A Pilot Benchmark for NL-to-FOL Translation in Planetary Exploration

Anonymous Authors*

***Abstract*— Future planetary exploration envisions autonomous robotic agents operating under severe communication constraints, without global positioning, and with minimal human intervention. In such environments, agents must not only perceive and act, but also reason over mission objectives, operational constraints, and evolving environmental conditions. While prior work has largely focused on perception and control, the translation of high-level mission knowledge into structured, machine-interpretable representations remains underexplored.**

We introduce a pilot benchmark for translating natural language (NL) into First-Order Logic (FOL) within the domain of planetary exploration. The dataset is constructed from real mission documentation sourced from NASA's Planetary Data System (PDS), spanning missions from 2003 to 2013. These documents describe mission phases such as launch, boost, coast, cruise, and orbital operations in rich natural language. We manually annotate these documents with corresponding FOL representations that capture temporal structure, agent roles, and operational dependencies. In addition, we provide structured predicate vocabularies and typed constants to enable controlled experimentation with varying levels of prior knowledge. This pilot benchmark provides a foundation for research at the intersection of language understanding and formal reasoning, grounded in real-world, safety-critical mission data. The dataset is provided for anonymous review at: `https://anonymous.4open.science/r/PMR-BF96/mission.json`.

I. INTRODUCTION

Autonomous robotic systems are expected to play a central role in the next generation of planetary exploration [1]. Unlike terrestrial systems, these agents must operate under extreme constraints, including delayed or intermittent communication with Earth [2], limited onboard computational resources, and the absence of global positioning systems. In such conditions, autonomy is not merely a convenience but a necessity. Agents must interpret mission objectives, reason about evolving environmental conditions, and make decisions that are both safe and effective without continuous human oversight.

A key challenge in achieving this level of autonomy lies in how mission knowledge is represented and utilized. In current practice, mission plans, operational procedures, and system constraints are largely expressed in natural language. These descriptions contain rich information about temporal sequences, dependencies between subsystems, and conditional behaviors. However, natural language lacks the precision and structure required for reliable execution and formal reasoning. As a result, there is a gap between how mission knowledge is specified and how it can be consumed by autonomous systems.

Bridging this gap requires transforming high-level natural language descriptions into structured, machine-interpretable representations. First-Order Logic (FOL) provides a natural candidate for this transformation, offering a formal framework for representing entities, relations, and temporal dependencies. Despite its relevance, the problem of translating natural language mission descriptions into FOL has received limited attention, particularly in domains characterized by long-horizon temporal reasoning and complex operational structure [3], [4], [5], [6], [7].

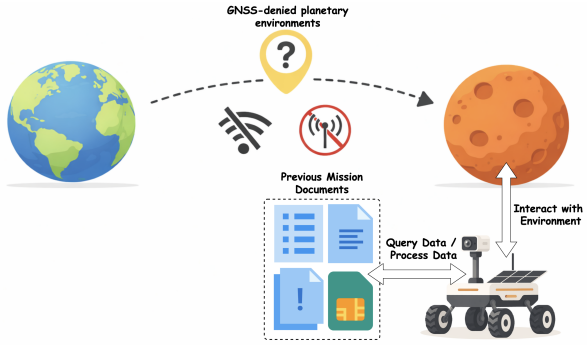

Fig. 1. Conceptual setting motivating this work. In long-duration, communication-constrained planetary exploration, an autonomous agent cannot rely on continuous connectivity to Earth and must instead use onboard mission documents and internal reasoning to interpret mission phases, understand constraints, and make decisions while interacting with its environment.

In this work, we introduce a pilot benchmark designed to support research on this problem. Our goal is to provide a dataset that captures the richness and complexity of real planetary mission documentation while enabling systematic exploration of natural language to logical representation. Rather than focusing on evaluating specific models, we position this benchmark as a foundational resource for studying the interaction between language understanding and formal reasoning in safety-critical planetary exploration environments. This complements prior work in autoformalization and structured translation tasks [5], [6], [7].

II. DATASET CONSTRUCTION

The dataset is constructed from publicly available mission documentation provided by NASA's Planetary Data System (PDS) [8]. These documents are written for scientific and operational purposes and contain detailed descriptions of mission phases, system behaviors, and observational activities. As such, they provide a realistic and challenging source of natural language data for structured reasoning tasks. The full dataset is available for anonymous review at: `https://anonymous.4open.science/r/PMR-BF96/mission.json`.

We collect mission documents spanning the period from

2003 to 2013, covering a diverse set of planetary missions and operational contexts. Each document describes the progression of a mission through multiple phases, including launch, boost, coast, cruise, orbit insertion, and science operations. These phases vary in duration from minutes to years and often involve nested temporal relationships and interdependent activities. The resulting dataset captures both short-duration procedural events and long-horizon mission planning, providing a broad range of reasoning challenges.

Each document is manually translated into a corresponding First-Order Logic representation. This annotation process is designed to preserve the semantic content of the original text while making its structure explicit. Temporal relationships are encoded using predicates that capture ordering, duration, and dependency, allowing events to be composed into coherent logical sequences. Agent roles are explicitly represented, distinguishing between spacecraft, subsystems, instruments, and external entities such as ground stations. Operational activities, such as maneuvers, observations, and system checks, are modeled as predicates that relate agents, actions, and outcomes.

An example from the dataset illustrates the level of detail captured in these annotations. The following excerpt shows a natural language description of a launch sequence alongside its corresponding First-Order Logic representation:

Natural Language:

> Spirit (MER-2 hardware) was launched June 10, 2003, at 1759 UTC (1359 EDT) from launch complex 17A at Cape Canaveral Air Force Station, Florida. The launch azimuth was 93 degrees. The boost portion of the launch vehicle trajectory took approximately 10 minutes, and was followed by a short coast phase in a parking orbit for approximately 15 minutes. After third stage burnout, the upper stage despun the stack using a yo-yo despin system. Separation of the third stage occurred approximately 36 minutes after launch.

First-Order Logic:

$$
\begin{aligned}
&\exists e \ (\text{LaunchEvent}(e) \wedge \text{Agent}(e, \text{Spirit\_MER2}) \wedge \\
&\text{DateUTC}(e, '\text{2003-06-10}') \wedge \\
&\text{TimeUTC}(e, '\text{17:59}') \wedge \\
&\text{LocalTime}(e, '\text{13:59}') \wedge \text{LocalTZ}(e, 'EDT') \wedge \\
&\text{From}(e, \text{LaunchComplex17A}) \wedge \\
&\text{At}(e, \text{CapeCanaveralAFS}) \wedge \text{In}(e, \text{Florida}) \wedge \\
&\text{AzimuthDeg}(e, 93)) \\
&\wedge \exists b \ (\text{BoostPhase}(b) \wedge \text{Agent}(b, \text{Spirit\_MER2}) \wedge \\
&\text{StartsAfterMinutesApprox}(b, e, 0) \wedge \\
&\text{DurationMinutesApprox}(b, 10)) \\
&\wedge \exists c \ (\text{CoastPhase}(c) \wedge \text{Agent}(c, \text{Spirit\_MER2}) \wedge \\
&\text{InOrbit}(c, \text{ParkingOrbit}) \wedge \\
&\text{StartsAfterMinutesApprox}(c, b, 0) \wedge \\
&\text{DurationMinutesApprox}(c, 15) \wedge \text{After}(c, b)) \\
&\wedge \exists d \ (\text{DespinEvent}(d) \wedge \text{Agent}(d, \text{Spirit\_MER2}) \wedge \\
&\text{After}(d, \text{ThirdStageBurnout}) \wedge \\
&\text{Uses}(d, \text{YoYoDespinSystem}) \wedge \\
&\text{ActsOn}(d, \text{UpperStageStack})) \\
&\wedge \exists s \ (\text{ThirdStageSeparation}(s) \wedge \\
&\text{Agent}(s, \text{Spirit\_MER2}) \wedge \\
&\text{AfterMinutesApprox}(s, e, 36))
\end{aligned}
$$

Beyond the paired natural language and FOL representations, the dataset includes additional structure that supports controlled experimentation. Predicate vocabularies are extracted and standardized across all documents, providing a consistent set of logical primitives. Constants are separated and typed according to their semantic roles, including entities, temporal values, and numerical quantities. This separation allows the dataset to be used in settings with varying degrees of prior knowledge, such as scenarios where the logical vocabulary is fixed or where type information is available to guide parsing.

III. BENCHMARK DESIGN

The pilot benchmark is defined as the task of translating natural language mission descriptions into First-Order Logic representations. This task requires capturing not only the meaning of individual statements but also the compositional structure that arises from temporal dependencies and operational constraints. Unlike traditional semantic parsing tasks, which often focus on short and relatively independent utterances, this benchmark involves longer documents with interconnected events and hierarchical structure.

A central feature of the benchmark is its modular design. By explicitly separating predicate vocabularies and typed constants from the core annotations, the dataset enables researchers to explore different problem formulations. In one setting, models may be required to generate fully unconstrained logical forms directly from natural language. In another, models may operate within a predefined schema, leveraging known predicates and type constraints to guide the translation process. These variations allow for systematic investigation of how structure and prior knowledge influence performance.

The benchmark is intentionally agnostic to specific evaluation protocols. Instead, it is designed to support a range of research directions, including symbolic methods that rely on rule-based parsing, neural approaches that learn mappings from data, and hybrid methods that combine both paradigms [9], [10]. By focusing on the dataset itself, we aim to provide a flexible foundation that can be adapted to different experimental goals.

IV. MOTIVATION AND APPLICATION

The motivation for this pilot benchmark is rooted in the operational requirements of future autonomous planetary systems. In scenarios where communication delays prevent real-time human intervention, agents must rely on internal representations of mission knowledge to make decisions. These representations must be both expressive enough to capture complex behaviors and structured enough to support reliable reasoning.

First-Order Logic offers a formalism that satisfies these requirements. In the context of planetary rover autonomy, this capability enables several concrete functions. Structured logical representations derived from natural language mission descriptions can be used to define constraints for semantic mapping, inform task planning and scheduling, enforce safety conditions during navigation, and support verification and validation (V&V) of autonomous decisions. By grounding high-level mission intent in formal representations, autonomous systems can more reliably interpret objectives, reason over constraints, and ensure that executed actions remain consistent with mission requirements.

The benchmark also has implications beyond planetary exploration. Many domains require the translation of natural language into structured representations, including robotics, aerospace systems, and industrial automation. In these settings, the ability to bridge language and logic can enable more interpretable and reliable systems. By grounding this problem in real-world mission data, the benchmark provides a realistic testbed for developing and evaluating such capabilities [11], [12].

V. BENCHMARK STATISTICS

The pilot benchmark currently consists of 11 multi-phase mission segments in natural language (NL) and their corresponding First-Order Logic (FOL). These examples are derived from 5 unique planetary mission sources spanning NASA Planetary Data System (PDS) documentation from 2003 to 2013. Across the dataset, we identify 251 unique predicate signatures and 125 distinct entity constants, along with 16 date literals, 13 time literals, and 51 numeric values. This reflects both the semantic diversity of mission descriptions and the structured variability required for formal reasoning.

The natural language inputs are moderately long and descriptive, with an average length of 257.6 words (median 249), ranging from 84 to 412 words. The corresponding FOL representations are similarly expressive, averaging 286.9 tokens (median 288), with lengths ranging from 174 to 434 tokens. This near parity in NL and FOL length highlights that the translation task preserves substantial structural detail rather than compressing information into shallow forms.

At the logical level, each example contains an average of 43.4 predicate mentions and 13.0 quantifiers, indicating a high degree of compositional structure. The number of unique predicates per example averages 30.3, suggesting that each mission segment introduces a rich and varied set of relations. Entity constants average 12.2 per example, while numeric constants average 4.6, reflecting the prevalence of temporal durations, orbital parameters, and measurement values in mission descriptions.

The dataset also exhibits significant variability in complexity. Natural language inputs range from short procedural descriptions (84 words) to long, multi-phase mission narratives (over 400 words). Correspondingly, FOL representations vary in both depth and breadth, with predicate vocabularies ranging from 17 to 50 per example and predicate mentions ranging from 27 to 65. Quantifier counts range from 5 to 25, further emphasizing differences in logical nesting and compositional depth across mission segments.

An analysis of predicate usage reveals consistent structural patterns across examples. The most frequent predicates include `During` (42 occurrences), `Agent` (26), `DateUTC` (13), and temporal relation predicates such as `StartsAt`, `EndsAt`, and `After`. These distributions highlight the central role of temporal reasoning and agent-centric event modeling in the dataset.

Overall, these statistics indicate that the benchmark captures both linguistic and logical complexity, combining long-form natural language descriptions with deeply structured formal representations. The diversity in predicate vocabularies, temporal constructs, and entity types makes this dataset well-suited for studying compositional generalization and structured reasoning in realistic, safety-critical and planetary exploration domains.

Statistic	Mean	Range
NL Length (words)	257.6	84–412
FOL Length (tokens)	286.9	174–434
Predicate Mentions	43.4	27–65
Unique Predicates / Example	30.3	17–50
Quantifiers	13.0	5–25
Entity Constants	12.2	6–18

TABLE I

SUMMARY STATISTICS OF THE NL–FOL PILOT BENCHMARK.

VI. PRELIMINARY RESULTS

To provide an initial characterization of the pilot benchmark, we conducted exploratory experiments using a set of publicly available local language models, including `Qwen/Qwen1.5-1.8B-Chat` [13], `mistralai/Mistral-7B-Instruct-v0.3` [14], `microsoft/Phi-3-mini-4k-instruct` [15], `meta-llama/Meta-Llama-3.1-8B-Instruct` [16], and `google/gemma-2-9b-it` [17]. For each model, we provided the full natural language mission description as input and prompted the model to directly generate a corresponding First-Order Logic (FOL) representation without additional explanation or intermediate steps. All model outputs were fixed to a maximum of 500 tokens.

This setup represents a straightforward, unconstrained translation task and is not intended to reflect optimized prompting or system design. Rather, it serves as a baseline to illustrate the challenges posed by long-form, structured mission descriptions.

Across all models, we observe consistent failure modes. First, models struggle to maintain temporal and logical consistency over long sequences, often producing incomplete or incoherent event structures. Temporal relationships such as ordering, duration, and dependency are frequently omitted, misrepresented, or contradicted within the generated outputs. Second, models tend to miss critical components embedded within the natural language descriptions, particularly when

these components appear in later portions of the text or are nested within complex sentences. Third, hallucination is prevalent, with models introducing predicates, entities, or relationships that are not grounded in the input description. These issues are amplified by the length and compositional structure of the task, which requires sustained reasoning across multiple interconnected events.

Some of the models demonstrate improved fluency and partial structure, but still fail to reliably capture the full set of temporal dependencies and operational constraints present in the source text. Notably, even when local consistency is maintained within short segments, global coherence across the full mission sequence remains a challenge.

Model-specific behaviors further highlight these limitations. Qwen [13] frequently fails to produce valid FOL representations and instead hallucinates by outputting natural language instructions or step-by-step procedural descriptions. Mistral [14] and Meta-LLaMA-3.1 [16] demonstrate stronger alignment with the task and are more likely to attempt FOL-like structures, but still tend to generate ordered lists of components that require additional processing to convert into valid logical forms. Interestingly, while these enumerated outputs are incorrect with respect to the target format, they often preserve temporal sequencing (e.g., step-wise progression), which partially maintains logical ordering despite not being explicitly requested.

Microsoft Phi-3 [15] exhibits severe hallucination behavior, frequently introducing unrelated space or NASA concepts not present in the input (e.g., references to the Hubble Space Telescope). In addition, it often repeats hallucinated temporal statements such as "Launch-40 min" and generates redundant natural language outputs describing tasks rather than producing FOL. These repetitions further degrade output quality and consistency.

In addition, Google Gemma [17] occasionally failed to produce any output when given the full mission descriptions, returning either empty responses or terminating generation prematurely. This behavior was observed more frequently for longer inputs and suggests limitations in handling extended context windows or internal decoding constraints under long-form generation.

Among the evaluated models, Meta-LLaMA-3.1 show the strongest ability to produce structured FOL-like outputs in some cases. However, even in these instances, the generated representations deviate significantly from the ground truth in both syntax and structure.

These preliminary results highlight the difficulty of direct, single-pass translation from long-form natural language into structured logical representations. They suggest that more effective approaches will likely require decomposition strategies, such as breaking mission descriptions into smaller units, as well as structured prompting techniques. Potential directions include few-shot prompting with schema constraints, guided decoding with formal grammars, and hybrid neuro-symbolic pipelines that enforce logical consistency during generation [18], [19], [20].

We emphasize that these experiments are not intended as a comprehensive evaluation, but rather as an initial demonstration of the challenges inherent in this pilot benchmark. All raw model outputs are available at the anonymized link: `https://anonymous.4open.science/r/PMR-BF96/output.txt`.

VII. LIMITATIONS

While this work introduces a curated pilot benchmark for NL-to-FOL translation in planetary exploration, several limitations should be noted.

First, the dataset is small in scale, consisting of 11 mission segments derived from 5 unique sources. This work should therefore be viewed as a pilot benchmark intended to motivate further data collection and expansion. Second, the annotations are manually constructed, which introduces potential subjectivity in how natural language descriptions are mapped to logical representations. Third, the evaluation presented in this work is preliminary and qualitative in nature. Future work should explore quantitative evaluation methods, including structural similarity, logical equivalence, and execution-based validation. Finally, the experimental setup relies on direct, single-pass generation with a fixed output length constraint. This setting does not reflect more advanced prompting strategies, decomposition methods, or constrained decoding approaches that may significantly improve performance.

Despite these limitations, we believe this benchmark provides a useful starting point for studying the intersection of natural language understanding and formal reasoning in long-horizon, structured environments.

VIII. CONCLUSION

We present a pilot benchmark for translating natural language mission descriptions into First-Order Logic, constructed from real-world planetary mission documentation. The dataset captures the complexity of mission phases, temporal dependencies, and operational constraints, and augments these representations with structured predicate vocabularies and typed constants. By focusing on the problem of translating high-level mission knowledge into formal representations, this work highlights an important direction for enabling interpretable and verifiable autonomy. We hope this pilot benchmark will serve as a foundation for future research at the intersection of language understanding and formal reasoning in complex, real-world environments.

REFERENCES

[1] S. Chien et al., "The EO-1 Autonomous Science Agent," in Proc. Int. Joint Conf. Autonomous Agents and Multiagent Systems (AAMAS), 2004.

[2] K. Fall, "A delay-tolerant network architecture for challenged internets," in Proc. ACM SIGCOMM, 2003.

[3] L. S. Zettlemoyer and M. Collins, "Learning to map sentences to logical form," in Proc. UAI, 2005.

[4] A. Kamath and R. Das, "A survey on semantic parsing," arXiv preprint arXiv:1812.00978, 2018.

[5] H. Moore and A. Shah, "Evaluating Autoformalization Robustness via Semantically Similar Paraphrasing," in AAAI 2026 Bridge LM-Reasoning Workshop.

[6] M. Safarzadeh, A. Oroojlooyjadid, and D. Roth, "Evaluating NL2SQL via SQL2NL," EMNLP, 2025.

[7] S. Han, X. Guo, J. Chen, and X. Sun, "FOLIO: Natural Language Reasoning with First-Order Logic," in Proc. EMNLP, 2024.
[8] NASA Planetary Data System (PDS), "Planetary Data System," 2024. [Online]. Available: https://pds.nasa.gov/
[9] P. Yin and G. Neubig, "A syntactic neural model for general-purpose code generation," in Proc. ACL, 2017.
[10] A. d'Avila Garcez et al., "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," arXiv, 2019.
[11] P. Rajpurkar et al., "SQuAD: 100,000+ questions for machine comprehension of text," in Proc. EMNLP, 2016.
[12] S. W. Squyres et al., "The Opportunity Rover's Athena science investigation at Meridiani Planum, Mars," Science, 2004.
[13] Qwen Team, "Qwen1.5: Open foundation models," 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen1.5-1.8B-Chat
[14] A. Mistral AI, "Mistral-7B-Instruct-v0.3," 2024. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[15] Microsoft, "Phi-3: Technical report," 2024. [Online]. Available: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
[16] Meta AI, "LLaMA 3.1: Open and efficient foundation models," 2024. [Online]. Available: https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
[17] Google DeepMind, "Gemma: Open models based on Gemini research," 2024. [Online]. Available: https://huggingface.co/google/gemma-2-9b-it
[18] J. Wei et al., "Chain-of-Thought prompting elicits reasoning in large language models," in Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS), 2022.
[19] N. F. Liu et al., "Lost in the middle: How language models use long contexts," Trans. ACL, 2024.
[20] Y. Bang et al., "A multitask, multilingual, multimodal evaluation of ChatGPT," in Proceedings of the 13th International Joint Conference on Natural Language (IJCNLP), 2023