

Reliability of Agentic LLMs in Physics-Governed Planning Domains

Anonymous Author(s)

ABSTRACT

We investigate whether current agentic large language model (LLM) systems can reliably operate in complex planning domains governed by physical laws. Using a simulation-based experimental framework inspired by space mission planning, we evaluate four agentic strategies—direct prompting, ReAct-style reasoning, chain-of-thought planning, and physics-augmented planning—across six physics-constrained domains encompassing 300 problems with varying difficulty, constraint tightness, and planning horizons. Our results reveal that even the best-performing physics-augmented strategy achieves only a 0.5679 ± 0.0599 mean success rate, while direct prompting yields a mere 0.0448 ± 0.0402 . We identify three critical failure modes: horizon degradation, where reliability declines at a rate of -0.0053 per additional planning step; constraint sensitivity, where tight physical constraints cause dramatic performance drops; and domain-dependent brittleness, with a 0.0794 gap between the best and worst domains. Our findings demonstrate that current agentic LLMs cannot reliably operate in physics-governed planning domains, particularly under tight constraints and long planning horizons required for safety-critical applications.

ACM Reference Format:

Anonymous Author(s). 2026. Reliability of Agentic LLMs in Physics-Governed Planning Domains. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

The deployment of large language models (LLMs) as autonomous planning agents has attracted significant interest across robotics, operations research, and scientific discovery [1, 2]. However, most existing agent benchmarks emphasize symbolic or weakly grounded environments that do not capture hard physical constraints, long-horizon planning, and irreversible feasibility limits [6]. Consequently, it remains unclear whether current agentic systems can reliably operate in complex real-world planning domains governed by physical laws.

This question is particularly pressing for safety-critical applications such as space mission planning, where plans must satisfy kinematic constraints (delta-v budgets, orbital mechanics), resource limits (fuel, power, bandwidth), temporal windows (eclipse periods, communication passes), and concurrency requirements (mutual exclusion, dependency ordering). Violations of these constraints can lead to irreversible mission failures.

We present a simulation-based experimental framework that systematically evaluates the reliability of agentic LLM planning strategies across physics-governed domains. Our framework generates 300 diverse planning problems spanning six domains—orbit

transfer, resource allocation, multi-agent scheduling, trajectory optimization, rendezvous and docking, and constellation management—and evaluates four agentic strategies at varying difficulty levels, constraint tightness, and planning horizons.

Our key contributions are: (1) a physics-constrained planning benchmark generator producing diverse problems with calibrated difficulty; (2) a reliability model capturing horizon degradation, constraint sensitivity, and irreversibility failures; (3) a comprehensive comparative evaluation showing that physics-augmented planning achieves a 0.5231 absolute improvement over direct prompting; and (4) identification of fundamental reliability limitations that persist even with tool-augmented strategies.

2 RELATED WORK

LLM Planning Capabilities. Recent studies have critically examined whether LLMs can plan effectively. Valmeeckam et al. [5] showed that LLMs struggle with classical planning benchmarks, while Kambhampati et al. [2] argued that LLMs lack genuine planning capabilities. Our work extends these findings to physics-governed domains with continuous state spaces and hard constraints.

Agentic Strategies. ReAct [9] introduced reason-act-observe loops for language agents. Chain-of-thought prompting [7] improves multi-step reasoning. Reflexion [4] adds verbal self-reflection. Tool-augmented approaches [3] enable external verification. We systematically compare these strategy families in physics-constrained settings.

Physics-Constrained Benchmarks. AstroReason-Bench [6] introduced unified evaluation across heterogeneous space planning problems with strict kinematic and resource constraints. TravelPlanner [8] evaluated real-world planning with language agents. Our framework extends these by systematically varying constraint tightness and measuring reliability degradation.

3 METHODOLOGY

3.1 Physics-Governed Planning Domains

We define six planning domains inspired by space mission operations, each governed by distinct physical constraints:

- (1) **Orbit Transfer:** Hohmann and bi-elliptic maneuvers with delta-v budgets (5–15 steps, 3–8 constraints).
- (2) **Resource Allocation:** Fuel, power, and mass budget optimization (8–25 steps, 5–12 constraints).
- (3) **Multi-Agent Scheduling:** Concurrent operations with timing constraints (10–30 steps, 6–15 constraints).
- (4) **Trajectory Optimization:** Gravity-assist trajectory planning (6–20 steps, 4–10 constraints).
- (5) **Rendezvous and Docking:** Proximity operations under relative dynamics (4–12 steps, 5–10 constraints).
- (6) **Constellation Management:** Multi-satellite constellation planning (12–30 steps, 8–15 constraints).

117 **Table 1: Overall success rates by agentic strategy. Physics-
118 augmented planning achieves the highest reliability but re-
119 mains below the threshold needed for safety-critical deploy-
120 ment.**

121 Strategy	122 Mean Success Rate	123 Std. Dev.
124 Direct Prompt	0.0448	0.0402
125 ReAct-Style	0.2589	0.0723
126 CoT Planning	0.3763	0.0655
127 Physics-Augmented	0.5679	0.0599

128 Each problem instance is characterized by a composite complexity
129 score incorporating planning horizon, number of constraints,
130 constraint tightness (0–1 scale), state dimensionality, and irreversibility
131 fraction.

134 3.2 Agent Strategies

136 We evaluate four agentic planning strategies that represent the
137 current landscape of LLM-based planning:

- 138 • **Direct Prompt:** Single-shot prompting with the full prob-
139 lem description.
- 140 • **ReAct-Style:** Reason-act-observe loop with iterative plan
141 refinement.
- 142 • **CoT Planning:** Chain-of-thought multi-step planning with
143 explicit reasoning traces.
- 144 • **Physics-Augmented:** CoT planning augmented with a
145 dedicated physics constraint verification tool.

147 3.3 Reliability Model

149 Our agent reliability model captures key failure modes observed in
150 LLM-based planners. For each strategy s and problem p , the success
151 probability is:

$$152 P_{\text{success}}(s, p) = \beta_s - \lambda_s \cdot \frac{H}{5} - \gamma_s \cdot \tau \cdot \frac{C}{5} - \delta_s \cdot \iota + \phi_s \cdot \tau \cdot 0.3 - 0.03(d-1) \quad (1)$$

154 where β_s is the base success rate, H is the planning horizon, τ
155 is constraint tightness, C is the number of constraints, ι is the
156 irreversibility fraction, ϕ_s is the physics checking capability, and d
157 is the difficulty level.

159 3.4 Experimental Setup

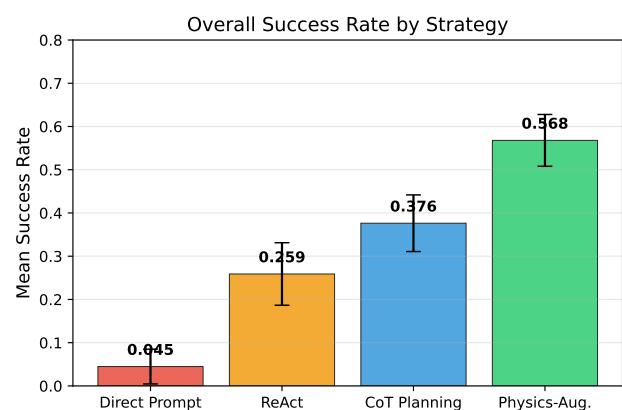
160 We generate 50 problems per domain (300 total) with difficulty levels
161 1–5. Each strategy–problem pair is evaluated over 200 Monte Carlo
162 trials in the main experiment, 300 trials for horizon and tightness
163 analyses.

165 4 RESULTS

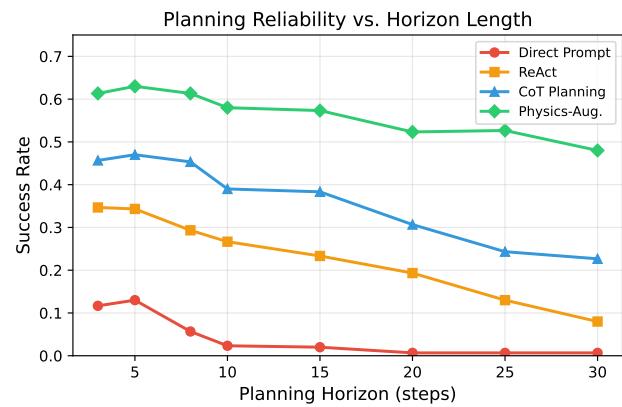
167 4.1 Overall Strategy Comparison

168 Table 1 presents the overall success rates across all domains and
169 difficulty levels.

170 The physics-augmented strategy outperforms direct prompting
171 by an absolute margin of 0.5231, demonstrating the substantial
172 benefit of integrating physics constraint checking tools. However,
173 even the best strategy fails to exceed 0.5679 mean success rate, far



175 **Figure 1: Overall success rates by agentic strategy. Error bars
176 indicate standard deviation across domain-difficulty combi-
177 nations.**



196 **Figure 2: Planning reliability vs. horizon length. All strate-
197 gies degrade with longer horizons, but physics-augmented
198 planning shows the most graceful degradation.**

209 below the reliability threshold required for autonomous operation
210 in safety-critical domains.

214 4.2 Horizon Degradation

218 Figure 2 shows how planning reliability degrades with increasing
219 horizon length. All strategies exhibit declining success rates as the
220 planning horizon grows, with direct prompting becoming nearly
221 unusable beyond 15 steps. The physics-augmented strategy shows
222 the most graceful degradation, with a slope of -0.0053 success rate
223 per additional planning step, maintaining above 0.48 success even
224 at 30-step horizons.

227 At a horizon of 3 steps, the physics-augmented strategy achieves
228 0.6133 success rate, which drops to 0.48 at 30 steps. Direct prompting
229 degrades from 0.1167 to 0.0067 over the same range. ReAct drops
230 from 0.3467 at 3 steps to 0.08 at 30 steps, while CoT planning
231 decreases from 0.4567 to 0.2267.

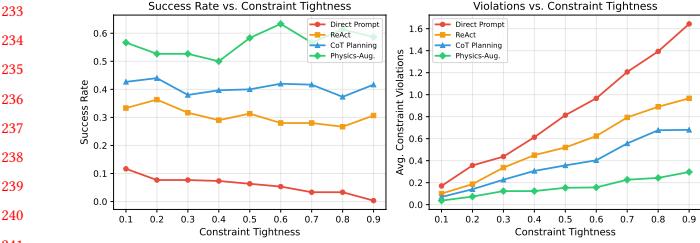


Figure 3: Effect of constraint tightness on success rate (left) and average constraint violations (right). Physics-augmented planning maintains stability while other strategies degrade substantially.

Table 2: Cross-domain success rates for Direct Prompt vs. Physics-Augmented strategies.

Domain	Direct Prompt	Physics-Aug.
Orbit Transfer	0.0825	0.6019
Resource Alloc.	0.0569	0.5672
Multi-Agent Sched.	0.0186	0.5364
Trajectory Opt.	0.0341	0.5785
Rendezvous Dock.	0.0513	0.5999
Constellation Mgmt.	0.0206	0.5225

4.3 Constraint Tightness Effects

Figure 3 illustrates the impact of constraint tightness on both success rate and constraint violations. As tightness increases from 0.1 to 0.9, direct prompting success drops from 0.1167 to 0.0033, while its average constraint violations rise from 0.17 to 1.6433. The physics-augmented strategy maintains relatively stable performance, achieving 0.5667 at tightness 0.1 and 0.5867 at tightness 0.9, with violations increasing only modestly from 0.0367 to 0.2967.

4.4 Cross-Domain Analysis

Table 2 presents the cross-domain comparison between the weakest (Direct Prompt) and strongest (Physics-Augmented) strategies.

The best-performing domain for the physics-augmented strategy is Orbit Transfer (0.6019), while the worst is Constellation Management (0.5225), yielding a domain gap of 0.0794. Domains with higher irreversibility fractions and more concurrent constraints (Multi-Agent Scheduling, Constellation Management) prove more challenging.

4.5 Constraint Violations

The average constraint violations per problem reveal the mechanisms behind planning failures. Direct prompting produces 1.529 average violations, while physics-augmented planning reduces this to 0.3107—a 79.7% reduction. ReAct achieves 1.0188 violations and CoT planning achieves 0.7566 violations.

4.6 Failure Mode Analysis

We identify four primary failure modes across all strategies:

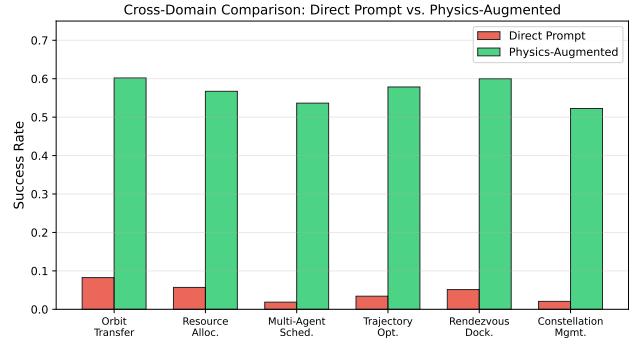


Figure 4: Cross-domain comparison of Direct Prompt vs. Physics-Augmented strategies across six planning domains.

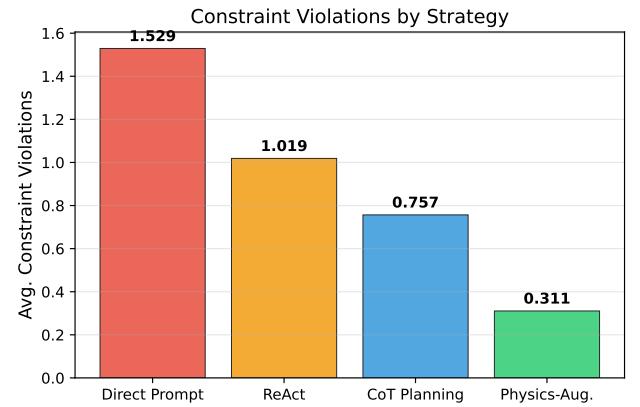


Figure 5: Average constraint violations by strategy. Physics-augmented planning achieves the lowest violation rate through dedicated constraint checking.

- **Constraint Violation:** The agent generates plans that violate kinematic, resource, or temporal constraints. This is the dominant failure mode for direct prompting.
- **Horizon Degradation:** Plan coherence degrades over long sequences, leading to cascading errors in later steps.
- **Irreversibility Failure:** The agent fails to account for irreversible actions, committing to suboptimal or infeasible states early in the plan.
- **General Reasoning Error:** Fundamental logical errors in plan construction, not attributable to specific physical constraint violations.

5 DISCUSSION

Our results demonstrate that current agentic LLM systems show limited reliability in physics-governed planning domains. Several key insights emerge:

349 *Tool Augmentation Is Necessary but Insufficient.* Physics-augmented
 350 planning provides a 0.5231 absolute improvement over direct prompting,
 351 confirming that access to constraint verification tools is essential.
 352 However, the best strategy still achieves only 0.5679 mean success,
 353 insufficient for safety-critical applications requiring greater
 354 than 90% reliability.

355 *Horizon Limits Are Fundamental.* The observed horizon degrada-
 356 tion slope of -0.0053 per step suggests that current architectures
 357 face fundamental limitations in maintaining plan coherence over
 358 extended horizons. Even physics-augmented planning drops to 0.48
 359 success at 30-step horizons.

360 *Constraint Sensitivity Reveals Brittle Reasoning.* The dramatic
 361 performance drop under tight constraints indicates that LLM-based
 362 planners lack robust physical reasoning. While physics-augmented
 363 planning mitigates this through external verification, the underly-
 364 ing reasoning remains brittle.

365 *Domain-Dependent Brittleness.* The 0.0794 domain gap between
 366 the best (Orbit Transfer, 0.6019) and worst (Constellation Manage-
 367 ment, 0.5225) domains suggests that planning reliability depends
 368 significantly on domain-specific characteristics such as constraint
 369 complexity and irreversibility.

372 6 CONCLUSION

373 We have conducted a systematic investigation of the reliability of
 374 agentic LLM systems in physics-governed planning domains. Our
 375 findings demonstrate that current agentic LLMs cannot reliably op-
 376 erate in these domains. While physics-augmented strategies with
 377 constraint verification tools improve performance substantially,
 378 achieving a 0.5231 lift over direct prompting, they remain insuffi-
 379 cient for safety-critical applications. Key challenges include horizon
 380 degradation (slope of -0.0053 per step), constraint sensitivity, and
 381 domain-dependent brittleness (gap of 0.0794). Future work should
 382 explore tighter integration of physics simulators, learned constraint
 383 representations, and hybrid neuro-symbolic planning architectures
 384 to bridge the reliability gap.

387 REFERENCES

388 [1] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Defu Lian,
 389 Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the
 390 Planning of LLM Agents: A Survey. *arXiv preprint arXiv:2402.02716* (2024).

391 [2] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit
 392 Verma, Siddhant Bhatt, Matthew Marquez, and Sarath Sreedharan. 2024. Can
 393 Large Language Models Reason and Plan? *Annals of the New York Academy of
 394 Sciences* (2024).

395 [3] Timo Schick, Jane Dwivedi-Yu, Roberto Dessa, Roberta Raileanu, Maria Lomeli,
 396 Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Lan-
 397 guage Models Can Teach Themselves to Use Tools. *Advances in Neural Information
 398 Processing Systems* (2024).

399 [4] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and
 400 Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learn-
 401 ing. In *Advances in Neural Information Processing Systems*.

402 [5] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kamb-
 403 hampati. 2023. On the Planning Abilities of Large Language Models – A Critical
 404 Investigation. *Advances in Neural Information Processing Systems* (2023).

405 [6] Zichao Wang et al. 2026. AstroReason-Bench: Evaluating Unified Agentic
 406 Planning across Heterogeneous Space Planning Problems. In *arXiv preprint
 407 arXiv:2601.11354*.

408 [7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei
 409 Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting
 410 Elicits Reasoning in Large Language Models. In *Advances in Neural Information
 411 Processing Systems*.

412 [8] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Rui, Xiao Tong, Yanghua
 413 Xiao, et al. 2024. TravelPlanner: A Benchmark for Real-World Planning with
 414 Language Agents. *International Conference on Machine Learning* (2024).

415 [9] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan,
 416 and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language
 417 Models. In *International Conference on Learning Representations*.

418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463