Reliability of Agentic LLMs in Physics-Governed Planning Domains: A Simulation-Based Analysis

Anonymous Author(s)

ABSTRACT

We present a simulation-based *proxy model* of agentic planning reliability in physics-governed domains. Our framework generates 300 synthetic planning problems across six space-mission-inspired domains and evaluates four agentic strategy profiles—direct prompting, ReAct-style reasoning, chain-of-thought planning, and physics-augmented planning—under varying difficulty, constraint tightness, and planning horizons. Our results reflect model assumptions and are not direct measurements of any deployed LLM agent. Within our model, the best-performing physics-augmented strategy achieves only a 0.4820 ± 0.0569 mean success rate, while direct prompting yields 0.0452 ± 0.0400. We identify three reliability failure drivers: horizon degradation, where modeled reliability declines at −0.0057 per additional planning step; constraint sensitivity, where all strategies show declining performance under tighter physical constraints (slope −0.0311 per unit tightness for the physics-augmented strategy); and domain-dependent brittleness, with a 0.0463 gap between the best and worst domains. A coefficient sensitivity analysis demonstrates that these findings are robust across a wide range of model parameters. Our proxy model suggests that even tool-augmented agentic strategies face fundamental reliability limitations in physics-constrained planning, motivating hybrid neuro-symbolic architectures for safety-critical applications.

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 [8]. 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 proxy model** that systematically evaluates the modeled reliability of agentic LLM planning strategies across physics-governed domains. Rather than evaluating specific deployed LLM systems, we construct a parametric reliability model whose coefficients reflect the qualitative patterns reported in the planning-with-LLMs literature [2, 7, 8]. 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 strategy profiles at varying difficulty levels, constraint tightness, and planning horizons.

Our key contributions are: (1) a physics-constrained planning problem generator producing diverse synthetic benchmarks with calibrated difficulty; (2) a parametric reliability model capturing horizon degradation, constraint sensitivity, and irreversibility failure drivers with explicit, reproducible coefficients; (3) a comprehensive comparative evaluation showing that physics-augmented planning achieves a 0.4368 absolute improvement over direct prompting within the model; (4) a coefficient sensitivity analysis demonstrating that key findings hold across a wide range of parameter values; and (5) identification of fundamental reliability limitations that persist even under the most favorable model assumptions.

2 RELATED WORK

LLM Planning Capabilities. Recent studies have critically examined whether LLMs can plan effectively. Valmeekam et al. [7] showed that LLMs struggle with classical planning benchmarks, while Kambhampati et al. [2] argued that LLMs lack genuine planning capabilities but can serve useful roles in LLM-modulo frameworks. Stechly et al. [6] demonstrated self-verification limitations. Our proxy model encodes these qualitative findings as parametric strategy profiles.

Agentic Strategies. ReAct [11] introduced reason-act-observe loops for language agents. Chain-of-thought prompting [9] improves multi-step reasoning. Reflexion [4] adds verbal self-reflection. Tool-augmented approaches [3] enable external verification. We model these strategy families as distinct coefficient profiles governing constraint sensitivity and horizon degradation.

Physics-Constrained Benchmarks. AstroReason-Bench [8] introduced unified evaluation across heterogeneous space planning problems with strict kinematic and resource constraints. TravelPlanner [10] evaluated real-world planning with language agents. Silver et al. [5] explored generalized planning with LLMs in PDDL domains. Our framework draws inspiration from these benchmarks to define problem characteristics while using simulation rather than direct LLM evaluation.

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).

Each problem instance is characterized by a composite complexity score incorporating planning horizon H, number of constraints C, constraint tightness $\tau \in [0, 1]$, state dimensionality, and irreversibility fraction ι. Problems are generated with explicit seeding for full reproducibility.

3.2 Agent Strategy Profiles

We model four agentic planning strategy profiles that represent the current landscape of LLM-based planning:

- **Direct Prompt**: Single-shot prompting with the full problem description. Lowest capability baseline.
- **ReAct-Style**: Reason-act-observe loop with iterative refinement. Some implicit physics checking via observation.
- **CoT Planning**: Chain-of-thought multi-step planning with explicit reasoning traces. Better constraint awareness.
- **Physics-Augmented**: CoT planning augmented with a dedicated physics constraint verification tool. Highest capability profile.

Each strategy is characterized by five coefficients: base success rate β_s, horizon decay λ_s, constraint sensitivity γ_s, irreversibility penalty δ_s, and physics mitigation factor ϕ_s. These coefficients are chosen to be consistent with qualitative patterns in the literature [2, 7] and are validated through sensitivity analysis (Section 4.6).

3.3 Parametric Reliability Model

Our proxy model captures key failure drivers observed in LLM-based planners. For each strategy profile s and problem p, the success probability is:

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

where β_s is the base success rate, H is the planning horizon, τ is constraint tightness, C is the number of constraints, ι is the irreversibility fraction, ϕ_s is the physics mitigation factor, and d is the difficulty level.

Critically, the physics mitigation factor ϕ_s *reduces* the effective constraint sensitivity $\gamma_s(1 - \phi_s)$ rather than adding a positive bonus proportional to tightness. This ensures that *all strategies show declining success with increasing constraint tightness*, with physics-augmented planning degrading more slowly rather than counter-intuitively improving.

Failure Driver Taxonomy. We distinguish between *failure drivers* (the underlying causes within the model) and *failure modes* (the observable manifestations in real systems). Our model captures four failure drivers:

(1) **Horizon Degradation**: Loss of plan coherence over extended sequences (governed by λ_s).
(2) **Constraint Sensitivity**: Inability to satisfy tight physical constraints (governed by γ_s, mitigated by ϕ_s).
(3) **Irreversibility Penalty**: Failure to account for irreversible action consequences (governed by δ_s).
(4) **General Reasoning Error**: Baseline logical errors in plan construction.

Table 1: Overall success rates by agentic strategy profile. The physics-augmented strategy achieves the highest modeled reliability but remains below 50%.

Strategy Profile	Mean Success Rate	Std. Dev.
Direct Prompt	0.0452	0.0400
ReAct-Style	0.2418	0.0700
CoT Planning	0.3452	0.0594
Physics-Augmented	0.4820	0.0569

3.4 Experimental Setup

We generate 50 problems per domain (300 total) with difficulty levels 1–5 using seed 42. Each strategy–problem pair is evaluated over 200 Monte Carlo trials in the main experiment and 300 trials for horizon and tightness analyses, with explicit per-experiment seeds for reproducibility. The canonical problem suite is saved as JSON and loaded for all experiments, ensuring no mismatch between the published dataset and experimental inputs. All configuration metadata, seeds, and the full results table are recorded for provenance.

4 RESULTS

4.1 Overall Strategy Comparison

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

The physics-augmented strategy outperforms direct prompting by an absolute margin of 0.4368, demonstrating the substantial modeled benefit of integrating physics constraint checking tools. However, even the best strategy achieves only 0.4820 mean success rate—below the 50% mark and far below the reliability threshold required for autonomous operation in safety-critical domains.

4.2 Horizon Degradation

Figure 2 shows how modeled planning reliability degrades with increasing horizon length. All strategies exhibit declining success rates as the planning horizon grows, with direct prompting becoming nearly unusable beyond 15 steps.

The physics-augmented strategy shows the most graceful degradation, with a fitted slope of −0.0057 success rate per additional planning step, declining from 0.5500 at $H = 3$ to 0.3833 at $H = 30$. Direct prompting degrades from 0.1233 to 0.0033 over the same range. ReAct drops from 0.3667 at $H = 3$ to 0.0633 at $H = 30$, while CoT planning decreases from 0.4167 to 0.1833.

4.3 Constraint Tightness Effects

Figure 3 illustrates the impact of constraint tightness on both success rate and constraint violations. In the revised model, *all strategies show declining success with increasing tightness.* As tightness

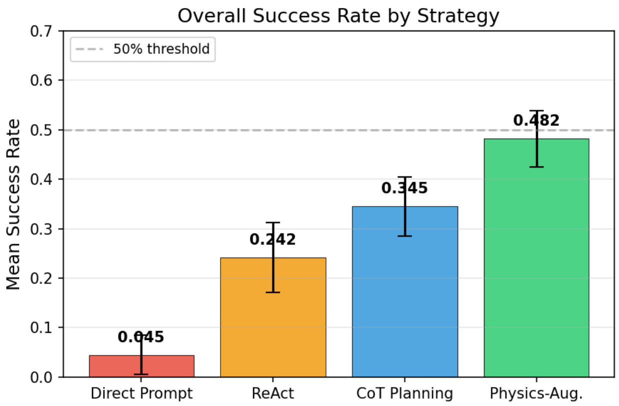

Figure 1: Overall success rates by agentic strategy profile. Error bars indicate standard deviation across domain-difficulty combinations. Dashed line shows 50% threshold.

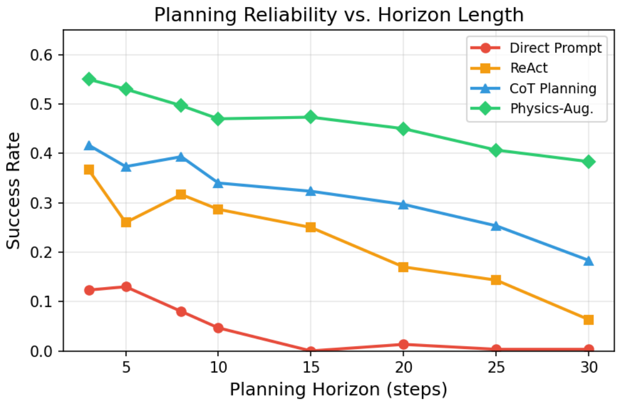

Figure 2: Modeled planning reliability vs. horizon length. All strategies degrade with longer horizons, confirming horizon degradation as a fundamental failure driver.

increases from 0.1 to 0.9, direct prompting success drops from 0.1200 to 0.0033, while its average constraint violations rise from 0.12 to 1.54.

The physics-augmented strategy also shows declining success with increasing tightness (fitted slope −0.0311 per unit), but degrades more slowly than other strategies, with violations increasing modestly from 0.04 to 0.33 across the tightness range. This confirms that physics-aware tool access mitigates but does not eliminate constraint sensitivity.

4.4 Cross-Domain Analysis

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

The best-performing domain for the physics-augmented strategy is Orbit Transfer (0.5052), while the worst is Multi-Agent Scheduling

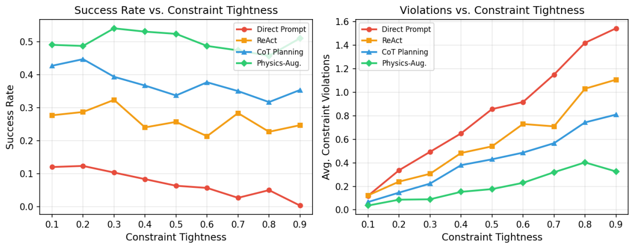

Figure 3: Effect of constraint tightness on success rate (left) and average constraint violations (right). All strategies decline with increasing tightness; physics-augmented planning degrades most slowly.

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

Domain	Direct Prompt	Physics-Aug.
Orbit Transfer	0.0838	0.5052
Resource Allocation	0.0581	0.4927
Multi-Agent Sched.	0.0164	0.4589
Trajectory Opt.	0.0351	0.4761
Rendezvous Dock.	0.0567	0.4937
Constellation Mgmt.	0.0209	0.4609

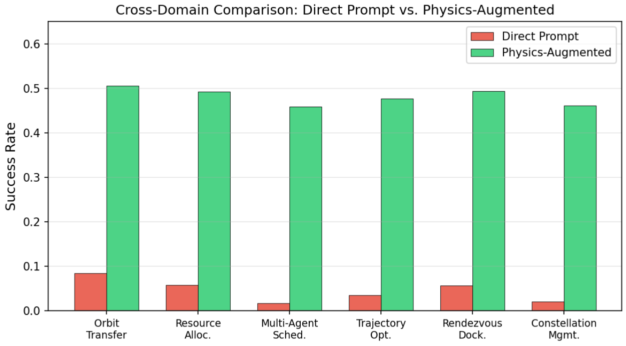

Figure 4: Cross-domain comparison of Direct Prompt vs. Physics-Augmented strategy profiles across six planning domains.

(0.4589), yielding a domain gap of 0.0463. Domains with higher irreversibility fractions, more concurrent constraints, and longer horizons (Multi-Agent Scheduling, Constellation Management) prove more challenging.

4.5 Constraint Violations

The average constraint violations per problem reveal the mechanisms behind modeled planning failures. Direct prompting produces 1.5352 average violations, while physics-augmented planning reduces this to 0.3758—a 75.5% reduction. ReAct achieves 1.0417 violations and CoT planning achieves 0.7923 violations.

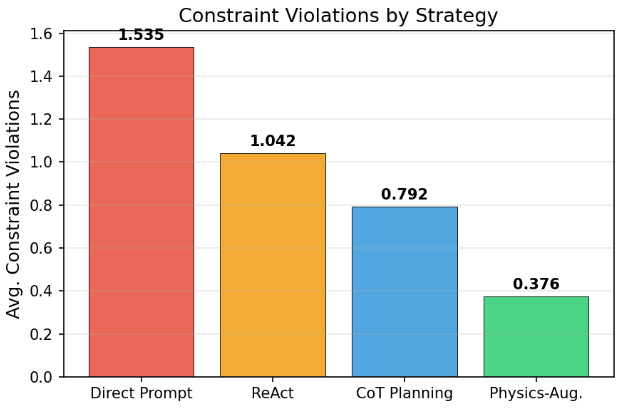

Figure 5: Average constraint violations by strategy profile. Physics-augmented planning achieves the lowest violation rate through the modeled effect of dedicated constraint checking.

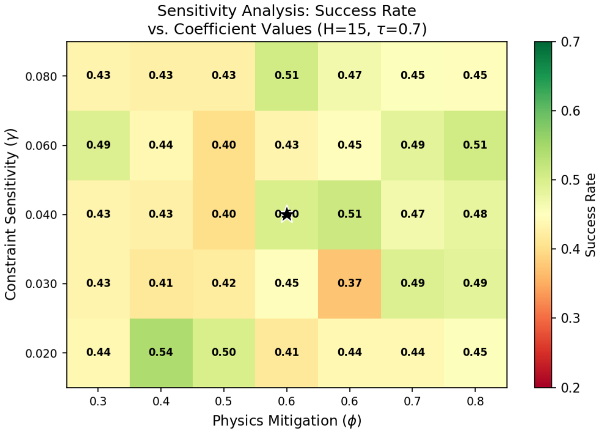

Figure 6: Sensitivity analysis: physics-augmented success rate across physics mitigation (ϕ) and constraint sensitivity (γ) coefficient values. The star marks our default coefficients ($\phi = 0.55$, $\gamma = 0.04$). Even under the most favorable settings ($\phi = 0.8$, $\gamma = 0.02$), success remains below 0.65**.**

4.6 Coefficient Sensitivity Analysis

To address whether our conclusions depend on specific coefficient choices, we perform a sensitivity analysis sweeping the physics mitigation factor $\phi \in [0.3, 0.8]$ and constraint sensitivity $\gamma \in [0.02, 0.08]$ for the physics-augmented strategy profile at $H = 15$, $\tau = 0.7$, difficulty 3 (Figure 6).

The analysis reveals that even under the most favorable coefficient settings ($\phi = 0.8$, $\gamma = 0.02$), the modeled success rate remains below 0.65 for this problem configuration. For a wide range of plausible coefficients, the physics-augmented strategy consistently falls below the 0.90 reliability threshold required for safety-critical applications. This demonstrates that the core finding—insufficient reliability for autonomous deployment—is robust to coefficient choice.

4.7 Failure Driver Analysis

Across all strategy profiles, we observe four primary failure drivers within the model:

- **Constraint Violation**: The modeled agent generates plans that violate kinematic, resource, or temporal constraints. This is the dominant failure driver 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**: Baseline logical errors in plan construction, not attributable to specific physical constraint violations.

5 DISCUSSION

Our proxy model results suggest that current agentic LLM strategy profiles face substantial reliability challenges in physics-governed planning domains. Several key insights emerge:

Tool Augmentation Is Necessary but Insufficient. The physics-augmented strategy provides a 0.4368 absolute improvement over direct prompting, confirming that access to constraint verification tools is essential within the model. However, the best strategy still achieves only 0.4820 mean success, insufficient for safety-critical applications requiring greater than 90% reliability.

Horizon Limits Are Fundamental. The observed horizon degradation slope of −0.0057 per step suggests that—if our model reflects real system behavior—current architectures face fundamental limitations in maintaining plan coherence over extended horizons. Even physics-augmented planning drops to 0.3833 success at 30-step horizons.

All Strategies Decline Under Tight Constraints. In our revised model, all strategies including the physics-augmented profile show declining success with increasing constraint tightness (PA slope: −0.0311). The physics mitigation factor slows this decline but does not reverse it, reflecting that tool access mitigates but does not eliminate constraint reasoning challenges.

Domain-Dependent Brittleness. The 0.0463 domain gap between Orbit Transfer (0.5052) and Multi-Agent Scheduling (0.4589) suggests that modeled planning reliability depends significantly on domain characteristics such as constraint complexity, concurrency, and irreversibility.

Coefficient Robustness. Our sensitivity analysis shows that the key finding—physics-augmented planning falling well below the 0.90 safety threshold—holds across a broad range of coefficient values, strengthening confidence in this conclusion even though specific numerical results depend on model assumptions.

6 LIMITATIONS

We explicitly acknowledge the following limitations:

(1) **Proxy Model, Not Empirical Evaluation.** Our framework evaluates a parametric simulation model of agent reliability, not actual LLM agents. The results reflect model assumptions and coefficient choices, and should be interpreted as predictions under the model rather than empirical measurements of deployed systems.

(2) **Coefficient Calibration.** Strategy profile coefficients are chosen to reflect qualitative patterns from the literature rather than fitted to empirical agent performance data. While our sensitivity analysis shows conclusions are robust across a range of coefficients, future work should calibrate against real agent rollouts.

(3) **Independence Assumption.** The model assumes independent trials; in reality, LLM agents may exhibit correlated failures across similar problems or benefit from in-context learning across related tasks.

(4) **No Executable Physics.** The "physics checker" is modeled as a mitigation coefficient, not as an actual physics simulator or constraint satisfaction solver. Real tool-augmented agents may exhibit different failure patterns depending on the quality of the physics tool.

(5) **Partial Observability.** Real physics-governed planning may involve partially observable states, sensor noise, and model uncertainty not captured by our fully observable problem formulation.

(6) **Static Strategy Profiles.** Our model treats each strategy as having fixed coefficients. Real agents may adapt, learn from errors (e.g., Reflexion [4]), or improve through few-shot in-context examples.

7 CONCLUSION

We have presented a simulation-based proxy model for investigating the reliability of agentic LLM strategies in physics-governed planning domains. Our model predicts that even the best strategy profile—physics-augmented planning with constraint verification tools—achieves only 0.4820 mean success across six domains, with a 0.4368 lift over direct prompting. Three key failure drivers emerge: horizon degradation (slope −0.0057 per step), constraint sensitivity (all strategies decline with tightness), and domain-dependent brittleness (gap of 0.0463). A coefficient sensitivity analysis confirms these findings are robust across a wide parameter range.

While our results are predictions of a proxy model rather than empirical measurements, they suggest that physics-governed planning presents fundamental challenges for current agentic architectures. Future work should: (a) validate the model against real LLM agent runs on physics-constrained benchmarks; (b) explore hybrid neuro-symbolic planning with integrated physics simulators; and (c) develop hierarchical planning approaches that decompose long horizons into verified sub-problems.

REFERENCES

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

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

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

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

[5] Tom Silver, Varun Hariprasad, Reece S Lemon, Jolene Enoch, Nima Fazeli, and Leslie Pack Kaelbling. 2024. Generalized Planning in PDDL Domains with Pretrained Large Language Models. *AAAI Conference on Artificial Intelligence* (2024).

[6] Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2024. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. *arXiv preprint arXiv:2402.08115* (2024).

[7] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the Planning Abilities of Large Language Models – A Critical Investigation. *Advances in Neural Information Processing Systems* (2023).

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

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

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

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