Sec.* 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See *Sec.* 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See *App.* B
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See *App.* D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total compute time and the resource type used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See *Sec.* 5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See *Sec.* 6
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See *Sec.* 6