Physics-Grounded Benchmark for AI-Driven Orbital Dynamics

Aaron Gyles[1]

Supervisor:
Kristen Menou[1]

[1] *Univeristy of Toronto, Scarborough*

December 1, 2024

Abstract

This paper introduces a benchmark designed to evaluate the performance of Large Language Models (LLMs) in the field of orbital dynamics. The benchmark employs a pipeline of data processing functions to compare of AI-generated solutions for orbital dynamic problems to reference human-verified solutions. The evaluation is quantized by a Root Mean Square Error (RMSE) score obtained from processed time-series data from AI and reference solutions. To demonstrate its use, the benchmark is applied to evaluate the performance of two well-established LLMs- Google's Gemini Flash and Pro - on a set of eight undergraduate-level designed orbital dynamic problems. The results demonstrated that while the Pro model performs significantly better than the Flash model in most problems, both models struggled with problems involving more complex setups and close interactions. These findings suggest that AI models like Gemini hold the potential to solve complex astrophysical problems, but are not yet capable of providing the required precision for research-grade results.

1 Introduction

The rapid advancements in Artificial Intelligence (AI), particularly in Large Language Models (LLMs) such as Google's Gemini, have significantly expanded their capabilities across a wide range of disciplines. From natural language processing to mathematical and scientific problem solving, these models have shown an intriguing potential in handling complex tasks. In the domain of physics, LLMs demonstrate the potential for advancing research in areas such as orbital dynamics, where accurate simulations are crucial for both theoretical studies and real-time analyses.

Orbital dynamic simulations play an essential role in the study of planetary systems, and theoretical models of our universe, as well as a preliminary step in many spacecraft missions. Incorporating the use of AI to generate reliable, physics-grounded simulations could revolutionize this field, saving valuable time for researchers with the potential of enhancing accuracy. Unfortunately, evaluating the quality of AI-generated solutions remains a challenge. Traditional assessment methods, such as multiple-choice evaluation or subjective analyses, can be inconclusive and insufficient to capture the complexities involved in its performance on physics problem-solving. This inadequacy necessitates the development of a more robust and quantitative "ground-truth" evaluation framework.

This paper introduces a benchmark specifically designed to address this need by evaluating LLMs' problem-solving skills in orbital dynamics. The benchmark employs a pipeline of data processing functions for comparisons of AI-generated solutions of orbital dynamic problems to reference human-verified solutions. The evaluation of the AI-generated solution is quantified by a Root Mean Square Error (RMSE) score calculated from processed time-series data of AI and reference solution for the orbital dynamic problem. To demonstrate its utility, the bench-

mark is applied to assess the performance of two well-established LLMs - Google's Gemini Flash and Pro - on a set of eight undergraduate-level designed orbital dynamic problems.

By focusing on a quantitative evaluation of processed time-series data, this study highlights the strengths and limitations of LLMs in solving complex physics problems. Beyond this demonstration, this benchmark represents a broader potential as a tool for assessing AI's role in scientific research as well as providing an objective standard for evaluating their solutions in contrast to existing subjective methods.

1.1 Motivation and Literature Review

The integration of AI into scientific research has transformed how discoveries are made, offering tools to accelerate and enhance the research process (Wang et al. 2023). AI, particularly LLMs, provides capabilities such as generating hypotheses, designing experiments, and interpreting vast datasets, enabling researchers to explore areas in science that were previously inaccessible. Recent advancements, including AI-driven simulations of molecular systems and solutions to long-standing challenges like the 50-year-old protein-folding problem, highlight the potential of AI to revolutionize science across various fields. Current studies are investigating the partial automation of the research process using LLMs, which includes tasks such as experimental design, simulation execution, and generating comprehensive reports (Liu et al. 2024). Some models even propose implementing AI systems capable of fully automating the research process and potentially contributing to scientific discoveries (Lu et al. 2024). According to a survey by Van Noorden and Perkel, many researchers anticipate that AI will play an increasingly crucial role in future research endeavours (Van Noorden and Perkel 2023).

The potential of AI extends beyond accelerating the research process; it can revolutionize simulations, a cornerstone of modern scientific experimentation. While simulations are undeniably invaluable in modern science, they often require extensive effort in coding and fine-tuning parameters to imitate real-world scenarios. This necessity introduces an unfortunate trade-off between accuracy and runtime, a challenge that researchers frequently encounter. These challenges are particularly evident in fields like orbital dynamics, where N-body simulations require a deep understanding of optimization and advanced time-integration techniques. Fortunately, the emergence of AI provides the opportunity to reduce these trade-offs by identifying new or optimized approaches to various problems.

However, the rapid adoption of AI in research necessitates robust methods to evaluate its performance. Current benchmarks, such as those assessing LLMs in machine learning (ML) tasks, often overlook their application in scientific contexts (Huang et al. 2024). While academic benchmarks exist for models like OpenAI's ChatGPT, studies revealed limitations in solving human-level difficulty problems as well as overlooking many of its capabilities, emphasizing the need for additional research and specialized evaluation frameworks (Laskar et al. 2023). In contrast to these evaluations, this study introduces a physics-grounded benchmark specifically designed to evaluate LLMs in solving complex orbital dynamics problems.

Orbital dynamics represents a field where accurate simulations are essential, yet it remains underexplored in AI-focused studies. This paper distinguishes itself from prior research, which has examined the application of AI in high-level computational physics across various domains (Ali-Dib and Menou 2024), by specifically targeting undergraduate-level problems in orbital dynamics. This study evaluates the performance of Gemini's models, Flash and Pro, to determine their effectiveness in tackling complex physics challenges at the undergraduate level. By quantifying its performance with a Root Mean Square Error (RMSE) metric, this work provides a ground truth for AI evaluation, moving beyond subjective assessments. In doing so, this study not only fills a critical gap in the benchmarking of LLMs but also highlights their potential to serve as reliable AI assistants in advancing physics research.

2 Methodologies

2.1 Designed Problems and Reference Solutions

To determine the benchmark of Gemini's Flash and Pro models, it was required to design various orbital dynamics problems at the undergraduate level. This was facilitated in Python using the latest version of the Python package `REBOUND`. The package `REBOUND` was specifically chosen so that Gemini's models would have plenty of pre-training data available online. The orbital dynamics problems are designed to have set initial conditions so that the solution can be reproduced. The problems are shown in Table 1.

Once problems were designed, they were then solved to provide a human-verified reference solution for data comparison. The orbital parameters selected for the study were based on their relevance to the problem and their ability to give clear comparisons between datasets. The solutions were specifically designed to return a dictionary type storing orbital element data, along with a corresponding time array. These outputted values were stored in Comma Separated Values (CSV) files to later combine them into scaled time-series data for comparative analysis.

To obtain the solution from the Gemini Model, the problems were carefully transformed into clear and informative prompts. The prompts were written to clearly state the coding language and integrating package (`REBOUND`) that should be used in the generated solution. Additionally, the initial conditions were explicitly defined, along with an example of the expected data output to ensure compatibility with the data processing pipeline. Preliminary testing was conducted on Gemini's Flash model to improve the prompt before initiating data collection. It is worth noting that the final prompt was made sure to only provide the necessary information required to solve the problem as well as the necessary data format. During preliminary testing, none of the outputted responses were saved by Gemini's Flash model to ensure an initial "first-run" evaluation. The prompts used are shown in Appendix A.

2.2 Data Pipeline and Benchmark

A critical step in evaluating the performance of Gemini's solutions involves processing the simulation data to ensure a meaningful and accurate comparison with the reference solutions. To achieve this, a comprehensive data processing pipeline was developed, which standardizes, aligns, and evaluates the orbital data and time-series outputs from both solutions.

Data Frame Creation, Standardization and Normalization

The pipeline begins by taking the orbital element data and the corresponding time arrays from both the Gemini-generated solution and the reference solution as input. These data are combined into two separate `pandas.DataFrames` types, one for each dataset. The time arrays are added as an explicit column to ensure precise alignment for future steps.

To create a unit-less comparison of orbital elements, the data are standardized using z-score normalization (centering the mean about zero, and unit variance) provided by `scikit-learn`. This ensures that differences in data sets are clear and are not diminished by the magnitude of the data involved. The time arrays are also normalized using Min-Max scaling. Normalization ensures that the time axes of both datasets are comparable, particularly when their simulation time spans differ due to time step selection. To maintain consistency, the dataset with the larger time range serves as the reference for scaling.

Interpolation for Resolution Matching

Given that Gemini's solution was generated with different densities of time points, integrator used, and time steps, the output resolution often differed significantly between the AI-generated and reference datasets. To address this, the pipeline performs interpolation to align the resolutions of the two datasets.

Using the `pandas.Dataframe.merge_asof()` function, the smaller dataset is merged with the larger one, ensuring alignment based on the nearest time points. Polynomial interpolation (order 5) is then applied to fill the gaps and

Table 1: *Summary of Orbital Dynamics Problems*

Problem	Description
1	Simulate a planetary system with a central star, a terrestrial planet, an inclined planet with 20° inclination, and a gas giant. Track the semi-major axes and inclinations of each planet.
2	Simulate a planetary system with a central star, a terrestrial planet, an inclined planet with 75° inclination, and a gas giant. Track the semi-major axes and inclinations of each planet.
3	Simulate a double-planet system in a 1:2 orbital resonance. Track the semi-major axes, orbital periods, and eccentricities of the planets.
4	Simulate a triple-planet system in a 1:2:4 orbital resonance. Track the semi-major axes, orbital periods, and eccentricities of the planets.
5	Simulate a binary star system with a stable planet orbiting both stars. Track the semi-major axes and eccentricities of the stars and the planet.
6	Simulate a binary star system with an unstable planet orbiting both stars. Track the semi-major axes and eccentricities of the stars and the planet, observing the planet's instability.
7	Simulate a planetary system with three planets and a rogue planet that makes a close encounter with the innermost terrestrial planet. Track the semi-major axes and eccentricities of all bodies.
8	Simulate a planetary system with three planets and a rogue planet that makes a close encounter with the second terrestrial planet. Track the semi-major axes and eccentricities of all bodies.

provide a smooth approximation of orbital elements. Polynomial interpolation and order were chosen to balance capturing the smooth variations in the orbital elements with avoiding overfitting or introducing further oscillations. Additionally, depending on which dataset has the higher resolution, the pipeline adapts to either interpolate the AI solution to the reference time points or vice versa.

Root Mean Square Error (RMSE) Calculation

Once the data sets are aligned and interpolated, the pipeline calculates the RMSE for the orbital elements. This metric quantifies the differences between the AI-generated and reference datasets. The RMSE is computed for each orbital element individually, excluding the time column. Additionally, a total RMSE is calculated by taking the roof of the mean of individual RMSEs across all orbital elements to give an overall score.

The final outputs are: 1. Two fully processed and aligned DataFrames representing the Gemini solution and the reference solution and 2. A set of RMSE scores for each orbital element and an aggregate RMSE score. The entire data processing structure is illustrated in Figure 1.

2.3 Determining Benchmark for Gemini's Models

To benchmark Gemini's Pro and Flash models, the data processing pipeline is applied to both of their solutions. The RMSE scores for each orbital element involved in the problem are then compared between the two models to evaluate their performance on the specific orbital dynamics problem. Once all problems were evaluated in both models, the total RMSE score for each problem is then compared to give an overall benchmark of Gemini's model performance in solving orbital dynamics problems.

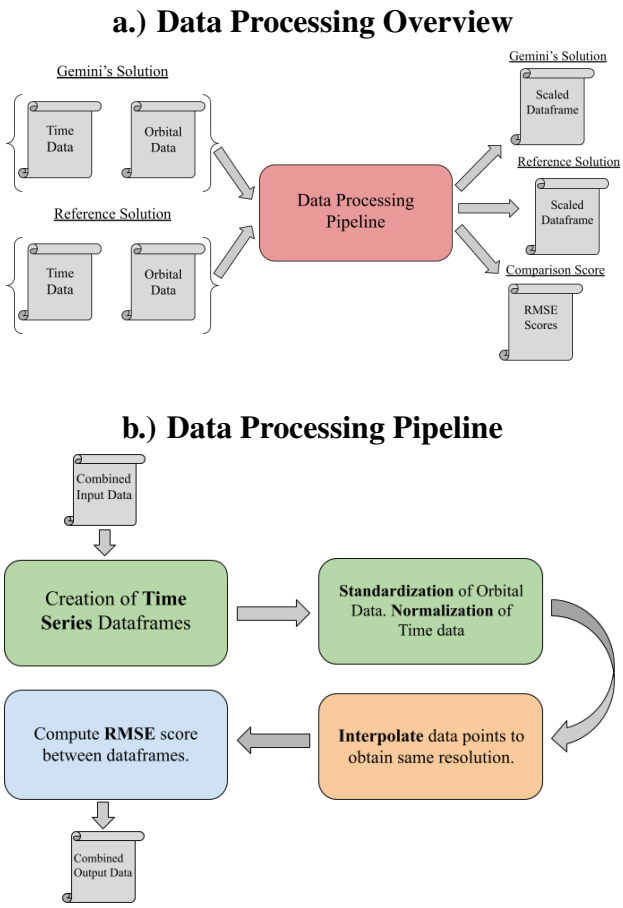


Figure 1: *The first illustration shown in a.) provides a high-level schematic of the pipeline's input (time and orbital data), processing, and output (scaled DataFrames and RMSE scores). The second illustration shown in b.) dives into the pipeline's internal processes, highlighting key functions: creation of time series DataFrames, standardization and normalization of data, interpolation to achieve uniform resolution, and RMSE computation to evaluate discrepancies between datasets. The color-coded blocks represent each processing step for clarity and emphasis on the pipeline's modular structure.*

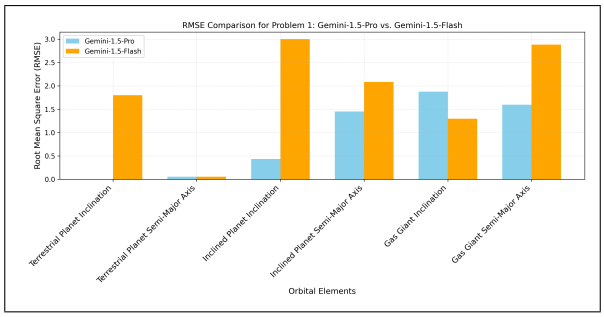
3 Results

The results presented in this section aim to quantify the performance of Google's Gemini Flash and Pro models in solving undergraduate-level orbital dynamics problems. Using the developed benchmark, the models were evaluated based on their ability to produce accurate solutions, with performance measured using the Root Mean Square Error (RMSE) metric across eight orbital dynamics problems. The specific models used from Google's Gemini were `gemini-1.5-flash` for the Flash model and `gemini-1.5-pro` for the Pro model.

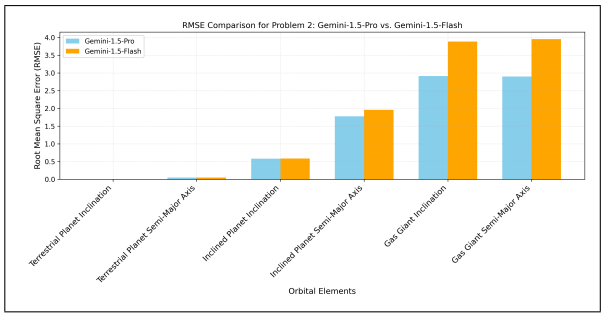
The results present the RMSE scores for the orbital elements calculated in each problem, such as semi-major axis, eccentricity, inclination, or specific parameters unique to the given scenario are presented. These metrics provide a detailed comparison of the performance of Gemini Flash and Pro across the eight problems. Additionally, the overall RMSE for each problem aggregated across all calculated elements, is included to provide a comprehensive measure of each model's accuracy. This analysis highlights both element-specific and overall performance trends between the two models. The calculated RMSE values for each orbital element can be seen in Figure 2.

The results presented in Figure 2 indicate that the Gemini Pro model achieves a notably lower RMSE score across many aspects of the designed problems, particularly in problems 1 through 6. However, in the final problems, 7 and 8, both models display a shared struggle in providing accurate data for the rouge planet problem. A pattern in the RMSE data was also identified: the Gemini Pro model tends to achieve lower RMSE scores for orbital elements associated with the orbital bodies most relevant to the scenario.

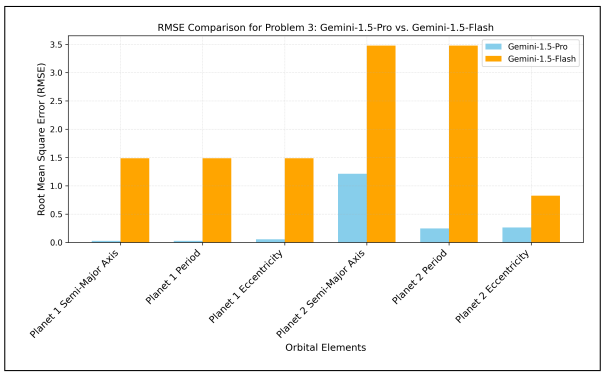
To provide a comprehensive comparison of the overall accuracy of Gemini Flash and Pro across the eight orbital dynamics problems, the total RMSE values for each problem are presented in Figure 3. This chart enables a clear visualization of the differences in accuracy between the two models across varying levels of problem complexity.



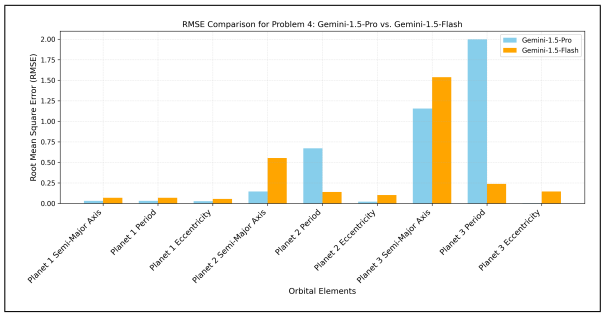
(a) *Problem 1*



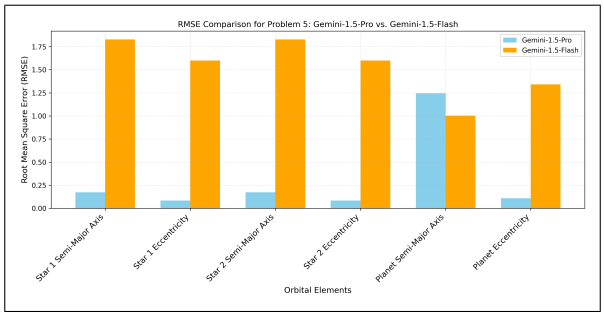
(b) *Problem 2*



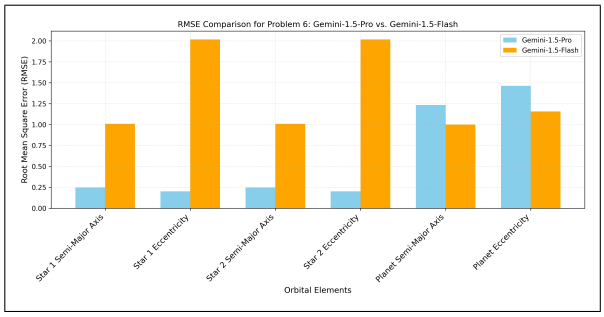
(c) *Problem 3*



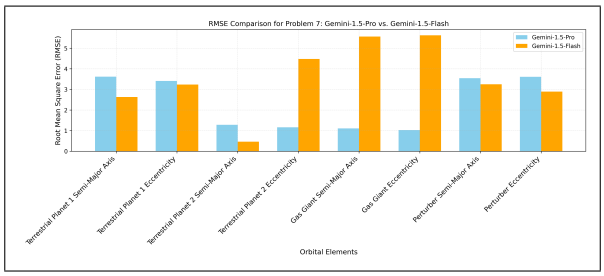
(d) *Problem 4*



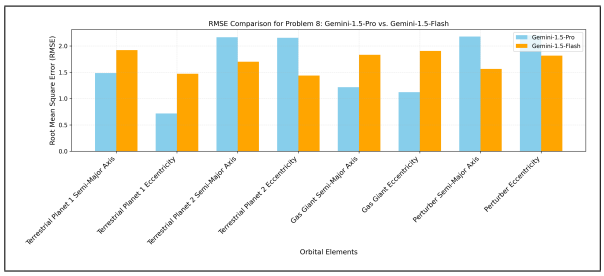
(e) *Problem 5*



(f) *Problem 6*



(g) *Problem 7*



(h) *Problem 8*

Figure 2: *RMSE comparison for the eight orbital dynamics problems. The orange bar graphs represent the Gemini Flash model and the blue bar graphs represent the Gemini Pro model. The superior capabilities of the Gemini Pro model are evident across several problems, particularly in problems 1 through 6. However, both models exhibit notably poor performance in problems 7 and 8.*

The data presented in Figure 3 was observed to illustrate an overall better performance in Gemini's Pro model for handling orbital dynamic problems. However, Gemini's Flash model consis-

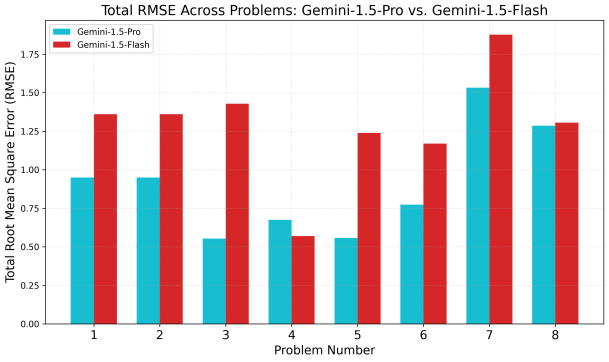


Figure 3: *Total RMSE values for Gemini Flash and Pro across the eight orbital dynamics problems. Each bar represents the aggregated RMSE across all tracked orbital elements in a single problem, providing a comprehensive comparison of model performance.*

tently obtained much higher RMSE values across most problems.

4 Discussion

4.1 Interpretation of Results

This study provided several significant findings regarding the evaluation of Gemini's models in the context of orbital dynamics. A clear trend emerging from various problems analyzed is that the `gemini-1.5-pro` model consistently outperforms the `gemini-1.5-flash` model. This is particularly evident in problems 1 through 6 shown in Figure 2. The same overall result is shown in the overall RMSE scores illustrated in Figure 3. This outcome is anticipated, as the Pro model is better equipped to manage complex reasoning tasks compared to the Flash model. However, both models seem to exhibit poor performance in problems 7 and 8, which involved scenarios of a rogue planet causing instability among the innermost planets. These particular problems may have introduced complexities into the simulation that the current models were unable to handle effectively. Another interesting pattern is that the Pro model tends to achieve lower RMSE scores for orbital elements that are nearer to the center of mass. This may be attributed to the nature of designed problems, where the bodies that are of most interest in the simulation are closer to the center of mass. This may prompt the Pro model to prioritize delivering accurate data for these orbiting bodies. In contrast, bodies that remain relatively unchanged exhibit significantly higher RMSE scores. This indicates that the Pro model is choosing a less precise approach when it comes to capturing fine details of these stable masses, such as the gas giants involved in many of the scenarios.

However, it is evident from numerous problems used in the evaluation that both models still fall short in delivering accurate data that aligns with the reference solution. This suggests that the Gemini large language models may not yet be fully prepared to tackle complex simulation challenges with the level of accuracy required by some researchers. This is especially important in astrophysical simulations where precise data is critical in many aspects of the field. It is worth noting that RMSE scores, while offering a measure of a model's accuracy, are relative values that depend on the scale and nature of the measurements being analyzed. For instance, a high RMSE score in fluctuations of eccentricity could represent either a significant difference in the results or a relatively minor one.

4.2 Challenges and Limitations

While this study provides valuable evaluations of the performance of Gemini's models in solving orbital dynamics problems, several challenges and limitations must be acknowledged. The variability of AI-generation solutions gives rise to many different approaches to a given problem. LLMs like Gemini generate different approaches every time the scenario is prompted, offering differing levels of accuracy. However, this study focused on evaluating the model's performance based on a single run per problem. This approach may not fully capture their abilities, as it limits the possible approaches it may generate to a problem.

Additionally, this study did not examine the performance of the models when incorporating feedback mechanisms. AI performance is significantly enhanced when providing iterative feedback, such as identifying errors, intuition faults, or providing its score compared to reference so-

lutions. Providing feedback can enable models to refine their responses and produce more accurate results over successive attempts. By excluding feedback loops, the benchmark provided may seriously undermine the performance of the models to solve complex orbital dynamics problems.

4.3 Future Endeavors

The findings and limitations of this study give rise to several promising directions for future research. By addressing aspects of LLM performance that appear to have been overlooked, many enhancements could be made to refine the benchmarking framework. In particular, incorporating many runs of an LLM on each problem allows for a more comprehensive evaluation of the model's ability to apply different approaches, potentially leading to a lower RMSE score. Furthermore, providing iterative feedback to this process would further enhance its ability to reduce the RMSE score. This approach could provide a more robust understanding of the true capabilities of the models. Finally, including more complex and realistic scenarios, such as long-term stability in planetary systems or chaotic multi-body systems, would further enhance the benchmark's relevance to real-world astrophysical research.

5 Conclusion

This study assessed the performance of Gemini's large language models, `gemini=1.5-pro` and `gemini-1.5-flash`, in their abilities to solve undergraduate-level orbital dynamics problems using a physics-grounded benchmark. The results demonstrated that while the Pro model performs significantly better than the Flash model in most of the designed astrophysical scenarios, both models struggled with problems involving more complex setups and close interactions, such as the rouge planet encounters. These findings suggest that although AI models like Gemini hold the potential to solve complex physics problems, they are not yet capable of providing the required precision for research-grade results.

In addition, this study highlights the potential of AI as a self-learning assistant in orbital dynamics research, particularly in providing initial approaches or aiding with computational tasks. However, the lack of precision in some of the evaluated scenarios emphasizes the need for further development of AI models and benchmarking techniques. Incorporation of multiple run systems on the scenario, in addition to iterative feedback mechanisms, couple provide a more comprehensive understanding of a models performance.

Overall, this research contributes to an ongoing exploration of AI's potential in scientific research and discovery. The insights gained from this study lay out the potential for future advancements in benchmarking frameworks as well as the foundation for quantifying their performance. By establishing a foundation for quantifying AI's performance, this study opens new horizons for future innovations that could redefine the role of AI in the scientific field.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Menou, for his invaluable guidance, continuous support, and insightful discussions, which were instrumental in deepening my understanding of this study. His mentorship has been truly essential to my academic growth. I would also like to extend my heartfelt thanks to my dear friend, Ivan Pakhomov, whose thoughtful advice, encouragement, and unwavering support were vital throughout the process of writing this paper. I am truly grateful for his friendship and assistance.

A Appendix: Code Repository

The Python code used for this study is available on GitHub. This includes all the simulated orbital dynamics scenarios as well as the data processing pipeline involved in the benchmark framework. You can access the code by clicking the following link: GitHub Repository: Orbital Dynamics Simulation Code

References

Ali-Dib, M., & Menou, K. (2024). Physics simulation capabilities of llms. https://arxiv.org/abs/2312.02091

Huang, Q., Vora, J., Liang, P., & Leskovec, J. (2024). Mlagentbench: Evaluating language agents on machine learning experimentation. https://arxiv.org/abs/2310.03302

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. https://arxiv.org/abs/2305.18486

Liu, Z., Chai, Y., & Li, J. (2024). Towards fully autonomous research powered by llms: Case study on simulations. https://arxiv.org/abs/2408.15512

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The ai scientist: Towards fully automated open-ended scientific discovery. https://arxiv.org/abs/2408.06292

Van Noorden, R., & Perkel, J. M. (2023). Ai and science: What 1,600 researchers think. *Nature News*. https://www.nature.com/articles/d41586-023-02980-0

Wang, H., Fu, T., & Du, Y. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*, 47–60. https://doi.org/10.1038/s41586-023-06221-2