
Harnessing the Power of Adversarial Prompting and Large Language Models for Robust Hypothesis Generation in Astronomy

Ioana Ciucă^{*1,2} Yuan-Sen Ting^{*1,2} Sandor Kruk³ Kartheik Iyer⁴

Abstract

This study investigates the application of Large Language Models (LLMs), specifically GPT-4, within Astronomy. We employ in-context prompting, supplying the model with up to 1000 papers from the NASA Astrophysics Data System, to explore the extent to which performance can be improved by immersing the model in domain-specific literature. Our findings point towards a substantial boost in hypothesis generation when using in-context prompting, a benefit that is further accentuated by adversarial prompting. We illustrate how adversarial prompting empowers GPT-4 to extract essential details from a vast knowledge base to produce meaningful hypotheses, signaling an innovative step towards employing LLMs for scientific research in Astronomy.

1. Introduction

Significant strides in Natural Language Processing (NLP) have been made possible through attention mechanisms and transformer architecture, leading to the development of Large Language Models (LLMs) such as GPT-4 (Vig, 2019; Brown et al., 2020; Ouyang et al., 2022). These models exhibit extraordinary aptitude in understanding, generating, and interacting with human language. They go beyond deciphering complex linguistic patterns to making non-trivial deductions and forming relationships across diverse contexts (e.g., Devlin et al., 2018; Elkins & Chun, 2020).

Two intriguing facets of these models have stirred excite-

^{*}Equal contribution ¹Research School of Astronomy & Astrophysics, Australian National University, Cotter Rd., Weston, ACT 2611, Australia ²School of Computing, Australian National University, Acton, ACT 2601, Australia ³European Space Astronomy Centre, European Space Agency, Villafranca del Castillo, Madrid 28692, Spain ⁴Columbia Astrophysics Laboratory, Columbia University, New York, NY 10027, USA. Correspondence to: Ioana Ciuca <ioana.ciuca@anu.edu.au>, Yuan-Sen Ting <yuan-sen.ting@anu.edu.au>.

ment for their potential that surpasses their initial intended applications. Firstly, despite LLMs’ propensity to sample posterior means of languages—a factor that can occasionally result in non-trivial hallucination problems—improved performance has been witnessed through in-context prompting (Wang et al., 2022; Wei et al., 2022; Zhang et al., 2022). This enhancement enables them to handle complex, domain-specific tasks (e.g., Radford & Narasimhan, 2018; Brown et al., 2020; Lu et al., 2022). Secondly, these models, when combined with revolutionary technologies like Langchain¹ to provide extensive context to the LLMs, expand their functionality across a wide range of fields.

While methods like the use of adapters (He et al., 2021; Karimi Mahabadi et al., 2021; Hu et al., 2021) can remarkably augment performance for domain-specific tasks through fine-tuning the LLMs, these approaches often prove challenging for institutions without sufficient resources. In this study, we delve into the application of low-cost in-context prompting (Chen et al., 2021; Xie et al., 2021) in the realm of astronomy.

Astronomy offers a compelling case study due to three key reasons. Firstly, although the field is rich in literature, the inclusion of such text in the vast corpus used to train GPT models is probably limited. This lack leads to noticeable hallucination problems when employing naive versions of LLMs (Ciucă et al., 2023). Secondly, unlike domains that focus more on intensive, detailed studies, advancements in astronomy often stem from “connecting the dots” across different subfields due to the universality of underlying physical processes at various scales. This feature fosters the hypothesis that extensive in-context prompting could significantly enhance hypothesis generation if LLMs are initially exposed to a broad range of literature.

Lastly, astronomy’s longstanding “open sky” policy makes it an ideal candidate for in-context prompting research. This policy ensures that most data sets are publicly available immediately or after a short proprietary period (Almeida et al., 2023; Fabricius et al., 2021). Further, the field possesses a comprehensive, well-curated literature database. The internet has enabled the archiving of astronomical knowledge,

¹<https://python.langchain.com>

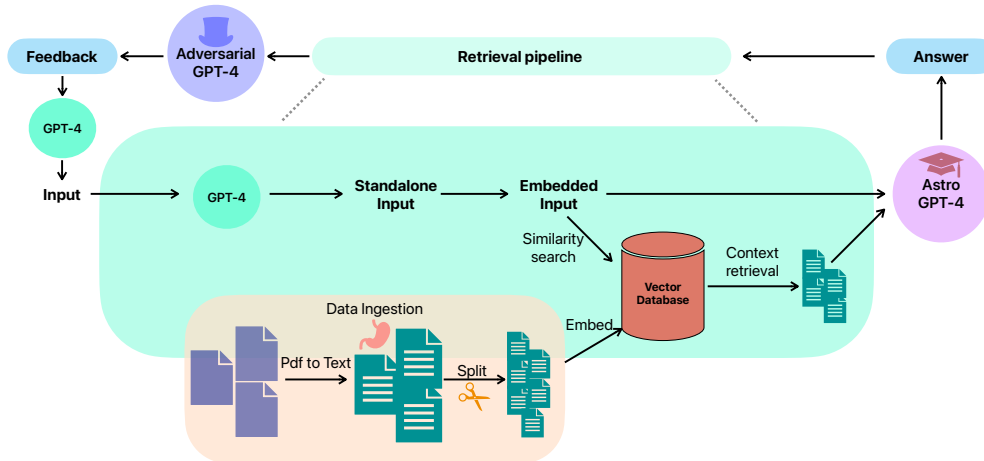


Figure 1. This figure illustrates the adversarial in-context prompting workflow using OpenAI’s GPT-4 model. The procedure begins with the pre-processing and embedding of Galactic Astronomy papers. A similarity search is conducted on the embedded query, and relevant document chunks are retrieved. A further contextual compression is performed to remove irrelevant information from the chunks. These compressed texts serve as input to a GPT-4 instance, which generates an idea. This idea is then critiqued by a second GPT-4 model, and the feedback is moderated by a third GPT-4 model.

with NASA’s Astrophysics Data System hosting over 15 million resources, effectively covering the entire spectrum of astronomical literature utilized by researchers (Accomazzi et al., 2015; Borgman & Wofford, 2021). This accessibility greatly aids our engagement with the astronomy database.

2. Literature retrieval and pre-processing

For this study, we focused our exploration on Galactic Astronomy, utilizing our domain expertise to assess the results. We selected Galactic Astronomy as our focal area due to its integrative nature, fusing knowledge from diverse subfields. The study of galaxy evolution not only incorporates the fundamental understanding of stars and stellar populations (Aouad et al., 2020; Sánchez et al., 2022) but it is also influenced by large-scale cosmological environmental factors (Singh et al., 2020; Whitney et al., 2021). Therefore, studying galaxy evolution provides both exciting challenges and abundant possibilities for harnessing implicit knowledge embedded within the vast network of literature.

Our study includes a selection of 1,000 papers related to Galactic Astronomy from the NASA ADS (Accomazzi et al., 2015) Astronomy collection. Our chosen papers were identified through a database query based on criteria such as ‘Gaia’ appearing in the abstract, publications from the last ten years (since the Gaia launch date), being refereed journal articles, and the inclusion of relevant keywords such as ‘galaxy kinematics and dynamics’, ‘galaxy structure’, ‘galaxy disk’, ‘galaxy halo’, ‘galaxy abundances’, and ‘galaxy evolution’. Our initial query yielded more than 1,000 papers, leading us to prioritize the most recent publications. Our curated collection contains details such as the ArxivID, Publication

Date, Authors, Title, Abstract, Citation, and Key, providing a comprehensive dataset for our analysis. The full dataset as well as the codebase used in our analysis can be found here for reproducibility².

3. Astro-GPT Workflow

Our exploration capitalizes on the abilities of OpenAI’s GPT-4 model (OpenAI, 2023). The first step in in-context prompting involves pre-processing 1,000 papers from the Galactic Astronomy corpus using the langchain library. Each paper, transformed from PDF to text, is subsequently segmented into ‘chunks’ of 1,000 tokens each. These segmented units are then embedded using OpenAI’s text-ada-002 embedding model.

The retrieval phase begins with converting the chat history and input query into a standalone input, which is then embedded. A similarity search is conducted between the embedded query and the vector database. We then use langchain’s contextual compression to filter out irrelevant information from the individual chunks. These final texts, combined with the standalone input, form the foundation upon which a GPT-4 model, having a context window of approximately 8,000 tokens, formulates ideas. To scrutinize the model’s prowess, we design an adversarial experiment. This involves a secondary GPT-4 model that critiques the idea, highlighting its frailties and suggesting potential enhancements. This feedback is reformulated within a feedback-question structure by a third GPT-4 instance and returned to the initial model.

²<https://github.com/errai34/IdeaGPT>

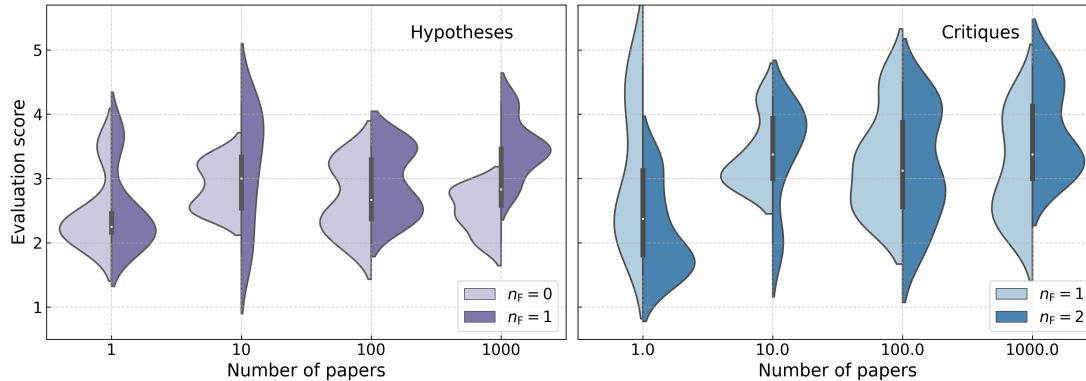


Figure 2. Adversarial prompting and domain-specific context enrichment significantly enhance hypothesis generation quality. 60 hypotheses and 40 critiques generated by the AI were evaluated by two human experts, with the mean scores reported for individual instances. The iterations of adversarial prompting (n_F) was instrumental in driving substantial enhancements in both the quality and consistency of the AI judge and AI generator outputs, particularly when they were supplied with an extensive context ($N = 1000$ papers). Crucially, in the absence of adversarial prompting ($n_F = 0$), the quality of the hypothesis remained stagnant despite the provision of ample context. This observation underscores the stark contrast and superior effectiveness of adversarial prompting.

Implementing our experimental setup, we use N_k papers, where $k \in \{1, 10, 100, 1000\}$. Each sample undergoes hypothesis generation by the ‘Generation GPT-4’ instance (our in-context prompted model on k papers). An adversarial response from ‘Adversarial GPT-4’ follows, which is reformulated by a moderator GPT-4 instance and fed back to the generator model. This cycle, yielding three hypotheses and two critiques per experiment, is repeated twice for each N_k and replicated five times in total. The same approach is applied to 1,000 papers, without resampling, accumulating a total of 60 hypotheses and 40 critiques.

4. Results

4.1. Human Evaluation

Given the qualitative nature of hypothesis generation, we needed an assessment process that, while inherently subjective, would match the expectations of human experts. For this, we involved two domain experts in the field of Galactic Astronomy to evaluate the quality of the generated hypotheses. These were graded based on the number of papers included within the domain-specific context, and we computed the average score from these dual-human evaluations for each hypothesis. The hypotheses are graded based on a rubric of three categories – scientific accuracy, creativity and feasibility, and the average score of these three domains assumed to be the final score. We also evaluated the critiques provided by the AI judge, which had access to the same contextual information.

As illustrated in the left panel of Fig. 2, adversarial prompting proved to be a critical tool in markedly improving hypothesis generation. The quality of hypothesis generation, without adversarial prompting, showed little dependence on the number of papers, suggesting that in-context prompting

alone, while helpful for mitigating hallucination, did not suffice for a comprehensive understanding of the corpus.

The introduction of adversarial prompting considerably altered this outcome. A significant improvement in the quality of hypothesis generation was observed both for the AI generator and the AI judge, even without explicitly aligning the models with human expectations. Notably, adversarial prompting introduced a strong correlation between hypothesis quality and the number of papers reviewed, especially at larger context ($N = 1000$). It also leads to a much more consistent in term of the quality of the hypotheses (and the critiques). The average quality score rose significantly from 2.5 (when 10 papers were used as context, where a score of 3/5 corresponds to a typical hypothesis by a competent PhD student) to a near-expert level of 4/5 when 1,000 papers were included, emphasizing the potential of adversarial prompting in enhancing the quality of scientific hypothesis generation. We refer to the Appendix for examples.

4.2. Exploration of Embeddings

To truly understand the power of adversarial prompting, we first passed the abstracts of our set of 1000 astronomy papers through the `text-ada-002` embedding model and arranged them into a 2D TSNE projection. This captured the contextual differences and similarities of these 1000 papers.

For each hypothesis generated, we determined which papers inspired it by querying the GPT model. In Fig. 3, we visualized this ‘knowledge footprint’ for each hypothesis as black polygons within a green hull representing all the papers GPT-4 had access to. From Fig. 3’s top panels, it’s clear that as the number of papers grows, GPT broadens its scope, drawing on diverse topics to build interconnected hypotheses. We note that, even with a small training pool

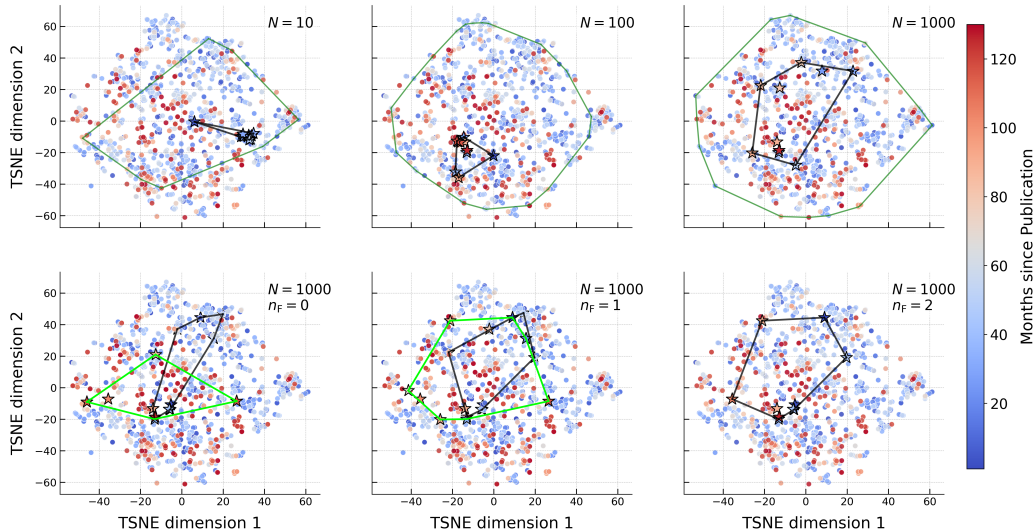


Figure 3. Visual representation of each hypothesis’s ‘knowledge footprint’, depicted as black polygons within the TSNE projection of the abstracts of our corpus comprising 1000 papers. As the quantity of papers consumed increases, the model leverages a more diverse array of topics, thus boosting the quality of the hypothesis (as seen in the top panels). The green hull in the top panel shows the overall knowledge base that the model have access through in-context prompting. The lower panel demonstrates how the ‘knowledge footprint’ evolves with varying quantities n_F of adversarial attacks in the case of 1000 papers. The black polygon signifies the footprint of the original response, while the lime indicates the information utilized by the AI judge for critique.

of 10 papers, we design our experiment in a way that the corpus still covers a wide range of topics, but GPT-4 lacks the context to connect them in a meaningful way, leading to more general hypotheses (see Appendix).

The bottom panel in Fig. 3, concentrating on the case with 1000 papers, explores how the knowledge footprint evolves with different numbers of adversarial attacks. In the preliminary iteration (bottom left), the judge ingeniously identifies areas of critique based on knowledge overlooked by the original response. This compels the generator to expand its scope further (as shown in the bottom middle and right panels), to appropriately address the criticism. As shown in some of the examples in the Appendix, adversarial prompting allows the GPT-4 model to genuinely benefit from a large number of contextual inputs, guiding the model towards a more coherent understanding of the topic, rather than creating a mere ‘fact jumble’. The specific examples of hypotheses and corresponding critiques are shown in the Appendix.

5. Conclusion and Future Direction

In this research, we delved into a detailed examination of the GPT-4 model’s ability to propose novel hypotheses in the domain of Galactic Astronomy using in-context prompting. Our findings confirm that in-context prompting significantly mitigates hallucination, leading to the generation of meaningful hypotheses that can compete with substantive thesis topics, as evaluated by domain experts. Importantly, we found that the inclusion of adversarial prompts enables the

model to progressively enhance its performance based on the number of contextual papers. However, a naive implementation of in-context prompting without adversarial judgment fails to replicate this improvement.

While our study marks the inception of a pioneering exploration of in-context prompting for scientific hypothesis generation, it’s clear that this dynamic field is rapidly evolving. Thus, we have identified several crucial areas for enhancement. These areas include (a) an improved and automated evaluation method for hypotheses. We have observed that while the AI judge can assist the AI generator, improvements are primarily in technical detail rather than deep insights. We propose leveraging well-curated question-and-answer pairs (e.g. Dugan et al., 2022) to better align the judge with human expectations. (b) Instead of focusing solely on hypothesis generation, integrating other downstream tasks and finite fine-tuning models with smaller adapter models could potentially improve inferences. We have commenced curating metadata from ADS to better design these tasks.

6. Broader Impact

In this study, our focus on in-context prompting, rather than the more computationally intensive fine-tuning, is inspired by the aim to democratize the utilization of LLMs for scientific inquiry. Current GPT models, due to their immense parameter sets, often render fine-tuning impractical. In the era of Large Language Models, it is crucial to determine whether all academic institutions, regardless of size or available computational resources, can keep pace with these

rapidly advancing technologies.

This critical inquiry forms the crux of our study, and our findings present a hopeful outlook. Our study indicates that, with the right strategies and approaches with ‘humans in the loop’ as domain experts, the barrier to harnessing the full power of these advanced LLMs can be lowered. As a result, we envision a future where all institutions, regardless of size or resources, can contribute to and benefit from the swift advancements in AI, enhancing the collective endeavour of scientific discovery. Our journey into this new frontier of Large Language Models is just beginning, and it promises a thrilling ride full of unexpected insights and revolutionary breakthroughs.

References

- Accomazzi, A., Kurtz, M. J., Henneken, E. A., Chyla, R., Luker, J., Grant, C. S., Thompson, D. M., Holachek, A., Dave, R., and Murray, S. S. ADS: The Next Generation Search Platform. In Holl, A., Lesteven, S., Dietrich, D., and Gasperini, A. (eds.), *Open Science at the Frontiers of Librarianship*, volume 492 of *Astronomical Society of the Pacific Conference Series*, pp. 189, April 2015. doi: 10.48550/arXiv.1503.04194.
- Almeida, A., Anderson, S. F., Argudo-Fernández, M., Badenes, C., Barger, K., Barrera-Ballesteros, J. K., Bender, C. F., Benitez, E., Besser, F., Bizyaev, D., Blanton, M. R., Bochanski, J., Bovy, J., Brandt, W. N., Brownstein, J. R., Buchner, J., Bulbul, E., Burchett, J. N., Díaz, M. C., Carlberg, J. K., Casey, A. R., Chandra, V., Cherinka, B., Chiappini, C., Coker, A. A., Comparat, J., Conroy, C., Contardo, G., Cortes, A., Covey, K., Crane, J. D., Cunha, K., Dabbieri, C., au2, J. W. D. J., Davis, M. C., Lee, N. D., Delgado, J. E. M., Demasi, S., Mille, F. D., Donor, J., Dow, P., Dwelly, T., Eracleous, M., Eriksen, J., Fan, X., Farr, E., Frederick, S., Fries, L., Frinchaboy, P., Gaensicke, B. T., Ge, J., Ávila, C. G., Grabowski, K., Grier, C., Guiglion, G., Gupta, P., Hall, P., Hawkins, K., Hayes, C. R., Hermes, J. J., Hernández-García, L., Hogg, D. W., Holtzman, J. A., Ibarra-Medel, H. J., Ji, A., Jofre, P., Johnson, J. A., Jones, A. M., Kinemuchi, K., Kluge, M., Koekemoer, A., Kollmeier, J. A., Kounkel, M., Krishnarao, D., Krumpe, M., Lacerna, I., Lago, P. J. A., Laporte, C., Liu, A., Liu, C., Liu, X., Lopes, A. R., Macktoobian, M., Malanushenko, V., Maoz, D., Masseron, T., Masters, K. L., Matijevic, G., McBride, A., Medan, I., Merloni, A., Morrison, S., Myers, N., Mészáros, S., Negrete, C. A., Nidever, D. L., Nitschelm, C., Oravetz, A., Oravetz, D., Pan, K., Peng, Y., Pinsonneault, M. H., Pogge, R., Qiu, D., de Andrade Queiroz, A. B., Ramirez, S. V., Rix, H.-W., Rosso, D. F., Runnoe, J., Salvato, M., Sanchez, S. F., Santana, F. A., Saydjari, A., Sayres, C., Schlafman, K. C., Schneider, D. P., Schwope, A., Serna, J., Shen, Y., Sobek, J., Song, Y.-Y., Souto, D., Spoo, T., Stassun, K. G., Steinmetz, M., Straumit, I., Stringfellow, G., Sánchez-Gallego, J., Taghizadeh-Popp, M., Tayar, J., Thakar, A., Tissera, P. B., Tkachenko, A., Toledo, H. H., Trakhtenbrot, B., Trincado, J. G. F., Troup, N., Trump, J. R., Tuttle, S., Ulloa, N., Vazquez-Mata, J. A., Alfaro, P. V., Villanova, S., Wachter, S., Weijmans, A.-M., Wheeler, A., Wilson, J., Wojno, L., Wolf, J., Xue, X.-X., Ybarra, J. E., Zari, E., and Zasowski, G. The eighteenth data release of the sloan digital sky surveys: Targeting and first spectra from sdss-v, 2023.
- Aouad, C. J., James, P. A., and Chilingarian, I. V. Coupling local to global star formation in spiral galaxies: the effect of differential rotation. *Monthly Notices of the Royal Astronomical Society*, 496(4):5211–5226, jul 2020. doi: 10.1093/mnras/staa1945.
- Borgman, C. L. and Wofford, M. F. From Data Processes to Data Products: Knowledge Infrastructures in Astronomy. *arXiv e-prints*, art. arXiv:2109.01707, September 2021. doi: 10.48550/arXiv.2109.01707.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.
- Chen, Y., Zhong, R., Zha, S., Karypis, G., and He, H. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*, 2021.
- Ciucă, I., Kawata, D., Ting, Y.-S., Grand, R. J. J., Miglio, A., Hayden, M., Baba, J., Fragkoudi, F., Monty, S., Buder, S., and Freeman, K. Chasing the impact of the Gaia-Sausage-Enceladus merger on the formation of the Milky Way thick disc. *Monthly Notices of the Royal Astronomical Society*, March 2023. doi: 10.1093/mnras/slad033.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, art. arXiv:1810.04805, October 2018. doi: 10.48550/arXiv.1810.04805.
- Dugan, L., Miltsakaki, E., Upadhyay, S., Ginsberg, E., Gonzalez, H., Choi, D., Yuan, C., and Callison-Burch, C. A feasibility study of answer-agnostic question generation for education. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1919–1926, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.151.

- Elkins, K. and Chun, J. Can gpt-3 pass a writer’s turing test? *Journal of Cultural Analytics*, 5(2), 2020.
- Fabricius, C., Luri, X., Arenou, F., Babusiaux, C., Helmi, A., Muraveva, T., Reylé, C., Spoto, F., Vallenari, A., Antoja, T., Balbinot, E., Barache, C., Bauchet, N., Bragaglia, A., Busonero, D., Cantat-Gaudin, T., Carrasco, J. M., Diakité, S., Fabrizio, M., Figueras, F., Garcia-Gutierrez, A., Garofalo, A., Jordi, C., Kervella, P., Khanna, S., Leclerc, N., Licata, E., Lambert, S., Marrese, P. M., Masip, A., Ramos, P., Robichon, N., Robin, A. C., Romero-Gómez, M., Rubele, S., and Weiler, M. <i>gaia</i>early data release 3. *Astronomy & Astrophysics*, 649:A5, apr 2021. doi: 10.1051/0004-6361/202039834.
- He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J.-W., Bing, L., and Si, L. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv e-prints*, art. arXiv:2106.09685, June 2021. doi: 10.48550/arXiv.2106.09685.
- Karimi Mahabadi, R., Henderson, J., and Ruder, S. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34: 1022–1035, 2021.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521. Curran Associates, Inc., 2022.
- OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. *arXiv e-prints*, art. arXiv:2203.02155, March 2022. doi: 10.48550/arXiv.2203.02155.
- Radford, A. and Narasimhan, K. Improving language understanding by generative pre-training. 2018.
- Sánchez, S. F., Gómez Medina, D. C., Barrera-Ballesteros, J. K., Galbany, L., Bolatto, A., and Wong, T. The local and global relations between Σ_* , Σ_{SFR} and Σ_{mol} that regulate star-formation. *arXiv e-prints*, art. arXiv:2212.03738, December 2022. doi: 10.48550/arXiv.2212.03738.
- Singh, A., Mahajan, S., and Bagla, J. S. Study of galaxies on large-scale filaments in simulations. *Monthly Notices of the Royal Astronomical Society*, 497(2):2265–2275, jul 2020. doi: 10.1093/mnras/staa1913.
- Fig, J. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*, 2019.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- Whitney, A., Ferreira, L., Conselice, C. J., and Duncan, K. Galaxy evolution in all five CANDELS fields and IllustrisTNG: Morphological, structural, and the major merger evolution to $z \sim 3$. *The Astrophysical Journal*, 919(2):139, oct 2021. doi: 10.3847/1538-4357/ac1422.
- Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.
- Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

A. Exemplary Hypotheses and Critiques Generated

This section presents three types of hypotheses generated through our in-context prompting and adversarial prompting techniques, along with the associated critiques provided by our adversarial GPT-4 model, which led to substantial improvements in hypothesis generation.

The first example demonstrates a hypothesis generated after subjecting the model to in-context prompting through 1000 papers, followed by two rounds of adversarial attacks.

I propose an investigation into the role of dark matter mini-halos in the formation and evolution of globular clusters (GCs) in the Milky Way. This study would combine the high-precision astrometric data from Gaia EDR3 and DR3 (Gaia Collaboration et al. 2016, 2020) with spectroscopic data from large ground-based surveys like APOGEE, GALAH, SDSS SEGUE, and LAMOST to characterize the dynamics of stars in the peripheral regions of GCs. The aim would be to determine whether these GCs are embedded in dark matter mini-halos, which could provide critical insights into their origins (Peebles 1984; Peñarrubia et al. 2017). The proposed research would build upon the probabilistic approach developed by Kuzma et al. (2021) for studying the peripheral regions of GCs, which utilizes a mixture model in spatial and proper motion space to model cluster, extra-tidal, and contaminant stellar populations. By extending this approach to include the effects of dark matter mini-halos on the kinematics of stars in GC outskirts, we can test the hypothesis that dark matter plays a significant role in the formation and evolution of GCs. Furthermore, this study would provide a better understanding of the distribution and properties of dark matter in the Milky Way, contributing to the broader field of near-field cosmology. Integrating this information with the existing knowledge of the hierarchical assembly of the Milky Way (Viswanathan et al. 2023) and the role of rapid gas accretion in the inner Galactic disc (Snaith et al. 2021) would help paint a more comprehensive picture of our Galaxy's formation history and its underlying dark matter distribution.

This hypothesis received high approval from the human jury due to its broad contextual understanding of Galactic Astronomy, a concrete thesis proposal, appropriate references for idea origins, and building upon existing methodologies and studies.

Conversely, the following hypothesis was produced when the in-context prompting was restricted to 10 papers. Hypotheses based on limited exposure to literature often appear generic, lacking substantial insights beyond minor improvements to existing work.

I propose a study investigating the role of intermediate-mass gas-rich mergers in the chemical and kinematic evolution of the Milky Way-like galaxies across a range of redshifts. This research would build upon the findings of Ciucă et al. (2023), which suggest that the Gaia-Sausage-Enceladus (GSE) merger significantly impacted the formation of the Galactic disc and the transition from the thick disc to the thin disc. The study could involve a combination of observational data from large spectroscopic surveys such as APOGEE, GALAH, and Gaia-ESO, as well as the utilization of high-resolution cosmological simulations like the Auriga project. Key objectives would include characterizing the chemical abundance patterns and kinematic properties of stars in the disc and halo components during and after the mergers and probing the impact of these mergers on the radial distribution and migration of stars across the Galactic disc. By focusing on intermediate-mass gas-rich mergers, this research would fill a knowledge gap in our understanding of the role these events play in shaping the structure and evolution of Milky Way-like galaxies. The findings would also provide valuable insights for refining galaxy formation models and contribute to a more comprehensive understanding of the complex history of our Galaxy.

Without sufficient context, the model tends to extrapolate based on existing literature, resulting in incremental, less innovative suggestions.

Importantly, despite the exposure to a corpus of 1000 papers, without adversarial attacks, the hypotheses can often seem like a jumble of facts, as shown in the example below:

I propose a comprehensive, multi-wavelength exploration of the uncharted regions of the Galactic plane, particularly in the South Galactic cap, including areas below a declination of -65° . This astronomical research initiative can be achieved by combining existing large photometric, spectroscopic, and astrometric surveys, such as Gaia, LAMOST, APOGEE, and GALAH, with new high-resolution surveys aimed at mapping the uncovered areas of the Milky Way. This undertaking should prioritize the detection and characterization of previously undiscovered ultra-faint dwarf galaxies, chemically peculiar stars, and stellar streams, as well as the evaluation of their origins and potential implications for our understanding of the Milky Way's formation and evolution. This research proposal is inspired by the unexplored regions mentioned in Antoja et al. (2015), along with the possibilities for improvement suggested by Gorski and Barmby (2020) and the advancements in large-scale surveys discussed in studies such as Yuan et al. (2014) and An and Beers (2020). Expanding our knowledge of the uncharted regions of the Galactic plane will significantly contribute to our understanding of the structure, formation, and evolution of the Milky Way. Additionally, it will allow astronomers to identify and assess previously unknown populations of stars and their kinematic properties, further refining our knowledge of the various components and phenomena that shape our Galaxy.

Demonstrating adversarial prompting is critical to ensure that the GPT can harness the broad context provided in a meaningful way.

Finally, an example of high-quality critique provided by the adversarial GPT-4 model is showcased. These adversarial critiques, after being primed with 1000 papers, are highly accurate and can rival the quality of expert human critique.

The interdisciplinary approach proposed aims to provide a comprehensive understanding of the Milky Way's gravitational potential by combining photometric data, high-resolution simulations, and novel techniques to analyze phase-space spirals. However, there are several weaknesses and limitations to consider. First, the proposal relies heavily on photometric data, which, although less affected by sampling biases, may still be prone to uncertainties due to dust obscuration and crowding (Rix and Bovy 2013). Additionally, the accuracy of photometric metallicity measurements may be limited compared to spectroscopic data (An and Beers 2021). Second, while high-resolution simulations may enable better modeling of the Milky Way's potential, they are computationally expensive, and current mass and spatial resolutions may not be sufficient for accurate analysis (Darling and Widrow 2018). Furthermore, the complexity of the Milky Way, including its non-steady-state nature and the presence of various structures such as spiral arms and bars, adds challenges to the modeling process (Kipper et al. 2018). Third, novel techniques to analyze phase-space spirals may be subject to different sources of systematic bias (Widmark et al. 2021). The proposal does not fully address how these biases will be accounted for or the potential limitations of these techniques in the context of a complex and evolving galaxy. Lastly, the proposal does not extensively discuss the implications of various selection biases in spectroscopic surveys (Sanders and Das 2018), nor the challenges in incorporating the complexity of our galaxy in large-scale cosmological models (Everall et al. 2021c). In conclusion, while the interdisciplinary approach has merits, further work is needed to address these weaknesses and limitations for a more accurate and comprehensive understanding of the Milky Way's gravitational potential.