Draft version April 13, 2023
Typeset using LaTeX default style in AASTeX631

Galactic ChitChat: Using Large Language Models to Converse with Astronomy Literature

Ioana Ciucă[1, 2] and Yuan-Sen Ting[1, 2]

[1] *Research School of Astronomy & Astrophysics, Australian National University,*
Cotter Rd., Weston, ACT 2611, Australia.

[2] *School of Computing, Australian National University,*
Acton, ACT 2601, Australia.

ABSTRACT

We demonstrate the potential of the state-of-the-art OpenAI GPT-4 large language model to engage in meaningful interactions with Astronomy papers using in-context prompting. To optimize for efficiency, we employ a distillation technique that effectively reduces the size of the original input paper by 50%, while maintaining the paragraph structure and overall semantic integrity. We then explore the model's responses using a multi-document context (ten distilled documents). Our findings indicate that GPT-4 excels in the multi-document domain, providing detailed answers contextualized within the framework of related research findings. Our results showcase the potential of large language models for the astronomical community, offering a promising avenue for further exploration, particularly the possibility of utilizing the models for hypothesis generation.

Keywords: Computational methods (1965) — Galactic archaeology (2178)

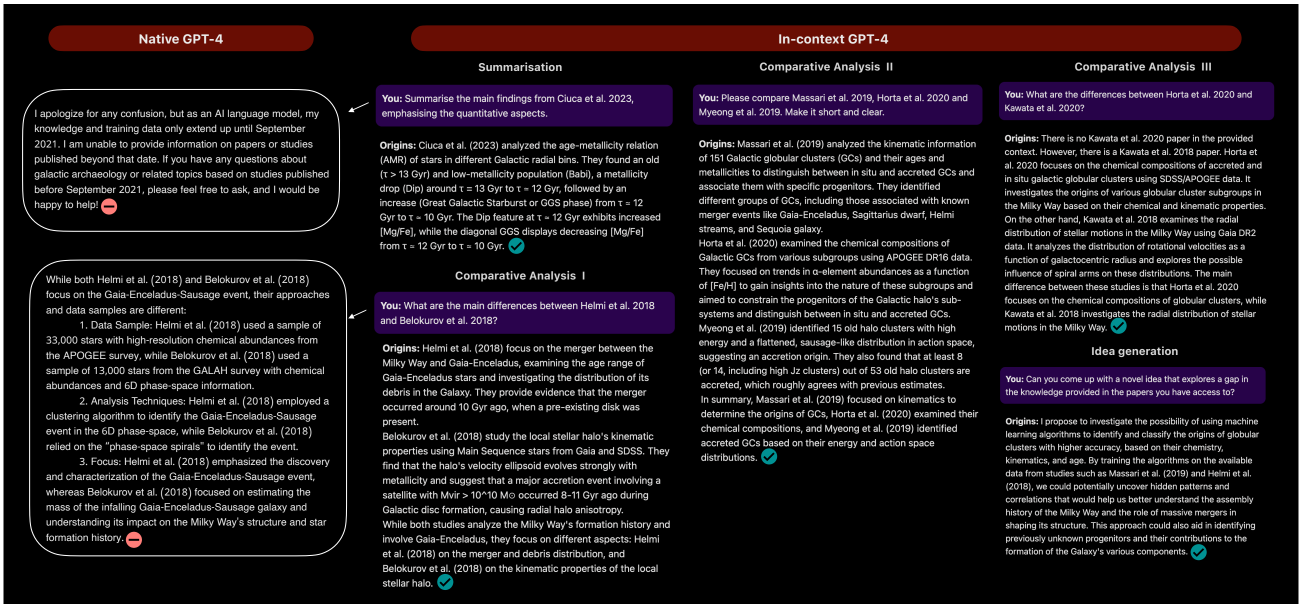
1. INTRODUCTION

Large language models (LLMs) have significantly advanced natural language processing, allowing machines to process and generate intricate text with remarkable quality (e.g., Devlin et al. 2018; Brown et al. 2020; Chowdhery et al. 2022; Bubeck et al. 2023). In this study, we employ 'in-context prompting' [1] for the SOTA GPT-4 model (OpenAI 2023) as a practical solution to tap into the model's capabilities without the need for fine-tuning (e.g., Howard & Ruder 2018; Wei et al. 2022) or incorporating specialized adapters (e.g, Pfeiffer et al. 2020; Hu et al. 2021). This technique enables us to harness the potential of GPT-4, which has not been specifically trained for Astronomy, to engage with Astronomy literature, a field characterized by its complexity and richness of information. Our goal for the future is to use LLMs in conjunction with expert knowledge to identify possible gaps in our knowledge base and generate new scientific hypotheses in Astronomy.

2. METHODS

We select ten papers in Galactic Archaeology, including Helmi et al. (2018); Belokurov et al. (2018); Kawata et al. (2018); Massari et al. (2019); Myeong et al. (2018); Horta et al. (2020); Ciucă et al. (2023). We extract the text from the PDF and divide it, then pass to the GPT-4 powered ChatGPT (OpenAI 2023) using a **distillation prompt** that contains: 'Distill each paragraph of the given text, maintaining the same number of paragraphs and structure. Limit the word count to 50% of the original, and ensure references are included.' We essentially 'compress' the paper to half its size. Unlike summarisation, this process ensures that the distilled paper inherits the original paper's paragraph structure and associated information. The distillation process allows us to use more processed papers efficiently for a context window with a given $\sim$8K token limit.

[1] While developing this project, we came across `paperqa` (White 2023), which performs QA on academic papers, similar to our approach. However, we employ a distillation technique and consider the chat history.



I apologize for any confusion, but as an AI language model, my knowledge and training data only extend up until September 2021. I am unable to provide information on papers or studies published beyond that date. If you have any questions about galactic archaeology or related topics based on studies published before September 2021, please feel free to ask, and I would be happy to help! ⊖

While both Helmi et al. (2018) and Belokurov et al. (2018) focus on the Gaia-Enceladus-Sausage event, their approaches and data samples are different:

1. Data Sample: Helmi et al. (2018) used a sample of 33,000 stars with high-resolution chemical abundances from the APOGEE survey, while Belokurov et al. (2018) used a sample of 13,000 stars from the GALAH survey with chemical abundances and 6D phase-space information.
2. Analysis Techniques: Helmi et al. (2018) employed a clustering algorithm to identify the Gaia-Enceladus-Sausage event in the 6D phase-space, while Belokurov et al. (2018) relied on the "phase-space spirals" to identify the event.
3. Focus: Helmi et al. (2018) emphasized the discovery and characterization of the Gaia-Enceladus-Sausage event, whereas Belokurov et al. (2018) focused on estimating the mass of the infalling Gaia-Enceladus-Sausage galaxy and understanding its impact on the Milky Way's structure and star formation history. ⊖

Summarisation

You: Summarise the main findings from Ciuca et al. 2023, emphasising the quantitative aspects.

Origins: Ciuca et al. (2023) analyzed the age-metallicity relation (AMR) of stars in different Galactic radial bins. They found an old (τ > 13 Gyr) and low-metallicity population (Babi), a metallicity drop (Dip) around τ = 13 Gyr to τ ≃ 12 Gyr, followed by an increase (Great Galactic Starburst or GGS phase) from τ ≃ 12 Gyr to τ ≃ 10 Gyr. The Dip feature at τ ≃ 12 Gyr exhibits increased [Mg/Fe], while the diagonal GGS displays decreasing [Mg/Fe] from τ ≃ 12 Gyr to τ ≃ 10 Gyr. ✓

Comparative Analysis I

You: What are the main differences between Helmi et al. 2018 and Belokurov et al. 2018?

Origins: Helmi et al. (2018) focus on the merger between the Milky Way and Gaia-Enceladus, examining the age range of Gaia-Enceladus stars and investigating the distribution of its debris in the Galaxy. They provide evidence that the merger occurred around 10 Gyr ago, when a pre-existing disk was present.
Belokurov et al. (2018) study the local stellar halo's kinematic properties using Main Sequence stars from Gaia and SDSS. They find that the halo's velocity ellipsoid evolves strongly with metallicity and suggest that a major accretion event involving a satellite with Mvir > 10^10 M⊙ occurred 8-11 Gyr ago during Galactic disc formation, causing radial halo anisotropy.
While both studies analyze the Milky Way's formation history and involve Gaia-Enceladus, they focus on different aspects: Helmi et al. (2018) on the merger and debris distribution, and Belokurov et al. (2018) on the kinematic properties of the local stellar halo. ✓

Comparative Analysis II

You: Please compare Massari et al. 2019, Horta et al. 2020 and Myeong et al. 2019. Make it short and clear.

Origins: Massari et al. (2019) analyzed the kinematic information of 151 Galactic globular clusters (GCs) and their ages and metallicities to distinguish between in situ and accreted GCs and associate them with specific progenitors. They identified different groups of GCs, including those associated with known merger events like Gaia-Enceladus, Sagittarius dwarf, Helmi streams, and Sequoia galaxy.
Horta et al. (2020) examined the chemical compositions of Galactic GCs from various subgroups using APOGEE DR16 data. They focused on trends in α-element abundances as a function of [Fe/H] to gain insights into the nature of these subgroups and aimed to constrain the progenitors of the Galactic halo's sub-systems and distinguish between in situ and accreted GCs.
Myeong et al. (2019) identified 15 old halo clusters with high energy and a flattened, sausage-like distribution in action space, suggesting an accretion origin. They also found that at least 8 (or 14, including high Jz clusters) out of 53 old halo clusters are accreted, which roughly agrees with previous estimates.
In summary, Massari et al. (2019) focused on kinematics to determine the origins of GCs, Horta et al. (2020) examined their chemical compositions, and Myeong et al. (2019) identified accreted GCs based on their energy and action space distributions. ✓

Comparative Analysis III

You: What are the differences between Horta et al. 2020 and Kawata et al. 2020?

Origins: There is no Kawata et al. 2020 paper in the provided context. However, there is a Kawata et al. 2018 paper. Horta et al. 2020 focuses on the chemical compositions of accreted and in situ galactic globular clusters using SDSS/APOGEE data. It investigates the origins of various globular cluster subgroups in the Milky Way based on their chemical and kinematic properties. On the other hand, Kawata et al. 2018 examines the radial distribution of stellar motions in the Milky Way using Gaia DR2 data. It analyzes the distribution of rotational velocities as a function of galactocentric radius and explores the possible influence of spiral arms on these distributions. The main difference between these studies is that Horta et al. 2020 focuses on the chemical compositions of globular clusters, while Kawata et al. 2018 investigates the radial distribution of stellar motions in the Milky Way. ✓

Idea generation

You: Can you come up with a novel idea that explores a gap in the knowledge provided in the papers you have access to?

Origins: I propose to investigate the possibility of using machine learning algorithms to identify and classify the origins of globular clusters with higher accuracy, based on their chemistry, kinematics, and age. By training the algorithms on the available data from studies such as Massari et al. (2019) and Helmi et al. (2018), we could potentially uncover hidden patterns and correlations that would help us better understand the assembly history of the Milky Way and the role of massive mergers in shaping its structure. This approach could also aid in identifying previously unknown progenitors and their contributions to the formation of the Galaxy's various components. ✓

Figure 1. The GPT-4 model responses to expert-level questioning in no-context vs. multi-document context settings. No-context responses may lack recent findings and sometimes hallucinate details. For example, Belokurov et al. (2018) did not use GALAH data. Context enhances performance in summarisation, study comparisons, and idea generation. For example, the model accurately identifies that it only has access to Kawata et al. (2018), as shown in Comparative Analysis III.

Our approach employs the `langchain` (Chase 2022) framework. The distilled input is embedded using OpenAI's `text-embedding-ada-002` model and stored in a large vector store. The expert query and chat history are processed by GPT-4 to generate a standalone question, which is also embedded. We then use FAISS (Facebook AI Similarity Search) (Johnson et al. 2019) to perform a similarity search between the embedded question and the input and retrieve a relevant context for the model. Finally, the GPT-4 model, accessed through the OpenAI API, uses this context and the standalone question to generate a response. The model uses the following **system prompt**: 'Engage in insightful conversations with humans, providing meaningful, concise answers based on the provided documents. Include pertinent study citations, such as Example et al. (2020).'

3. RESULTS AND CONCLUSION

We explore the performance of GPT-4 when responding to expert-level inquiries concerning summarization, comparative analysis, and idea generation, as illustrated in Figure 1. The native model may generate imprecise information through a phenomenon called 'hallucination' (e.g., Shuster et al. 2021; Ji et al. 2022; Peng et al. 2023), or fail to provide answers when presented with recent papers beyond its training data. However, when given access to context, the model's performance significantly improves, as shown by the different responses in the first two left panels of Figure 1 between the native and in-context GPT-4, which we denote by Origins[2]. The in-context prompting allows the model to explore connections and differences across multiple papers, as exemplified by the Comparative Analysis II in Figure 1, where the model correctly identifies the difference in the focus of the two papers provided.

The Comparative Analysis III in Figure 1 demonstrates that the in-context GPT-4 model can identify if it has access to a particular paper, indexed by the first author(s) and year of publication, and then uses that to answer the question. We recognize the potential for idea generation as shown in the lower right panel, with the model identifying a possible new link between machine learning, which was employed to study the Milky Way disc by Ciucă et al. (2023), with the vast available data for globular clusters.

[2] We provide an example of a conversation with Origins, the in-galactic-archaeology-context GPT-4 model, at https://www.youtube.com/watch?v=cufBNDDBgJ4.

To conclude, this research note emphasizes the importance of utilizing in-context prompting with large language models to engage with Astronomy papers effectively. In the future, we plan to investigate the model's response quality as a function of the number of input papers and other variables, such as the focus of the papers and the broadness of the research. On the technical aspect, we plan to explore how the responses vary with the distillation level and how they compare to those from a fine-tuned model.

The authors thank the OpenAI Team for access to the GPT-4 model. IC is grateful for the Joint Jubilee Fellowship at the ANU. YST acknowledges financial support from the Australian Research Council through DECRA Fellowship DE220101520. IC is grateful to Lachezar Todorov, who provided access to the GPT-4 API and to the `langchain` community, in particular Hamel Husain and Mayo Oshin, for their open-source code. IC is grateful to Dr Josh Peek and Dr Kartheik Iyer for their insightful comments. IC and YST thank the Kavli Institute for Theoretical Physics, where this research was undertaken as part of the 'Building a Physical Understanding of Galaxy Evolution with Data-driven Astronomy'. This research was in part supported by the National Science Foundation under Grant No. NSF PHY-1748958.

Software: OpenAI GPT-4 API (OpenAI 2023), `langchain` (Chase 2022), FAISS (Johnson et al. 2019).

REFERENCES

Belokurov, V., Erkal, D., Evans, N. W., Koposov, S. E., & Deason, A. J. 2018, MNRAS, 478, 611, doi: 10.1093/mnras/sty982

Brown, T. B., Mann, B., Ryder, N., et al. 2020, Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., et al. 2023, Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://arxiv.org/abs/2303.12712

Chase, H. 2022, LangChain. https://github.com/hwchase17/langchain

Chowdhery, A., Narang, S., Devlin, J., et al. 2022, arXiv e-prints, arXiv:2204.02311, doi: 10.48550/arXiv.2204.02311

Ciucă, I., Kawata, D., Ting, Y.-S., et al. 2023, MNRAS, doi: 10.1093/mnrasl/slad033

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018, arXiv e-prints, arXiv:1810.04805, doi: 10.48550/arXiv.1810.04805

Helmi, A., Babusiaux, C., Koppelman, H. H., et al. 2018, Nature, 563, 85–88, doi: 10.1038/s41586-018-0625-x

Horta, D., Schiavon, R. P., Mackereth, J. T., et al. 2020, MNRAS, 493, 3363, doi: 10.1093/mnras/staa478

Howard, J., & Ruder, S. 2018, arXiv e-prints, arXiv:1801.06146, doi: 10.48550/arXiv.1801.06146

Hu, E. J., Shen, Y., Wallis, P., et al. 2021, arXiv e-prints, arXiv:2106.09685, doi: 10.48550/arXiv.2106.09685

Ji, Z., Lee, N., Frieske, R., et al. 2022, Survey of Hallucination in Natural Language Generation. https://arxiv.org/abs/2202.03629

Johnson, J., Douze, M., & Jégou, H. 2019, IEEE Transactions on Big Data, 7, 535

Kawata, D., Baba, J., Ciucă, I., et al. 2018, MNRAS, 479, L108, doi: 10.1093/mnrasl/sly107

Massari, D., Koppelman, H. H., & Helmi, A. 2019, A&A, 630, L4, doi: 10.1051/0004-6361/201936135

Myeong, G. C., Evans, N. W., Belokurov, V., Sanders, J. L., & Koposov, S. E. 2018, ApJL, 863, L28, doi: 10.3847/2041-8213/aad7f7

OpenAI. 2023, ArXiv, abs/2303.08774

Peng, B., Galley, M., He, P., et al. 2023, arXiv e-prints, arXiv:2302.12813, doi: 10.48550/arXiv.2302.12813

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. 2020, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Online: Association for Computational Linguistics), 7654–7673, doi: 10.18653/v1/2020.emnlp-main.617

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. 2021, Retrieval Augmentation Reduces Hallucination in Conversation. https://arxiv.org/abs/2104.07567

Wei, J., Bosma, M., Zhao, V. Y., et al. 2022, Finetuned Language Models Are Zero-Shot Learners. https://arxiv.org/abs/2109.01652

White, A. 2023, Paper QA. https://github.com/whitead/paper-qa