




Astronomical Knowledge Entity Extraction in Astrophysics Journal Articles via Large Language Models

Wujun Shao^{1,2,3} , Rui Zhang^{4,7}, Pengli Ji⁴, Dongwei Fan^{1,2,3,5}, Yaohua Hu⁶, Xiaoran Yan⁴, Chenzhou Cui^{1,2,3}, Yihan Tao^{1,2,3}, Linying Mi^{1,2,3}, and Lang Chen^{1,2,3}

¹ National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; shaowj@bao.ac.cn

² University of Chinese Academy of Sciences, Beijing 100049, China

³ National Astronomical Data Center, Beijing 100101, China

⁴ Research Institute of Artificial Intelligence, Zhejiang Lab, Hangzhou 311100, China

⁵ Guilin University, Guangxi 541006, China

⁶ Xidian University, Xi'an 710126, China

Received 2024 January 1; revised 2024 April 2; accepted 2024 April 8; published 2024 May 24

Abstract

Astronomical knowledge entities, such as celestial object identifiers, are crucial for literature retrieval and knowledge graph construction, and other research and applications in the field of astronomy. Traditional methods of extracting knowledge entities from texts face numerous challenging obstacles that are difficult to overcome. Consequently, there is a pressing need for improved methods to efficiently extract them. This study explores the potential of pre-trained Large Language Models (LLMs) to perform astronomical knowledge entity extraction (KEE) task from astrophysical journal articles using prompts. We propose a prompting strategy called Prompt-KEE, which includes five prompt elements, and design eight combination prompts based on them. We select four representative LLMs (Llama-2-70B, GPT-3.5, GPT-4, and Claude 2) and attempt to extract the most typical astronomical knowledge entities, celestial object identifiers and telescope names, from astronomical journal articles using these eight combination prompts. To accommodate their token limitations, we construct two data sets: the full texts and paragraph collections of 30 articles. Leveraging the eight prompts, we test on full texts with GPT-4 and Claude 2, on paragraph collections with all LLMs. The experimental results demonstrate that pre-trained LLMs show significant potential in performing KEE tasks, but their performance varies on the two data sets. Furthermore, we analyze some important factors that influence the performance of LLMs in entity extraction and provide insights for future KEE tasks in astrophysical articles using LLMs. Finally, compared to other methods of KEE, LLMs exhibit strong competitiveness in multiple aspects.

Key words: astronomical databases: miscellaneous – virtual observatory tools – methods: data analysis

1. Introduction

The advent of multi-band and multi-messenger observations marks a new epoch in the field of astronomy. This further attracted more and more scholars to devote themselves to astronomical research, resulting in an increasing cumulative amount of astronomical literature. Renowned repositories such as the Astrophysics Data System (ADS)⁸ and ArXiv⁹ furnish a plethora of astrophysics journal articles, which are replete with invaluable specialized knowledge, including but not limited to, celestial object identifiers, and telescope names (Grezes et al. 2022). These knowledge entities are crucial to the research and application of literature retrieval (Marrero et al. 2013; Yadav & Bethard 2019), text mining, information association and recommendation, knowledge graph construction (Al-Moslmi

et al. 2020; Hogan et al. 2021), publication management, etc. Effectively extracting these knowledge entities from the literature has become one of the keys to improve the efficiency and depth of astronomy research.

Knowledge Entity Extraction (KEE), a subtask of Named Entity Recognition (NER), emphasizes the extraction of professional knowledge entities from texts and outputs them in a structured format, negating the need for sequential entity annotation (Wang et al. 2023). Currently, KEE and NER are extensively studied across various domains. For instance, in the field of biology, this involves extracting information about genes, proteins, and biological processes from texts. In the medical field, it encompasses the identification of symptoms, diagnostic opinions, and drug information. In the field of astronomy, as astronomers' demand for entity information in texts increasingly grows, researchers have also embarked on some exploratory studies.

Early efforts in the field of astronomy have largely depended on rule-based methods (Grishman & Sundheim 1996;

⁷ Corresponding author.

⁸ <https://ui.adsabs.harvard.edu/>

⁹ <https://arxiv.org/archive/astro-ph/>

Cardie 1997; Cucerzan & Yarowsky 1999) and dictionaries (Riloff et al. 1999; Cohen & Sarawagi 2004; Torisawa et al. 2007) for entity extraction. The Detection in Journals of Identifiers and Names (DJIN) system (Lesteven et al. 2010) is one example that uses the *Dictionary of Nomenclature of Celestial Objects* (Lortet et al. 1994) to design more than 50,000 regular expressions for identifying celestial object identifiers and names in articles. This system’s most successful implementation is evidenced at the Strasbourg Astronomical Data Centre (CDS), where it is seamlessly integrated with literature and catalog queries, providing a service within Set of Identifications, Measurements, and Bibliography for Astronomical Data (SIMBAD)¹⁰. This integration facilitates users in navigating through the celestial object identifiers and names mentioned in literature and accessing their detailed data during literature searches; conversely, it enables literature retrieval through celestial object identifiers and names. SIMBAD not only optimizes the process of knowledge acquisition for astronomers but also underscores the pivotal role of KEE in bridging various information resources within the field of astronomy.

With the development of machine learning (; Jordan & Mitchell 2015; Mahesh 2020), statistics-based NER, such as maximum entropy models (Bender et al. 2003; Curran & Clark 2003) and hidden Markov models (Shen et al. 2003; Morwal et al. 2012), have emerged as the prevailing strategy. Murphy et al. (2006) pioneered the development of a specialized entity extraction system tailored for astronomical literature. Specifically, they utilized the maximum entropy model to train an application which is capable of identifying key entities such as source types, source names, and equipment names. By learning from extensive corpora, this approach can glean contextual and distributional information about entities, thereby enhancing recognition performance.

Employing Google’s Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2018) deep neural network architecture, Grezes et al. (2021) have developed a domain-specific model for astronomy, termed astroBERT, through the training on a corpus comprising 395,499 astronomical research papers. Subsequently, this model was used for the development of the NER tool in ADS, which includes identifying specific organizations, projects, terms, etc., in the literature. Moreover, in the evaluation results, astroBERT performed better than the standard BERT model in the KEE tasks on ADS data. Following the success of this work, the Detecting Entities in Astrophysics Literature (DEAL) shared task was proposed at the first Workshop on Information Extraction from Scientific Publications (WIESP)¹¹ at ACL-IJCNLP 2022 (Grezes et al. 2022). The DEAL challenge mandates participants to construct systems capable of

automatically extracting astronomically named entities. Some researchers have attempted to use pre-trained language models such as mT5 (Ghosh et al. 2022) and BERT (Alkan et al. 2022) to extract knowledge entities from text, achieving considerable results.

Recently, Large Language Models (LLMs) with hundreds of billions of parameters, such as GPT-3.5, have demonstrated exceptional zero-shot and few-shot learning capabilities across a multitude of tasks (Li et al. 2023a; Li & Zhang 2023; Wang et al. 2023). Owing to their extensive training samples, these models can rapidly comprehend the rich semantic knowledge embedded in text without the need for large annotated data. Their robust transfer learning capabilities also enable them to swiftly adapt to new domains (Ciucă & Ting 2023; Ciucă et al. 2023; Nguyen et al. 2023). Therefore, LLMs are also actively being applied to KEE tasks by researchers. Sotnikov & Chaikova (2023) harnessed LLMs such as InstructGPT-3 (Ouyang et al. 2022) and Flan-T5-XXL (Chung et al. 2022) for the extraction of astronomical knowledge entities, including event IDs and object names, from Astronomical Telegrams and GCN Circulars. They explored various methods to enhance the capabilities of LLMs, including prompt engineering and model fine-tuning. Their research highlights the potential of LLMs in NER tasks within the field of astronomy.

Owing to the increasing specialization and diversity of astronomical knowledge entities within articles, annotating and training copious samples for each type of entity to develop a functional extraction model is evidently inefficient and unsustainable. Therefore, this paper attempts to explore a more effective method for extracting astronomical knowledge entities. In this paper, we focus on two representative knowledge entities within the field of astronomy: celestial object identifiers and telescope names. We select four mainstream LLMs (Llama-2-70B, GPT-3.5, GPT-4 and Claude 2) and carefully design a new strategy called Prompt-KEE to explore the potential of pre-trained LLMs for KEE in astrophysical articles.

The rest of this paper is structured as follows. In Section 2, we describe the Prompt-KEE strategy and the four LLMs, as well as the set of prompts that we design based on this strategy. In both Section 3 and Section 4, we detail the data set, the design of the combination prompts, the specific experimental procedures, and experimental results and analysis. In Section 5, we introduce some of the main differences between LLMs and other methods. In Section 6, we briefly discuss our work. In Section 7, we conclude this paper.

2. Method

Astrophysical journal articles contain a wide variety of astronomical knowledge entities. Table 1 shows two types of knowledge entities mentioned in sentences from different articles: celestial object identifiers and telescope names. Extracting them from articles is challenging. Inspired by

¹⁰ <https://simbad.u-strasbg.fr/simbad/>

¹¹ <https://ui.adsabs.harvard.edu/WIESP/2022/SharedTasks>

Table 1
Examples of Sentences that Contain Celestial Object Identifiers and Telescope Names

Entity Category	Sentence	Reference
Object Identifier	We performed a detailed chemical analysis for a few objects from this list and showed that the estimated abundances of the CEMP-r/s star LAMOST J151003.74+305407.3 (hereafter J151) could be well explained by the model yields ($[X/Fe]$) of i-process nucleosynthesis of heavy elements, and LAMOST J091608.81+230734.6 (hereafter J091) ...	Purandardas et al. (2022)
	Only a handful of SySts exhibit noticeable signs of such variations in their SEDs (e.g., 2MASS J17391715-3546593, 356.04+03.20, AS 245, H 2-34, PN H 2-5, RT Cru, SMP LMC 88, UV Aur, BI Cru, Hen 2-127, AS 221, Hen 2-139, K 3-9, RR Tel, V347 Nor, V835 Cen, 354.98-02.87).	Akras et al. (2019)
	However, no apparent periods have been detected in the millisecond to second range for either FRB 20121102A or FRB 20201124A, two of the most well-studied repeaters ...	Niu et al. (2022)
Telescope Name	...which was identified from the LAMOST spectrum. The photometric data were collected with the Tsinghua-NAOC 0.8 m telescope (TNT), Transiting Exoplanet Survey Satellite (TESS), Zwicky Transient Facility (ZTF), and ASAS-SN ...	Li et al. (2023b)
	Gaia measurements of G29-38 will build on existing observations with Keck, the Hubble Space Telescope, Herschel, and ALMA ...	Sanderson et al. (2022)
	The first observation for this pulsar was from the Arecibo telescope at 327 and 430 MHz..., Although FAST is the largest and most sensitive radio telescope in the world ...	Shang et al. (2022)

NATURAL-INSTRUCTIONS (Mishra et al. 2021), we propose a prompting strategy, Prompt-KEE, to explore the potential of using general LLMs to extract knowledge entities directly from astrophysical journal articles in a prompt-based method. In this section, we describe the five prompt components of Prompt-KEE, the specific prompts designed based on the Prompt-KEE framework, and four LLMs (Llama-2-70B, GPT-3.5, GPT-4 and Claude 2) used.

2.1. Prompt-KEE

Prompt-KEE is structured as a two-stage conversation process. In the first stage, the prompt comprises four components: Task Descriptions, Entity Definitions, Task Emphasis, and Task Examples. During the second stage, a partial utilization of Task Emphasis is employed specifically for the self-verification of LLMs. We follow the Prompt-KEE strategy to design a set of specific prompts, as shown in Figure 1.

2.1.1. Task Descriptions

We carry out a general design of the Task Description to satisfy the subsequent comparative experiments. First, to fully exploit the astronomical knowledge captured by LLMs, we ask them to take on the role of an experienced astronomer and inform them about the working ability they need to master (Kong et al. 2023). Second, we explicitly specify that the task is to extract astronomical knowledge entities and output the results in a JSON format. Third, we provide three basic requirements. Considering that both the abbreviated and full forms of an astronomical knowledge entity might concurrently appear in an article, and professional scholars exhibit a

preference for using abbreviations in academic writing, therefore, the first prompt asks LLMs to prioritize the extraction of entities in their abbreviated form. Most LLMs struggle with recognizing and processing structured tabular data (Bisercic et al. 2023). Thus, our study skips any tables and figures, and gives LLMs the other prompt, i.e., ignoring any information in tables or figures in the article.

2.1.2. Entity Definitions

LLMs often face challenges in distinguishing highly specialized and detailed terminology, as many similar terms can disperse LLMs' attention and thereby affect final performance (Zhao et al. 2023). To address this issue, we provide definitions for the knowledge entities that we expect to be extracted (celestial object identifiers and telescope names) in an attempt to use these detailed definitions to guide LLMs in the distinction of professional terminology. Specifically, we reference the definitions of celestial object¹² and telescope¹³ from Wikipedia. We also define celestial object identifier and telescope name. Note that celestial objects may have different identifier formats (e.g., Vega, LAMOST J004936.62+375022.8, AS Ser) due to different naming conventions and standards, so we use celestial object identifier as a general term for the convenience of LLMs understanding. Additionally, we do not make a strict distinction between telescope names, other observational facility names, or sky survey names (e.g., LAMOST, Gaia, and SDSS). In fact, astronomers usually do not emphasize the differences among them during scientific research.

¹² https://en.wikipedia.org/wiki/Astronomical_object

¹³ <https://en.wikipedia.org/wiki/Telescope>

<Task Descriptions>

You are an experienced astronomer, capable of easily recognizing knowledge entities ("celestial object names" and "telescope names") in a paragraph of astrophysics paper. Specifically, your task is to perform Knowledge Entity Extraction (KEE) task and meet the following basic requirements: 1) The output should be provided in JSON format. JSON format example: {"Celestial objects": ["XXX"], "Telescopes": ["XXX"]}. 2) Knowledge entities in the paragraph may have both full names and abbreviations, please prioritize abbreviations based on the semantic context. 3) Please do not extract data from the tables and focus on the unstructured textual data.

<Entity Definitions>

A celestial object is a naturally occurring physical entity, association, or structure that exists within the observable universe. A celestial identifier is a unique tag or code used to identify and classify celestial objects. These identifiers typically consist of a combination of letters and numbers, uniquely distinguishing one celestial object from another. A telescope is a device used to observe distant objects by their emission, absorption, or reflection of electromagnetic radiation. Telescope name refers to the unique designation given to a specific telescope.

<Task Emphasis>

1) The paragraph may contain some telescope names represented by their aperture length and address information. 2) The paragraph may also contain some celestial object names in forms such as "several letters or numbers + constellation abbreviation" or "abbreviation of telescope name + coordinate". 3) Don't create entities that's not in the given paragraph. 4) The given paragraph may not contain corresponding knowledge entities. If not exist, do not output such entities. 5) Entities should be verified repeatedly before returning them. 6) Please ensure that all entities from the paragraph have been extracted. 7) After outputting in JSON format, please provide your reasons for selecting these celestial objects and telescope names.

<Task Examples>

1) Example1: Input: In order to study the periods and period variations of CVs, we carried out photometric follow-up observations for several CVs using the SARA RM 1.0 meter telescope, Xinglong 85-cm telescope, and Lijiang 2.4-m telescope. Due to the limiting magnitudes of our telescopes and observing times, we selected five bright CVs as our photometric follow-up objects (UU Aqr, TT Tri, PX And, BP Lyn and RW Tri). Output: {"Celestial objects": ["UU Aqr", "TT Tri", "PX And", "BP Lyn", "RW Tri"], "Telescopes": ["SARA RM 1.0 meter telescope", "Xinglong 85-cm telescope", "Lijiang 2.4-m telescope"]}; 2) Example2: Input: LAMOST spectra of a PN candidate LAMOST J004936.62+375022.8 (upper panel) and a H II region candidate LAMOST J003947.69+402059.1 (bottom panel). Vertical lines with different colors mark the positions of the different emission lines. Output: {"Celestial objects": ["LAMOST J004936.62+375022.8", "LAMOST J003947.69+402059.1"], "Telescopes": ["LAMOST"]}; 3) Example3: Input: Based on this method, many EBs with a third light have been discovered, for instance, AS Ser, AO Ser, KIC 9532219, KIC 5621294, KIC 9007918, MQ UMa, V548 Cyg, EP And and VZ Psc. Output: {"Celestial objects": ["AS Ser", "AO Ser", "KIC 9532219", "KIC 5621294", "KIC 9007918", "MQ Uma", "V548 Cyg", "EP And", "VZ Psc"], "Telescopes": []}

<Second Conversation>

The knowledge entities you extracted may not be complete and accurate, please re-extract them in combination with the extraction result of the previous stage and experience. Emphasis: 1) The paragraph may contain some telescope names represented by their aperture length and address information. 2) The paragraph may also contain some celestial object names in forms such as "several letters or numbers + constellation abbreviation" or "abbreviation of telescope + coordinate". 3) After outputting in JSON format, please provide your reasons for selecting these celestial objects and telescope names.

Figure 1. A set of specific prompts that follow the Prompt-KEE strategy. The sentences from the three provided examples in the Task Examples are cited from Han et al. (2018), Zhang et al. (2020b), and Zhang et al. (2020a).

2.1.3. Task Emphasis

Task Emphasis provides domain knowledge to LLMs for the target entities, and uses prompts to activate the self-improvement capabilities of LLMs. We indicate that telescopes may be named by their aperture length and address information, and celestial objects may be encoded as “several letters or numbers + constellation abbreviation” or “abbreviation of telescope name + coordinate.” We employ self-check prompts to rectify potential errors in the output, thereby encouraging a more profound contextual understanding (Gero et al. 2023). It also contains some other prompts that should not be ignored, which come from our previous practical experience.

2.1.4. Task Examples

Task Examples are designed to enable LLMs to learn the mapping from the inputs to outputs for the celestial objects and telescope names KEE task (Min et al. 2022). Task Examples are meticulously crafted to facilitate the learning process of LLMs in mapping inputs to outputs for celestial object identifiers and telescope names within the KEE task. Specifically, the first example focuses on the form of telescope names, like “address + aperture” format often found in articles. The second one is designed for celestial object identifiers named in “abbreviation of telescope + coordinate” pattern. To reduce the likelihood of LLMs outputting non-existent entities, the third example uses an input without telescope names and an output with no telescope name as guidance.

2.1.5. Second Conversation

We assume that the extracted entities in the first stage may be incomplete and inaccurate. To make up for the errors and omissions of the extraction in the first stage, we provide another stage conversation prompts to ask LLMs to validate the results of the previous conversation and re-extract knowledge entities (Ji 2023).

2.2. LLMs Used

Recent advancements in computational capabilities, combined with the accumulation of extensive textual data sets, have propelled the development of powerful LLMs. Notably, Llama-2-70B, GPT-3.5, GPT-4 and Claude 2 represent cutting-edge systems that have garnered significant attention.

Llama-2-70B,¹⁴ a prominent member of the Llama 2 series (Touvron et al. 2023), is characterized by its expansive 70 billion parameters and a 4096-token (100 tokens = 75 words) context window, making it adept at understanding and processing complex language structures. Its training on a massive 2 trillion token data set significantly enhances its context comprehension, a critical factor in KEE tasks. The

model’s proficiency in external benchmarks, especially in areas requiring deep reasoning and knowledge understanding, positions it as an ideal candidate for testing KEE.

GPT-3.5,¹⁵ an advanced natural language processing model from OpenAI built upon the GPT architecture, is a high-performance iteration that inherits the exceptional language generation and comprehension capabilities of GPT-3¹⁶ while optimizing for faster response and lower cost through parameter efficiency. Through massive pre-training and self-supervised learning, this model has acquired mastery over linguistic patterns and structures, conferring immense application potential across diverse natural language processing tasks including text generation, question answering and KEE.

GPT-4, including its standard 8K and extended 32K token models, has been significantly advanced with the introduction of GPT-4 Turbo, capable of handling a 128K token context.¹⁷ This enhancement makes GPT-4 particularly adept at long-text KEE tasks. The extended token capacity allows for the processing of information equivalent to over 300 pages in a single prompt, thereby enabling more comprehensive analysis and understanding of extensive texts, vital for accurate and in-depth entity extraction from large documents or data sets.

Claude 2,¹⁹ developed by Anthropic, is an LLM with a substantial token limit of 100,000 tokens. This extensive token capacity enables Claude 2 to handle and analyze large volumes of text in a single prompt, making it equally suitable for long-text KEE tasks. The ability to process such a high number of tokens allows Claude 2 to maintain context over lengthy documents, ensuring more accurate and comprehensive extraction of knowledge entities. This capability is crucial for analyzing and understanding extensive data sets or documents, where context and detailed comprehension are crucial.

3. Experiments

In this section, we first introduce the experimental data sets. Then, we describe the specific experimental settings. Finally, we present the evaluation metrics and experimental results.

3.1. Dataset

In order to evaluate the capability of LLMs in performing KEE tasks within astrophysics articles, we have established a data set based on a set of specific selection criteria. Our focus was on selecting articles rich in distinct knowledge entities, such as celestial object identifiers and telescope names, to provide a diverse range of entity samples for the experiment.

¹⁵ <https://openai.com/chatgpt>

¹⁶ <https://openai.com/blog/gpt-3-apps>

¹⁷ <https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models>

¹⁸ <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>

¹⁹ <https://claude.ai/chats>

¹⁴ <https://ai.meta.com/llama/>

We ensured that the research subjects of the articles, covering galaxies, stars, planets, and more, as well as the observational bands like optical, radio, and X-ray, were as broad as possible. This breadth is crucial to fully reflect the diversity and complexity of astrophysical research. In addition, the selected articles needed to be logically structured and content-rich, which is essential for models to comprehend the text context and effectively identify and extract knowledge entities. By adhering to these criteria, our goal was to construct a data set that truly showcases the potential of LLMs in conducting KEE tasks in astrophysical articles.

Therefore, we carefully selected 30 astrophysics journal articles from authoritative publications, including the Astrophysical Journal (ApJ),²⁰ Astrophysical Journal Supplement Series,²¹ Astronomy & Astrophysics (A&A),²² Monthly Notices of the Royal Astronomical Society (MNRAS),²³ and Research in Astronomy and Astrophysics (RAA).²⁴

Despite GPT-4 and Claude 2's unparalleled long-context capabilities allowing them to effortlessly parse texts of the level of astrophysical journal articles, Llama-2-70B and GPT-3.5 currently face strict token limitations. Direct segmentation of lengthy texts based on the maximum token support of models can severely disrupt contextual semantic information, limiting the understanding and reasoning abilities of LLMs. Considering that the maximum token supported by Llama-2-70B and GPT-3.5 is sufficient to cover each paragraph in the articles, we segmented all articles in the order of their paragraphs to ensure maximum preservation of contextual semantic integrity. After processing, these articles were divided into segments ranging from 20 to over 100 paragraphs, forming 30 paragraph collections.

In total, we collected two data sets: the full texts and paragraph collections of the 30 articles. Following the principle of prioritizing abbreviations, we annotated the celestial object identifiers and telescope names that appeared in them. The DOIs of the articles we selected and the specific annotation data are available in the Paperdata Repository of National Astronomical Data Center at <https://nadc.china-vo.org/res/r101358/>.

3.2. Experiment Setup

We designed the comparative experiment from the following aspects. The Figure 2 illustrates our experimental pipeline. The terms Descriptions, Definitions, Emphasis, Examples, and Second Conversation in the figure represent Task Descriptions, Entity Definitions, Task Emphasis, Task Examples, and Second Conversation respectively in the prompt.

First, since we crafted Task Description in a general way, we combine Task Description with other prompt elements to

explore the influence of each prompt element on KEE task. These combinations include: (1) Des_Only: Task Descriptions only. (2) Des_Def: Task Descriptions combined with Entity Definitions. (3) Des_Emp: Task Descriptions combined with Task Emphasis. (4) Des_Exa: Task Descriptions combined with Task Examples. (5) Des_Def_Emp: Task Descriptions, Entity Definitions, and Task Emphasis combined. (6) Des_Def_Emp_Exa: Task Descriptions, Entity Definitions, Task Emphasis, and Task Examples combined. (7) Des_Def_Emp_Con: Task Descriptions, Entity Definitions, Task Emphasis, and Second Conversation combined. (8) All: Task Descriptions, Entity Definitions, Task Emphasis, Task Examples, and Second Conversation combined. Based on our experience, we observed that the definitions of entities and the emphasis on the task usually have a positive and stable impact on the outputs of LLMs. However, the output performance of some current LLMs may exhibit uncertainty when examples are included (Zhao et al. 2021). Therefore, we incorporated Task Examples in combinations (4), (6), and (8) to verify this possibility. Furthermore, Second Conversation is used to re-emphasize the task focus, leading us to accordingly construct combinations (7) and (8).

Second, we fed the full texts of 30 articles, each paired with the eight different combination prompts, into GPT-4 and Claude 2 for KEE. For the 30 paragraph collections of articles, we similarly combined them with these prompts and inputted them into Llama-2-70B, GPT-3.5, GPT-4, and Claude 2. It is important to note that for Llama-2-70B and GPT-3.5, the knowledge entities extracted from the paragraph collections underwent a specific post-processing procedure, which involved merging the results and then removing duplicates.

Finally, all experimental results will be compared with the corresponding annotated knowledge entities of each article.

3.3. Evaluation Metrics

In the realm of KEE task, the evaluation metrics of precision, recall, and F1-score are pivotal in ascertaining the efficacy of models in accurately discerning and extracting pertinent entities from textual data.

Precision: This metric quantifies the accuracy of the model in its entity extraction endeavors. Within the task of KEE, precision is defined as the proportion of accurately identified entities (true positive, TP) relative to the aggregate number of entities (TP + false positive (FP)) extracted by the model. Elevated precision indicates a substantial ratio of correct entity identifications, signifying a reduction in FPs, i.e., erroneous extraction of non-entities or incorrect entities. The formula is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (1)$$

Recall: It assesses the model's ability to extract a comprehensive set of relevant entities. It measures the fraction

²⁰ <https://iopscience.iop.org/journal/0004-637X>

²¹ <https://iopscience.iop.org/journal/0067-0049>

²² <https://www.aanda.org/>

²³ <https://academic.oup.com/mnras/>

²⁴ <https://www.raa-journal.org/>

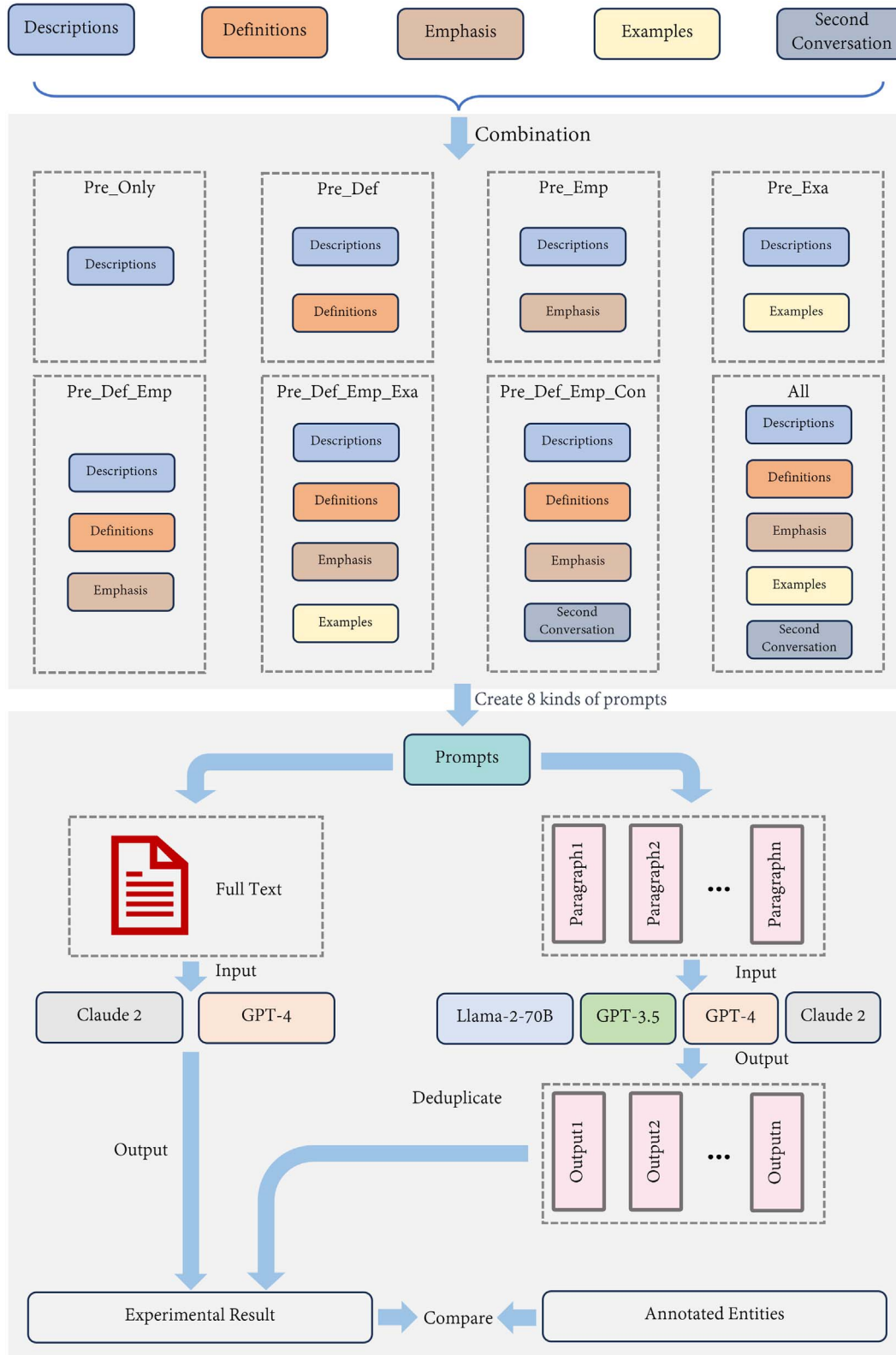
**Figure 2.** The Experimental pipeline.

Table 2

The Results of GPT-4 and Claude 2 in Extracting Celestial Object Identifier and Telescope Name Knowledge Entities from the Full Text of 30 Articles Using Each of the Eight Combination Prompts Individually

Combination Prompt		Celestial Object Identifier			Telescope Name		
		Precision	Recall	F1-score	Precision	Recall	F1-score
GPT-4	Des_Only	0.7913	0.4081	0.5385	0.8112	0.5179	0.6322
	Des_Def	0.7813	0.4484	0.5698	0.8145	0.5449	0.6530
	Des_Emp	0.8118	0.6480	0.7207	0.8540	0.7054	0.7726
	Des_Exa	0.8309	0.5179	0.6381	0.8125	0.5223	0.6359
	Des_Def_Emp	0.8504	0.6502	0.7369	0.8549	0.7366	0.7914
	Des_Def_Emp_Exa	0.8420	0.6569	0.7380	0.8684	0.7411	0.7997
	Des_Def_Emp_Con	0.8713	0.6682	0.7563	0.8763	0.7589	0.8134
	All	0.8739	0.6839	0.7673	0.8769	0.7634	0.8162
Claude 2	Des_Only	0.7456	0.1906	0.3036	0.6951	0.2545	0.3726
	Des_Def	0.6912	0.2108	0.3231	0.7142	0.3571	0.4761
	Des_Emp	0.7083	0.4193	0.5267	0.7952	0.5893	0.6769
	Des_Exa	0.6987	0.3587	0.4703	0.7059	0.3750	0.4898
	Des_Def_Emp	0.7410	0.3722	0.4955	0.6927	0.6741	0.6833
	Des_Def_Emp_Exa	0.7892	0.3946	0.5261	0.6748	0.7411	0.7064
	Des_Def_Emp_Con	0.7500	0.4507	0.5630	0.7255	0.6964	0.7107
	All	0.7955	0.4798	0.5986	0.7652	0.7277	0.7459

of correctly extracted entities (TP) out of the total number of correct entities (TP+FN) that are inherently present and should be extracted from the textual corpus. A model exhibiting high recall is indicative of its proficiency in extracting the majority, if not the entirety, of pertinent entities, thereby minimizing instances of missed identifications. The formula is

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (2)$$

F1-score: This metric, representing the harmonic mean of precision and recall, functions as an integrative measure that encapsulates both accuracy and completeness in the KEE task. It is particularly salient in balancing the model's performance in avoiding entity omissions (high recall) and avoiding inaccurate extractions (high precision). The ideal scenario encompasses a model demonstrating concurrently high precision and recall, although a trade-off between these metrics is often observed in practical applications. The formula is

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (3)$$

4. Results and Analysis

In this section, we comprehensively evaluate the effectiveness of the Prompt-KEE strategy and the extraction capabilities of LLMs. Our analysis is bifurcated into two distinct parts on the full texts and paragraph collections.

4.1. For Full Texts

As shown in Table 2, the results of GPT-4 and Claude 2 in extracting two types of knowledge entities (celestial object

identifier and telescope name) from the full texts of the 30 articles are presented under eight different combination prompts. And in Figure 3, we compare precision, recall, and F1-score from top to bottom respectively. In a comparative analysis of these results, several key insights emerge.

GPT-4 consistently outperforms Claude 2 across all metrics. This superiority is particularly evident in the recall and F1-scores, indicating GPT-4's great capability in accurately identifying a wider range of relevant entities. Claude 2, while competent, shows lower performance, especially in terms of recall.

The diversity of prompts significantly influences the performance of both models. Task Descriptions integrating other elements, especially Task Emphasis, lead to better results compared to simpler ones. GPT-4 shows a pronounced ability to leverage diverse prompts for more accurate entity extraction, with the All prompt (encompassing a combination of all elements) yielding the highest F1-scores. Claude 2 also benefits from more elaborate prompts, but the improvement is less dramatic than with GPT-4. In addition, the overall improvement of the recall based on combination prompts is remarkably greater than that of the precision.

Moreover, a comparison between the extraction of celestial object identifiers and telescope names reveals GPT-4's relatively balanced performance for both entities, while Claude 2 exhibits more variance. GPT-4 slightly favors the extraction of telescope names, whereas Claude 2 shows a notable preference for telescope names, particularly in recall and F1-score.

4.2. For Paragraph Collections

Table 3 provides the results of four models—Llama-2-70B, GPT-3.5, GPT-4, and Claude 2—in extracting celestial object

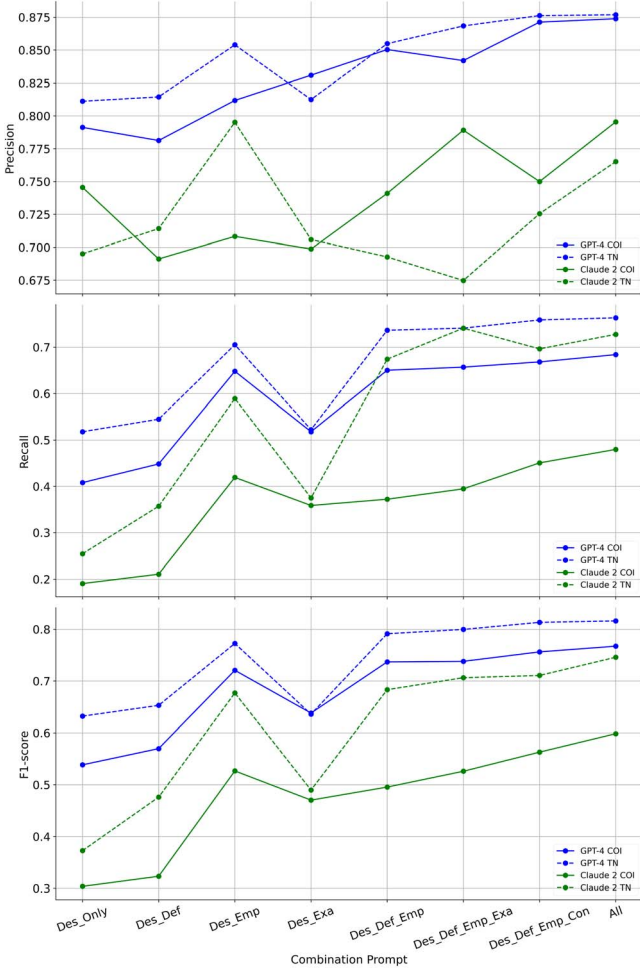


Figure 3. The comparison of precision, recall, and F1-score for extracting celestial object identifiers (COI) and telescope names (TN) in the full texts between GPT-4 and Claude 2.

identifiers and telescope names from the paragraph collections of the articles. And in Figure 4, we compare three metrics.

Llama-2-70B, while showing a respectable recall, particularly in identifying telescope names, falls behind by a wide margin in precision, leading to lower F1-scores. This suggests Llama-2-70B’s tendency to correctly identify a large number of relevant knowledge entities but at the cost of including more FPs. The pattern is consistent across all prompts, indicating a fundamental characteristic of the model’s extraction strategy. Furthermore, the performance of Task Examples in Llama-2-70B validates the possibility that the example introduces uncertainty. This phenomenon arises from the fact that prompts with examples may bias the model’s knowledge assessment, making it tend to prefer knowledge entities contained in the examples when extracting, while reducing the attention to information not present in the examples.

GPT-3.5 demonstrates a marked improvement over Llama-2-70B, with notably higher precision and F1-scores. Its performance peaks with the All prompt, suggesting an ability to effectively utilize prompt information for entity extraction. This trend implies that GPT-3.5 balances accuracy and comprehensiveness better than Llama-2-70B.

Claude 2, while slightly trailing behind GPT-4, shows impressive results. They outperform both Llama-2-70B and GPT-3.5, exhibiting high scores in all metrics for both entity types. Their highest F1-scores are observed with the All prompt, indicating exceptional proficiency in handling complex prompts. Moreover, the results suggest an advanced understanding of the text and a more nuanced extraction capability, making them particularly suitable for tasks requiring high precision and recall.

By analyzing the F1-scores across the above two tables, it is evident that the Prompt-KEE strategy significantly activates the ability of the four LLMs to identify celestial object identifiers and telescope names. In particular, the inclusion of Task Emphasis greatly improves the extraction performance.

We also note that four LLMs are much better at recognizing telescope names than they are at identifying celestial object identifiers. This disparity can be attributed to the fact that telescope names often appear alongside distinctive vocabulary, such as “telescope,” “survey,” etc., which aids in the understanding and judgment of LLMs. Additionally, the number of telescope names is still within a manageable range, possibly already encompassed within the prior knowledge of LLMs. In contrast, celestial object identifiers are diverse in format and vast in quantity, posing a great challenge for LLMs.

Figure 5 compares the three evaluation metrics for extracting object identifiers and telescope names in the full texts and paragraph collections between GPT-4 and Claude 2. We observe distinct disparities in their performance across the texts of two different lengths. Specifically, the former consistently maintains high precision while showing noticeable improvement in recall, whereas the latter demonstrates consistently high recall and precision. We attribute these differences primarily to the following factors:

Contextual Information: The full texts of the 30 journal articles are imbued with a wealth of contextual information, which plays a pivotal role in enabling models like GPT-4 and Claude 2 to accurately comprehend the semantic nuances associated with celestial object identifiers and telescope names, thereby facilitating high precision in entity extraction. While paragraphs inherently provide a narrower context for knowledge entities, the sheer volume and diversity of training data underpinning GPT-4 and Claude 2 equip these models with the capability to maintain notable precision.

Entity Distribution: Knowledge entities within full texts typically exhibit a sparser distribution pattern, whereas entities

Table 3
The Results of Llama-2-70B, GPT-3.5, GPT-4 and Claude 2 in Extracting Celestial Object Identifier and Telescope Name Knowledge Entities from the Paragraph Collections of 30 Articles Using Each of the Eight Combination Prompts Individually

Combination Prompt		Celestial Object Identifier			Telescope Name		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Llama-2-70B	Des_Only	0.0450	0.6687	0.0843	0.1341	0.7098	0.2256
	Des_Def	0.0680	0.6816	0.1237	0.1381	0.7232	0.2319
	Des_Emp	0.1100	0.7085	0.1904	0.1861	0.7768	0.3003
	Des_Exa	0.0320	0.5897	0.0607	0.1121	0.6429	0.1909
	Des_Def_Emp	0.1330	0.7197	0.2245	0.2100	0.7500	0.3281
	Des_Def_Emp_Exa	0.0930	0.6099	0.1614	0.1450	0.7634	0.2437
	Des_Def_Emp_Con	0.1563	0.7197	0.2568	0.1912	0.7723	0.3065
	All	0.1337	0.6300	0.2206	0.1591	0.7188	0.2605
GPT-3.5	Des_Only	0.2902	0.7197	0.4136	0.3605	0.7500	0.4869
	Des_Def	0.3322	0.6928	0.4491	0.3707	0.7232	0.4902
	Des_Emp	0.5105	0.7646	0.6122	0.4101	0.7946	0.5410
	Des_Exa	0.3101	0.7287	0.4351	0.3723	0.7679	0.5015
	Des_Def_Emp	0.5404	0.7803	0.6386	0.5112	0.8125	0.6276
	Des_Def_Emp_Exa	0.5703	0.8094	0.6691	0.5903	0.8170	0.6853
	Des_Def_Emp_Con	0.5505	0.7332	0.6288	0.6111	0.7857	0.6875
	All	0.5906	0.8184	0.6861	0.6301	0.8214	0.7131
GPT-4	Des_Only	0.7804	0.7489	0.7632	0.8026	0.8170	0.8097
	Des_Def	0.8455	0.7242	0.7802	0.8251	0.8214	0.8232
	Des_Emp	0.8474	0.8094	0.8280	0.8414	0.8527	0.8470
	Des_Exa	0.8313	0.7511	0.7892	0.8326	0.8214	0.8270
	Des_Def_Emp	0.8518	0.8117	0.8313	0.8458	0.8571	0.8514
	Des_Def_Emp_Exa	0.8414	0.8206	0.8309	0.8727	0.8571	0.8648
	Des_Def_Emp_Con	0.8535	0.8363	0.8449	0.8744	0.8393	0.8564
	All	0.8536	0.8632	0.8584	0.8694	0.8616	0.8655
Claude 2	Des_Only	0.8208	0.7085	0.7605	0.7702	0.8080	0.7886
	Des_Def	0.8029	0.7399	0.7701	0.7883	0.7813	0.7848
	Des_Emp	0.8005	0.7915	0.7960	0.8210	0.8393	0.8300
	Des_Exa	0.8234	0.7108	0.7630	0.7712	0.8125	0.7913
	Des_Def_Emp	0.8009	0.8206	0.8106	0.8000	0.8571	0.8276
	Des_Def_Emp_Exa	0.8408	0.8408	0.8408	0.8430	0.8393	0.8411
	Des_Def_Emp_Con	0.8518	0.8117	0.8313	0.8514	0.8438	0.8475
	All	0.8444	0.8520	0.8482	0.8319	0.8616	0.8465

in paragraphs tend to be more concentrated. This distribution variance has some influence on the model's recognition capabilities. The sparse knowledge entities embedded within the full texts elevate the likelihood of models overlooking certain entities, consequently diminishing LLMs' recall. Conversely, paragraphs present a denser knowledge entity environment for the models. This attribute allows them to extract more potential knowledge entities, which greatly improves recall rates.

These distinct attributes of full texts and paragraphs greatly affect how the models allocate their attention and process information, leading to varied performances in extracting knowledge entities from the texts of two different lengths within the astrophysical journal articles.

5. Comparison with other Methods

We further compare the LLMs with other extraction methods. These methods are rule-based, machine learning-based, and small-scale pre-trained language model-based methods mentioned in Section 1.

5.1. Comparative Methods

Rule-based Methods: The DJIN system has constructed a robust set of regular expressions based on the *Dictionary of Nomenclature of Celestial Objects* to extract as many celestial object identifiers as possible from the text. Regrettably, the availability of these regular expressions remains a challenge, presenting a significant hurdle in the broader application of the

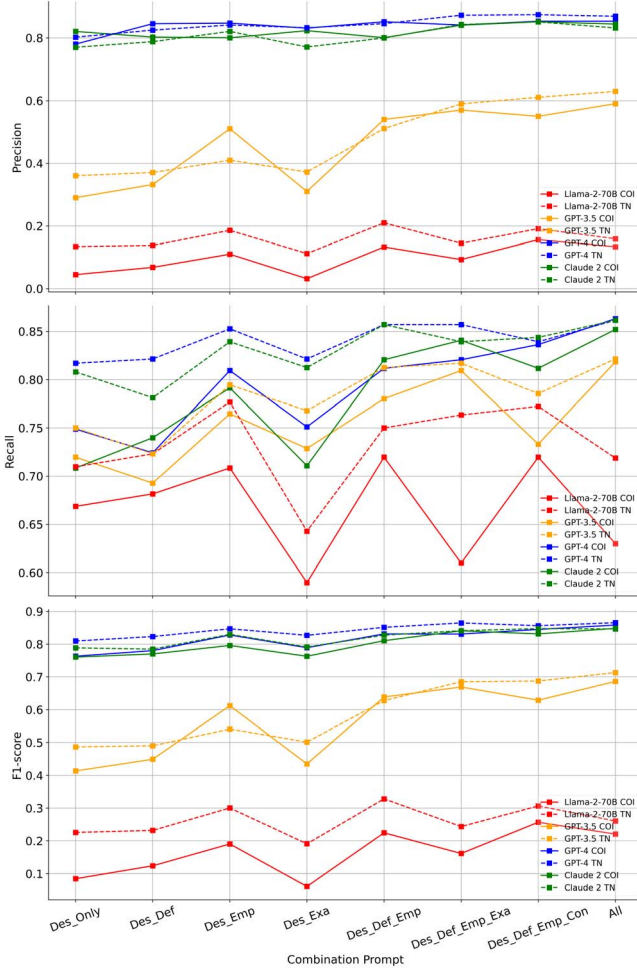


Figure 4. The comparison of precision, recall, and F1-score for extracting celestial object identifiers (COI) and telescope names (TN) in the paragraph collections between Llama-2-70B, GPT-3.5, GPT-4, and Claude 2.

system. Moreover, while the global inventory of telescopes is finite and quantifiable, there is no comprehensive and authoritative list of telescope names available for reference. Therefore, taking inspiration from the development experience of Lesteven et al. (2010), we designed a set of extraction rules for both conventional celestial object identifiers (such as LAMOST J151003.74+305407.3 and NGC 1866) and telescope names (such as Hubble Space Telescope and Arecibo Telescope).

Machine Learning-based Methods: We employed the maximum entropy model (MaxEnt), a prevalent method for information extraction based on machine learning techniques. The application of this model for the extraction of astronomical knowledge entities necessitates an ample supply of high-quality, annotated astronomical data. Consequently, we have selected the DEAL data set as our training corpus, which encompasses annotations for both telescope names and celestial

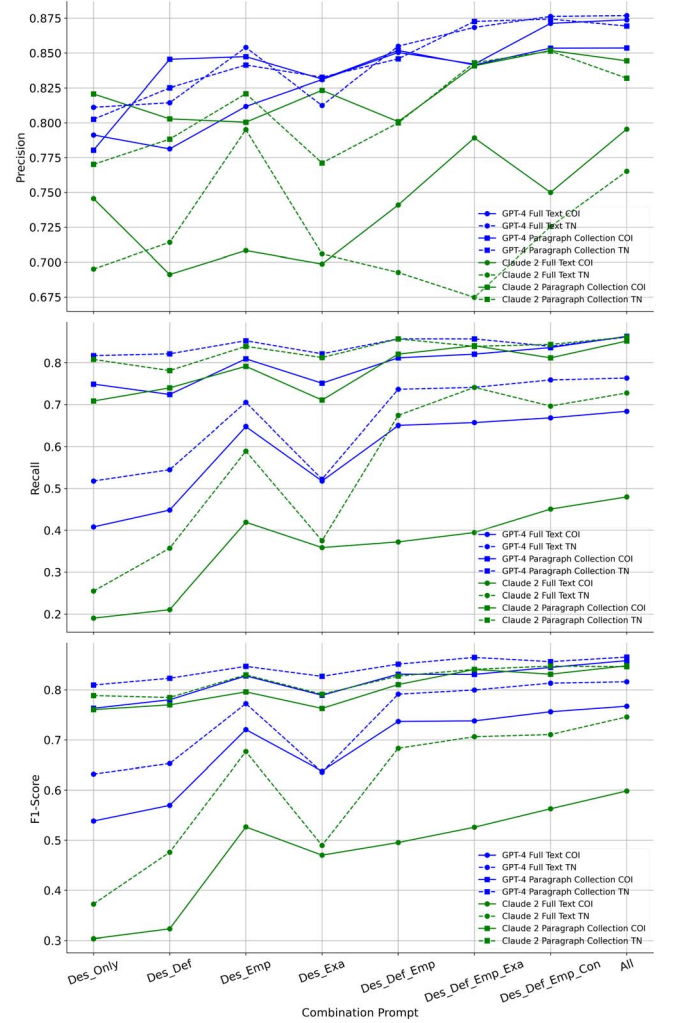


Figure 5. The comparison of precision, recall, and F1-score for extracting celestial object identifiers (COI) and telescope names (TN) in the full texts and paragraph collections between GPT-4 and Claude 2.

object identifiers. It is important to note that within this data set, the term “CelestialObject” corresponds to the “celestial object identifier” in this paper, while “Telescope” and “Survey” are aligned with the “telescope name” category. Detailed information regarding this data set can be accessed at <https://ui.adsabs.harvard.edu/WIESP/2022/LabelDefinitions> and <https://ui.adsabs.harvard.edu/blog/ads-models-and-data-sets>, and it is publicly available at <https://huggingface.co/data-sets/adsabs/WIESP2022-NER>. We implemented this task in Python using the MaxentClassifier class and related functions from the nltk library.²⁵

Smaller Language Model-based Methods: Building upon the work of Alkan et al. (2022), we opted for the

²⁵ <https://www.nltk.org/>

Table 4
The Performance Comparison of Four LLMs and other Methods in Extracting Celestial Object Identifier and Telescope Name Knowledge Entities

Method	Celestial Object Identifier			Telescope Name		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Llama-2-70B	0.1563	0.7179	0.2568	0.2100	0.7500	0.3281
GPT-3.5	0.5906	0.8184	0.6861	0.6301	0.8214	0.7131
GPT-4	0.8536	0.8632	0.8584	0.8694	0.8616	0.8655
Claude 2	0.8444	0.8520	0.8482	0.8514	0.8438	0.8475
Rule	0.5185	0.1211	0.1963	0.8529	0.1518	0.2577
MaxEnt	0.2817	0.1592	0.2034	0.2898	0.3080	0.2986
SciBERT	0.4624	0.3879	0.4218	0.5517	0.6473	0.5956

scibert_scivocab_cased version²⁶ of the SciBERT model (Beltagy et al. 2019) from PyTorch HuggingFace. This version is particularly adept at processing scientific texts, as it offers enhanced recognition and comprehension of scientific terminology and concepts. Through fine-tuning, the model can be further optimized to cater to specific tasks of entity recognition within scientific literature. We have continued to utilize the DEAL data set (Grezes et al. 2022) and implemented a sliding window strategy to accommodate the maximum sequence length constraint of the BERT model, which is typically 512 tokens. Additionally, we have set a training regimen of 30 epochs to ensure that the model thoroughly learns the characteristics of the data set.

5.2. Comparison

To ensure that the extraction performance of these methods can be compared with that of four LLMs, we employed the paragraph collections as the experimental text. For these LLMs, we selected their experimental results from their respective combination prompts that yielded the highest F1-score in Table 3 as the benchmarks for comparison. Table 4 shows their performance. The following outlines the main differences between LLMs and these methods.

Performance: From the F1-scores presented in Table 3, it is evident that the overall performance of rule-based method and maximum entropy model is significantly lower than that of LLMs. In contrast, SciBERT, which has been pre-trained on a substantial corpus of scientific literature and fine-tuned using the DEAL data set, exhibits superior performance compared to the Llama-2-70B without fine-tuning. Specifically, SciBERT outperforms Llama-2-70B by 0.1650 and 0.2675 in the extraction of celestial object identifiers and telescope names, respectively. In Section 5.1, we delineate some objective limitations; it must be acknowledged that the performance of these methods has not reached the pinnacle, equating to that of mature astronomical KEE methods such as the DJIN system. Hence, we do not extensively focus on the performance

differences between LLMs and these methods. Instead, our focus is primarily on the key differences during the extraction process. Table 5 shows these differences.

1. The comparative analysis of the six examples in Table 5 reveals the LLMs exhibit a superior capacity for comprehensively identifying celestial object identifiers and telescope names, two types of knowledge entities, when compared to other methods.
2. From all the examples, it can be observed that LLMs perform better in handling knowledge entity boundaries, such as compound nouns or abbreviations. For instance, “Small Magellanic Cloud” is a compound noun referring to a specific galaxy. Rule-based methods require specific rules to identify such structures, and both the maximum entropy model and SciBERT may encounter difficulties in handling these compound nouns. On the contrary, LLMs are capable of recognizing “Small Magellanic Cloud” as a whole entity, even when it spans across multiple words.
3. The examples presented in Table 5, particularly Example 3 and 4, demonstrate the better generalization capabilities of LLMs. Despite the presence of entities in the text that are either uncommon in the pre-training data or entirely novel, LLMs are adept at inferring the meaning and category of unknown entities by capturing semantic relationships and contextual information between words. This enables them to perform more consistently when confronted with new texts. In contrast, other methodologies struggle to adapt to such situations.
4. LLMs demonstrate strong competitiveness in entity disambiguation as well. In Example 5, both “Andromeda” and “M31” refer to the same galaxy. LLMs are capable of understanding the equivalence between the two entities and correctly associating them with the same celestial object, outputting it as “Andromeda galaxy (M31).” In Example 6, despite the presence of several personal names (Page, Grupe, and Giommi) and telescopes named after individuals (XMM-Newton, Swift, Fermi, Planck), all four LLMs are still able to accurately distinguish them.

²⁶ <https://github.com/allenai/scibert>

Table 5
Comparison Examples of LLMs and other Methods in Extracting Celestial Object Identifiers and Telescope Names

Example	Sentence	Method	Output
1	...indicated that the reddening of the quasar is steeper than in the Small Magellanic Cloud, and perhaps even steeper than for the galaxy IRAS 14026+4341. (Marculewicz et al. 2022)	Expected	[Small Magellanic Cloud, IRAS 14026+4341]
		Llama-2-70B	[Small Magellanic Cloud, IRAS 14026+4341]
		GPT-3.5	[Small Magellanic Cloud, IRAS 14026+4341]
		GPT-4	[Small Magellanic Cloud, IRAS 14026+4341]
		Claude 2	[Small Magellanic Cloud, IRAS 14026+4341]
		Rule	[IRAS 14026+4341]
		MaxEnt	[IRAS 14026+4341]
		SciBERT	[Magellanic Cloud, IRAS 14026+4341]
2	Compared to the limits placed by Scholz et al. (2017) on X-ray emission at the time of radio bursts from FRB 121102 using Chandra and XMM, the limits placed here using NuSTAR are not as constraining for the low absorption ...(Cruces et al. 2021)	Expected	[Chandra, XMM, NuSTAR]
		Llama-2-70B	[Chandra, XMM, NuSTAR]
		GPT-3.5	[Chandra, XMM, NuSTAR]
		GPT-4	[Chandra, XMM, NuSTAR]
		Claude 2	[Chandra, XMM, NuSTAR]
		Rule	[Chandra]
		MaxEnt	[Chandra]
		SciBERT	[Chandra, XMM]
3	For J1334, the primary component is near the 0.04 Gyr isochrone and the secondary component is not far below the 10 Gyr isochrone. (Lu et al. 2018)	Expected	[J1334]
		Llama-2-70B	[J1334]
		GPT-3.5	[J1334]
		GPT-4	[J1334]
		Claude 2	[J1334]
		Rule	[]
		MaxEnt	[]
		SciBERT	[]
4	The first time was on 2014 November 15 and 16. We used the 60 cm reflecting telescope to perform <i>R</i> -band photometry. The second time was on 2019 January 17 and 18. (Lu et al. 2020)	Expected	[60 cm reflecting telescope]
		Llama-2-70B	[60 cm reflecting telescope]
		GPT-3.5	[60 cm reflecting telescope]
		GPT-4	[60 cm reflecting telescope]
		Claude 2	[60 cm reflecting telescope]

Table 5
(Continued)

Example	Sentence	Method	Output
5	In this Letter, we report the discovery of a new LBV - LAMOST J0037+4016 (R.A.: 00:37:20.65, decl.: +40:16:37.70) in the Andromeda galaxy (M31). (Huang et al. 2019)	Rule	[cm reflecting telescope]
		MaxEnt	[reflecting telescope]
		SciBERT	[60 cm reflecting telescope]
		Expected	[LAMOST J0037+4016, Andromeda galaxy(M31)]
		Llama-2-70B	[LAMOST J0037+4016, Andromeda galaxy(M31)]
		GPT-3.5	[LAMOST J0037+4016, Andromeda galaxy(M31)]
		GPT-4	[LAMOST J0037+4016, Andromeda galaxy(M31)]
		Claude 2	[LAMOST J0037+4016, Andromeda galaxy(M31)]
		Rule	[LAMOST J0037+4016, M31]
		MaxEnt	[LAMOST J0037+4016, M31]
6	...XMM-Newton (Page et al. 2004), and Swift (Grupe et al. 2010) ... Giommi et al. (2012) did not detect LAMOST J1131+3114 in γ -ray or submillimeter ranges using Fermi and Planck. (Shi et al. 2014)	Expected	[XMM-Newton, Swift, Fermi, Planck]
		Llama-2-70B	[XMM-Newton, Swift, Fermi, Planck]
		GPT-3.5	[XMM-Newton, Swift, Fermi, Planck]
		GPT-4	[XMM-Newton, Swift, Fermi, Planck]
		Claude 2	[XMM-Newton, Swift, Fermi, Planck]
		Rule	[LAMOST]
		MaxEnt	[Swift, LAMOST, Fermi, Planck]
		SciBERT	[XMM-Newton, Swift, Giommi, Fermi, Planck]

Furthermore, the telescope name “LAMOST” included in the celestial object identifier “LAMOST J1131+3114” is correctly excluded from “LAMOST” and “J1131+3114” constitutes a single entity without conveying the meaning of “telescope.” However, other methods seem to be less satisfactory in these aspects.

Extraction Pattern: The performance differences in extracting entities are closely related to the respective working patterns of these methods. LLMs can learn a vast amount of general language knowledge during pre-training, thus requiring minimal external knowledge for identifying celestial identifiers and telescope names. They can accomplish KEE tasks with just a few prompts. In contrast, a rule-based entity extraction

method heavily relies on pre-defined sets of rules for pattern matching, resulting in limited capability to handle unknown or complex entities (such as “Small Magellanic Cloud”). Maximum entropy model and SciBERT seem to alleviate these issues, but they require extensive high-quality astronomical text data sets to enhance their ability to extract celestial object identifiers and telescope names. The labor-intensive and tedious nature of these tasks makes them less competitive compared to LLMs.

Update and Maintenance: For the practical application of astronomical entity extraction tools based on these methods, updating and maintenance become necessary tasks. The rule-based method and maximum entropy model require frequent

updates to adapt to new astronomical terms and terminology, ensuring sustained accuracy. This process can be labor-intensive. SciBERT benefits from less frequent updates due to its general scientific pre-training but still needs occasional fine-tuning. In contrast, LLMs are typically updated and maintained by professional companies, providing a more sustainable and scalable solution for entity extraction in astronomy. The remarkable capabilities of LLMs reduce the need for ongoing maintenance, making them a more efficient choice for accomplishing the task of extracting astronomical knowledge entities.

6. Discussion

This study validates the effectiveness of our carefully designed Prompt-KEE and highlights the potential of pre-trained LLMs for KEE in astrophysical journal articles. This further underscores the importance of prompts as a viable strategy to enable models to rapidly adapt to new domains and tasks. Furthermore, compared to other methods of extracting knowledge entities, these LLMs demonstrate stronger competitiveness, such as their robust generalization capabilities. However, it is essential to acknowledge the limitations of this study. First, we solely focused on two typical categories of astronomical knowledge entities, and the recognition of more complex and specialized astronomical entity types necessitates further investigation. Second, different models and astronomical task scenarios may require different prompt optimization strategies, and research in this area still offers extensive exploration opportunities. Third, due to certain challenges, the alternative methods used for comparison, while exhibiting some performance, have not reached the level comparable to mature systems.

Meanwhile, we also recognize several issues that warrant further investigation:

1. Domain knowledge is key to augmenting the understanding, reasoning, and generalization abilities of LLMs; thus, training LLMs with extensive astronomical knowledge can further enhance their capability in extracting astronomical knowledge entities.
2. The inclusion of examples may impact the model's extraction of new entities, suggesting that prompts need careful design based on specific models and scenarios. More information does not necessarily lead to better results, hence, further optimization of the Prompt-KEE strategy is possible.
3. Differences in extraction results between full texts and paragraph collections indicate that contextual information and data set characteristics influence model attention allocation. Investigating how to guide the model to perform better across the texts of two different lengths merits further research.
4. Astrophysical journal articles often present specialized knowledge entities such as celestial object identifiers in structured tables. Existing models exhibit limited capabilities in extracting structured table data from articles. Future research can address this issue specifically.

7. Conclusion

In this paper, we proposed the Prompt-KEE strategy to explore the potential of pre-trained LLMs for KEE in astrophysical journal articles. We focused on the two most typical astronomical knowledge entities, celestial object identifier and telescope name. Based on the Prompt-KEE strategy, we designed eight combination prompts with five elements: Task Descriptions, Entity Definitions, Task Emphasis, Task Examples, and Second Conversation. Furthermore, we collected two data sets: the full texts and paragraph collections of the 30 articles, and employ four LLMs (Llama-2-70B, GPT-3.5, GPT-4 and Claude 2) for our experiments. Leveraging the eight combination prompts, we tested on full texts with GPT-4 and Claude 2, on paragraph collections with all LLMs. The experimental results demonstrated that pre-trained LLMs can perform KEE tasks in the astrophysics journal articles, but there are differences in their performance. Moreover, we introduced areas that require further exploration and improvement, including the design of prompts and the use of contextual information. Finally, we compared LLMs to other methods, showing the advantages of LLMs in terms of performance, as well as their competitiveness in terms of working patterns, updates and maintenance. This study provides valuable insights for using prompt engineering to adapt LLMs for KEE tasks in astrophysical articles.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC, Grant Nos. 12273077, 72101068, 12373110, and 12103070), National Key Research and Development Program of China under grants (2022YFF0712400, 2022YFF0711500), and the 14th Five-year Informatization Plan of Chinese Academy of Sciences (CAS-WX2021SF-0204). Data resources are supported by China National Astronomical Data Center (NADC), CAS Astronomical Data Center and Chinese Virtual Observatory (China-VO). This work is supported by Astronomical Big Data Joint Research Center, co-founded by National Astronomical Observatories, Chinese Academy of Sciences and Alibaba Cloud.

This research has made use of NASA's Astrophysics Data System Bibliographic Services and SIMBAD database, operated at CDS, Strasbourg, France. Data Publishing is supported by China National Astronomical Data Center (NADC) through Chinese Virtual Observatory (China-VO) PaperData Repository.

References

- Akras, S., Guzman-Ramirez, L., Leal-Ferreira, M. L., & Ramos-Larios, G. 2019, *ApJS*, **240**, 21
- Alkan, A. K., Grouin, C., Schüssler, F., & Zweigenbaum, P. 2022, in First Workshop on Information Extraction from Scientific Publications (Taiwan: Association for Computational Linguistics), 145
- Al-Moslimi, T., Ocaña, M. G., Opdahl, A. L., & Veres, C. 2020, *IEEEA*, **8**, 32862
- Beltagy, I., Lo, K., & Cohan, A. 2019, arXiv:1903.10676
- Bender, O., Och, F. J., & Ney, H. 2003, in Proc. Seventh Conf. Natural Language Learning at HLT-NAACL 2003 (Edmonton: Association for Computational Linguistics), 148
- Bisercic, A., Nikolic, M., van der Schaar, M., et al. 2023, arXiv:2306.05052
- Cardie, C. 1997, *AI Mag.*, **18**, 65
- Chung, H. W., Hou, L., Longpre, S., et al. 2022, arXiv:2210.11416
- Ciucă, I., & Ting, Y.-S. 2023, *RNAAS*, **7**, 193
- Ciucă, I., Ting, Y.-S., Kruk, S., & Iyer, K. 2023, arXiv:2306.11648
- Cohen, W. W., & Sarawagi, S. 2004, in Proc. Tenth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (Seattle WA: Association for Computing Machinery), 89
- Cruces, M., Spitler, L., Scholz, P., et al. 2021, *MNRAS*, **500**, 448
- Cucerzan, S., & Yarowsky, D. 1999, in 1999 Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora (College Park, MD: Association for Computational Linguistics), 90
- Curran, J. R., & Clark, S. 2003, in Proc. Seventh Conf. Natural Language Learning at HLT-NAACL 2003 (Edmonton: Association for Computational Linguistics), 164
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018, arXiv:1810.04805
- Gero, Z., Singh, C., Cheng, H., et al. 2023, arXiv:2306.00024
- Ghosh, M., Santra, P., Iqbal, S. A., & Basuchowdhuri, P. 2022, in Proc. First Workshop on Information Extraction from Scientific Publications (Taiwan: Association for Computational Linguistics), 100
- Giommi, P., Polenta, G., Lähteenmäki, A., et al. 2012, *A&A*, **541**, A160
- Grezes, F., Blanco-Cuaresma, S., Accomazzi, A., et al. 2021, arXiv:2112.00590
- Grezes, F., Blanco-Cuaresma, S., Allen, T., & Ghosal, T. 2022, in Proc. First Workshop on Information Extraction from Scientific Publications (Taiwan: Association for Computational Linguistics), 1
- Grishman, R., & Sundheim, B. M. 1996, in COLING 1996 Volume 1: The 16th Int. Conf. Computational Linguistics (Copenhagen: Association for Computational Linguistics), 466
- Grupe, D., Komossa, S., Leighly, K. M., & Page, K. L. 2010, *ApJS*, **187**, 64-106
- Han, X. L., Zhang, L.-Y., Shi, J.-R., et al. 2018, *RAA*, **18**, 068
- Hogan, A., Blomqvist, E., Cochez, M., et al. 2021, *ACM Computing Surveys (CSur)*, **54**, 1
- Huang, Y., Zhang, H.-W., Wang, C., et al. 2019, *ApJL*, **884**, L7
- Ji, B. 2023, arXiv:2305.03253
- Jordan, M. I., & Mitchell, T. M. 2015, *Sci*, **349**, 255
- Kong, A., Zhao, S., Chen, H., et al. 2023, arXiv:2308.07702
- Lesteven, S., Bonnin, C., Derriere, S., et al. 2010, Library and Information Services in Astronomy VI: 21st Century Astronomy Librarianship, From New Ideas to Action, **433**, 317
- Li, J., Li, H., Pan, Z., & Pan, G. 2023a, arXiv:2305.12212
- Li, M., & Zhang, R. 2023, arXiv:2307.00186
- Li, X., Wang, X., Liu, J., et al. 2023b, arXiv:2306.07529
- Lortet, M.-C., Borde, S., & Ochsenbein, F. 1994, *A&AS*, **107**, 193
- Lu, H.-p., Michel, R., Zhang, L.-y., & Castro, A. 2018, *AJ*, **156**, 88
- Lu, H.-p., Zhang, L.-y., Michel, R., & Han, X. L. 2020, *ApJ*, **901**, 169
- Maresh, B. 2020, *International Journal of Science and Research (IJSR)*, **9**, 381
- Marculewicz, M., Nikolajuk, M., & Różańska, A. 2022, *A&A*, **668**, A128
- Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbis, J. M. 2013, *Comput. Stand. Interfaces*, **35**, 482
- Min, S., Lyu, X., Holtzman, A., et al. 2022, arXiv:2202.12837
- Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. 2021, arXiv:2104.08773
- Morwal, S., Jahan, N., & Chopra, D. 2012, International Journal on Natural Language Computing (IJNLC), **1** (4) 15
- Murphy, T., McIntosh, T., & Curran, J. R. 2006, in Proc. Australasian Language Technology Workshop 2006 (Sydney: ALTW Press), 59
- Nguyen, T. D., Ting, Y.-S., Ciucă, I., et al. 2023, arXiv:2309.06126
- Niu, J.-R., Zhu, W.-W., Zhang, B., et al. 2022, *RAA*, **22**, 124004
- Ouyang, L., Wu, J., Jiang, X., et al. 2022, *Advances in Neural Information Processing Systems*, **35**, 27730
- Page, K. L., Scharrel, N., Turner, M. J. L., & O'Brien, P. T. 2004, *MNRAS*, **352**, 523-534
- Purandardas, M., Goswami, A., Shejeelammal, J., et al. 2022, *MNRAS*, **513**, 4696
- Riloff, E., & Jones, R. 1999, *AAAI/IAAI*, **474**
- Sanderson, H., Bonsor, A., & Mustill, A. 2022, *MNRAS*, **517**, 5835
- Shang, L.-H., Bai, J.-T., Dang, S.-J., & Zhi, Q.-J. 2022, *RAA*, **22**, 025018
- Shen, D., Zhang, J., Zhou, G., Su, J., & Tan, C. L. 2003, in Proc. ACL 2003 Workshop on Natural Language Processing in Biomedicine (Sapporo: Association for Computational Linguistics), 49
- Shi, Z., Comte, G., Luo, A. L., et al. 2014, *A&A*, **564**, A89
- Sotnikov, V., & Chaikova, A. 2023, *Galaxies*, **11**, 63
- Torisawa, K. 2007, in Proc. 2007 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (Prague: Association for Computational Linguistics), 698
- Touvron, H., Martin, L., Stone, K., et al. 2023, arXiv:2307.09288
- Wang, S., Sun, X., Li, X., et al. 2023, arXiv:2304.10428
- Yadav, V., & Bethard, S. 2019, arXiv:1910.11470
- Zhang, B., Qian, S.-B., Wang, J.-J., et al. 2020a, *RAA*, **20**, 047
- Zhang, M., Chen, B.-Q., Huo, Z.-Y., et al. 2020b, *RAA*, **20**, 097
- Zhao, X., Lu, J., Deng, C., et al. 2023, arXiv:2305.18703
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. 2021, *ICML, PMLR*, **139**, 12697