中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于大语言模型的天文文献知识实体抽取方法研究

作者姓名：邵务俊

指导教师：樊东卫 副研究员 中国科学院国家天文台

学位类别：理学硕士

学科专业：天文技术与方法

培养单位：中国科学院国家天文台

2024 年 6 月

Research on Method of Knowledge Entity Extraction in Astronomical Literature Based on Large Language Models

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Natural Science

in Astronomical Technology and Method

By

Shao Wujun

Supervisor: Associate Professor Fan Dongwei

National Astronomical Observatories, Chinese Academy of Sciences

June, 2024

中国科学院大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：邵尧俊

日　　期：2024年5月31日

中国科学院大学

学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：邵尧俊　　导师签名：樊东卫

日　　期：2024年5月31日　　日　　期：2024年5月31日

摘 要

随着新一代高分辨率望远镜的投入使用以及一系列巡天计划的推进，天文学领域获取到了前所未有的大规模数据。天文数据的不断涌现促进了天文文献数量的持续攀升，这些文献是科研人员开展新研究不可或缺的资源。然而，目前天文数据与文献之间关联程度较低，给天文研究人员搜集天体相关信息带来了很多不便。诸如天体标识符、望远镜名称等天文知识实体作为天文数据与文献的关键纽带，是实现天文数据与文献关联融合的基本要素。准确、快速抽取天文知识实体对于天文学研究具有重要意义。传统知识实体抽取方法在处理大规模、复杂天文文献时存在诸多局限性，例如处理周期长、识别实体边界困难、泛化能力差等。而最近大语言模型的出现，为诸多领域自然语言处理任务带来了新的机遇。针对天文文献中复杂且多样的实体，本文探索如何利用大语言模型克服传统抽取方法所面临的挑战，更有效地完成文献的天文知识实体抽取任务，以实现天文数据与文献之间快速的关联融合，主要工作如下：

Prompt-KEE 提示策略的设计与应用。本文提出了一种名为 Prompt-KEE 的新型提示策略，其中包含任务描述、实体定义、任务强调、任务示例和二轮对话五个提示要素，用于指导大语言模型更有效地执行天文知识实体抽取任务。这一策略通过细致的提示要素设计，激活大语言模型在天文领域的潜在能力，减少对大规模标注数据的依赖，从而加快实体抽取工作。

基于大语言模型的天文文献知识实体抽取研究。本文尝试使用大语言模型来抽取天文知识实体，因此本文选择了四种具有代表性的大语言模型（Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2）。为了适应这些模型的 token 限制，本文构建了两种数据集，即 30 篇文献的全文文本数据集以及这些文献的段落文本集合数据集。通过 Prompt-KEE 策略，本文设计了八种组合提示，并对这些模型在不同提示组合以及不同数据集下的性能进行了测试实验和评估。实验结果表明，Prompt-KEE 策略能够有效提升大语言模型抽取天文知识实体的效果，并且大语言模型在执行天文知识实体抽取任务时表现出了显著的能力。

其它实体抽取方法与大语言模型方法的对比分析。为了进一步探索大语言模型在抽取任务上的各方面优势，本文还将基于规则、基于机器学习和基于小规模预训练语言模型的方法应用到天文知识实体抽取任务中，并且将其实验结果与大语言模型的实验结果进行了详细的对比。对比结果表明，大语言模型在处理复杂实体边界、实体消歧以及泛化能力等诸多方面存在显著优势。

本文通过系统地研究大语言模型在天文文献知识实体抽取任务中的应用，不仅为天文知识实体抽取研究提供了有益的参考和启示，也为提升天文数据与文献关联的工作效率做了铺垫。

关键词：天文文献，数据融合，关联检索，实体抽取，大语言模型，提示策略

Abstract

With the advent of new high-resolution telescopes and a series of sky surveys, the field of astronomy has seen an unprecedented surge in data. The continuous emergence of astronomical data has promoted the rapid increase in the astronomical literature, which is an indispensable resource for researchers to carry out new research. However, currently, there is a low level of correlation between astronomical data and literature, which poses significant challenges for astronomers in gathering relevant celestial information. Astronomical knowledge entities such as celestial object identifiers and telescope names serve as crucial links between astronomical data and literature, and they are fundamental elements for achieving the correlation and retrieval between astronomical data and literature. Accurately and quickly extracting astronomical knowledge entities is of great significance to astronomical research. Traditional knowledge entity extraction method have many limitations when dealing with large-scale and complex astronomical literature. The recent emergence of large language models has brought new opportunities to complex natural language processing tasks in various fields. Facing the complex and diverse entities in astronomical literature, this thesis explores how to leverage large language models to overcome the challenges faced by traditional extraction methods and more effectively complete the task of astronomical knowledge entity extraction from literature, in order to achieve a rapid association and integration between astronomical data and literature. The main contributions of this thesis are as follows:

Design and application of prompt-KEE prompting strategy. This thesis proposes a novel prompting strategy called Prompt-KEE, which consists of five prompting elements: Task Description, Entity Definition, Task Emphasis, Task Examples, and Second Conversation. These elements are designed to guide large language models in performing astronomical knowledge entity extraction tasks more effectively. This strategy, through meticulous prompt design, activates the latent capabilities of large language models in the field of astronomy, reducing the reliance on large-scale annotated data and thus accelerating entity extraction tasks.

Research on knowledge entity extraction in astronomy literature based on large language models. This thesis attempts to use large language models to extract astronomical knowledge entities, and therefore, four representative large language models (Llama-2-70B, GPT-3.5, GPT-4, and Claude 2) are selected. To accommodate the token limitations of these models, two datasets were constructed: a full-text dataset of 30 literature and a paragraph collection dataset of these literature. Based on Prompt-KEE strategy, eight kinds of combination prompts are designed, and the performance of these models under different combination prompts and different data sets is tested

and evaluated. The experimental results demonstrate that the Prompt-KEE strategy effectively improves the performance of large language models in extracting astronomical knowledge entities, and large language models exhibit significant ability in performing astronomical knowledge entity extraction tasks.

Comparison analysis between other entity extraction methods and large language model methods. To further explore the advantages of large language models in the extraction task, this thesis also applys rule-based, machine learning-based, and small-scale pre-trained language model methods to the task of astronomical knowledge entity extraction and compared their experimental results with those of large language models. The comparison results show that large language models have significant advantages in multiple aspects, such as dealing with complex entity boundaries, entity disambiguation, and generalization capabilities.

By systematically studying the application of the large language model in the task of extracting astronomical knowledge entities, this thesis not only provides a useful reference and inspiration for the research of astronomical knowledge entities extraction, but also lays the groundwork for improving the efficiency of correlating astronomical data with literature.

Key Words: Astronomical Literature, Data Fusion, Correlation and Retrieval, Entity Extraction, Large Language Model, Prompting Strategy

目 录

图目录

表目录

第 1 章　绪论

1.1　研究背景及意义

在信息时代的浪潮中，天文学这一古老的科学正在经历一场由数据驱动的革命。自伽利略将世界首台望远镜指向星空，揭开了天文观测的新篇章以来，人类对宇宙的探索已经从地面延伸到太空，从单一的光学波段拓展到了包括射电、红外、X 射线等在内的整个电磁波段，乃至引力波、中微子及宇宙线。同时，随着新一代高分辨率望远镜的投入使用和一系列雄心勃勃的巡天计划的推进，人类获取到了前所未有的大规模天文数据，涵盖了光谱数据、测光数据等众多不同的数据类型。此外，技术的飞速进步，特别是计算能力、数据存储和管理技术的发展，极大地增强了人类获取和处理天文数据的能力。这些因素促进了天文学迅速转变为一门高度依赖数据的学科。

随着海量的天文数据资源不断涌现，天文学领域吸引了更多研究者投身于这项探索宇宙奥秘的事业。这种趋势进一步促使天文文献数量持续攀升。学术期刊、会议论文、以及预印本网站上，关于天文观测、理论模型、数据分析等研究成果层出不穷，形成了一个庞大且不断拓展的天文知识体系。

天文数据与其衍生的天文文献并不是各自独立的资源，而是相互依存、互为补充的宝贵资产。数据为天文研究的文献产出提供了实证基础，而科学文献则将这些数据表征的天文现象及其背后的科学原理进行系统化、理论化的阐述。天文学的发展不仅依赖于观测技术的进步，也依赖于科学文献以及其对数据深入理解的广泛传播。因此，实现天文数据与文献的深度融合、关联检索成为了天文数据驱动科学研究范式下的一项重要任务。

1.1.1　国内外天文数据与文献检索主要平台简介

研究人员经常需要在研究工作中对天文档案数据和科学文献进行检索，这既能避免做重复的工作，又能够在前人的研究成果之中得到启发。当前，国内外已经存在诸多天文数据和文献检索平台，诸如 ADS、SIMBAD、NED、ESASky、NADC 天文数据检索平台等，为天文学研究提供了宝贵的资源和工具。

ADS（The SAO/NASA Astrophysics Data System）[1] (Kurtz 等, 2000; Accomazzi 等, 2000; Accomazzi, 2024) 由美国航天局（National Aeronautics and Space Administration，NASA）[2] 资助，并由哈佛史密松天体物理中心（Harvard-Smithsonian Center for Astrophysics，CfA）[3]下的史密松天体物理台（Smithsonian Astrophysical Observatory，SAO）[4]管理和维护，是全球极为重要的文献检索系统。ADS 广泛

[1]ADS 官网https://ui.adsabs.harvard.edu/

[2]NASA 官网https://www.nasa.gov/

[3]CfA 官网https://www.cfa.harvard.edu/

[4]SAO 官网https://www.si.edu/about/astrophysical-observatory

收集各类科学文献，包括同行评审的期刊文章、会议论文、观测数据报告以及预印本等，同时也覆盖了包括天文学、物理学、天体物理学、太阳物理学、行星科学、地球物理学以及科学教育等众多学科领域，截至 2023 年共收录了超过 1500 万篇文献[5]。ADS 的强大之处不仅在于其丰富的文献储备，还在于其高效的检索能力和友好的用户界面设计。用户可以通过多种信息进行检索，例如按照作者、出版年份、关键词或标题等。此外，ADS 还提供高级搜索功能，允许用户进行更精确的查询，以及定制化的搜索结果展示选项 (Eichhorn 等, 2000)。

SIMBAD（Set of Identifications, Measurements, and Bibliography for Astronomical Data）[6] (Wenger 等, 2000) 是一个专注于太阳系外天体的重要天文数据库，由法国斯特拉斯堡天文数据中心（Strasbourg Astronomical Data Center，CDS）[7] 负责开发和维护。截至 2024 年 3 月，该数据库已累积包含了超过 1700 万颗天体，这些天体涵盖了恒星、星系、星团、行星、超新星等众多类型；同时，也包含了 6200 余万个天体标识符和 43 多万条参考文献的信息。SIMBAD 允许用户通过多种方式检索天体信息，包括天体的名称、坐标或星等等。在用户检索的结果中，SIMBAD 通过与 VizieR 星表数据库[8]以及 Aladin[9]的集成，使用户能够轻松访问相关的星表数据、图像资源以及观测日志。此外，SIMBAD 中天体的标识符直接链接至《天体命名词典》（*Dictionary of Nomenclature of Celestial Objects*[10]），通过该链接可以得到该天体的完整描述，并能够直接访问 CDS 提供的相应星表。SIMBAD 还支持通过检索天体标识符或者其它天体检索方式来获得包含本天体相关文献的 Bibcode，这些 Bibcode 会指向 CDS、ADS 和其它可用期刊网站上的基础书目，大多数情况下都能够直接获取论文全文。相反，如果某篇文献中包含天体标识符，通过 Bibcode 同样也可以获取到其天体标识符列表。

NED（NASA/IPAC Extragalactic Database）[11](Helou 等, 1991; Mazzarella 等, 2007) 是一个专注于河外天体和相关研究的综合性天文数据库。由 NASA 的喷气推进实验室（Jet Propulsion Laboratory，JPL）[12]和加州理工学院（California Institute of Technology，Caltech）[13]的红外处理与分析中心（Infrared Processing & Analysis Center，IPAC）[14]共同维护。NED 的主要目标是支持天文学家在进行河外星系研究时能够快速获取信息和数据。它不仅收录了来自于各个天文台和研究机构的观测数据，还包括了理论模型和模拟研究的结果。其中包含的数据类型有天体坐标、红移、图像、多波段数据和相关文献等。基于这些数据，NED 提供了多种

[5]ADS 简介https://ui.adsabs.harvard.edu/about/

[6]SIMBAD 官网，https://simbad.u-strasbg.fr/simbad/

[7]CDS 官网https://cdsweb.u-strasbg.fr/

[8]VizieR 官网https://vizier.cds.unistra.fr/

[9]Aladin 官网https://aladin.cds.unistra.fr/aladin.gml

[10]天体命名词典https://vizier.cfa.harvard.edu/viz-bin/Dic

[11]NED 官网https://ned.ipac.caltech.edu/

[12]NASA 喷气推进实验室https://www.jpl.nasa.gov/

[13]加州理工学院https://www.caltech.edu/

[14]红外处理与分析中心https://www.ipac.caltech.edu/

搜索工具和接口，使用户能够根据不同的需求进行详细的数据检索，如通过天体名称、坐标或其他属性进行搜索。此外，NED 也支持与其它相关外部数据库无缝对接。例如，NED 与 SIMBAD、ADS 有着良好的互操作性，能够一键式查阅相关文献和数据，以便用户进一步探索和分析数据。

ESASky[15]是一个由欧洲空间局（European Space Agency, ESA）[16](Giordano 等, 2018; Baines 等, 2016) 开发和欧洲空间天文中心（European Space Astronomy Centre，ESAC）[17]科学数据中心（ESAC Science Data Centre, ESDC）[18]团队维护的天文数据门户，旨在为天文学家和爱好者提供一个统一的访问入口，以便于探索和分析来自 ESA 各种空间望远镜和任务的天文数据以及对应的文献资源。这个平台集成了多个天文观测项目的数据，包括但不限于盖亚（Gaia）任务、赫歇尔（Herschel）空间天文台任务、普朗克（Planck）卫星任务以及欧几里得（Euclid）任务。针对这些数据，ESASky 内置了多种数据分析工具，如视图阅览器、数据挖掘工具和虚拟观测工具，支持用户进行深入的数据分析和可视化。对于以上工具，ESASky 提供了直观的用户界面，使得用户可以通过简单的图形界面来浏览和查询天文数据，无需深入了解复杂的数据库查询语言。例如，用户可以直接点击某一个天文对象来快速获取相关的科学文献和研究报告。这些服务可以帮助用户更好地理解和利用天文数据，成为公众了解宇宙奥秘的一扇窗口。

国家天文科学数据中心（National Astronomical Data Center，NADC）[19](Cui 等, 2020)，作为国内天文数据管理和共享的重要门户，提供了一个高效、便捷的在线检索服务平台。该系统不仅为国内天文数据的归档和镜像提供了强大的支持，而且成为了国际天文数据交流的重要桥梁。截至 2022 年 3 月，NADC 已经成功整合并发布了 30 多套中国自主建立和国际公认的重要天文数据资源。这些数据集涵盖了从地面观测到空间探测任务，包括但不限于 LAMOST、FAST、AST3、SCUSS、BASS、SkyMapper、Gaia、WISE 等。用户可以通过 NADC 的检索系统，轻松访问和检索这些丰富的天文数据资源。检索结果不仅可以在网页上直观展示，还可以直接下载保存，极大地方便了科研人员的数据获取和使用。无论是进行学术研究还是教育普及，NADC 都提供了强有力的数据支持。尽管 NADC 在天文数据检索方面取得了显著成就，但目前尚缺乏一套能将天文数据与科学文献紧密结合实现关联检索的系统。

1.1.2　天文数据与文献关联检索面临的挑战

通过观察以上平台可以发现，天文数据与文献的关联检索主要有两种模式，分别为非结构化数据库检索模式和结构化数据库检索模式。

非结构化数据库检索模式是指通过平台提供的窗口进行检索，筛选出与某

[15]ESASky 官网https://sky.esa.int/esasky

[16]ESA 官网https://www.esa.int/

[17]欧洲空间天文中心https://www.cosmos.esa.int/web/esdc/home

[18]欧洲空间天文中心科学数据中心https://www.cosmos.esa.int/web/esdc/home

[19]中国国家天文科学数据中心https://nadc.china-vo.org/

表 1-1　国内外支持天文数据与文献检索的主要平台

Table 1-1　Major platforms supporting astronomical data and literature retrieval domestically and internationally.

序号	平台名称	运行单位
1	ADS	美国史密松天体物理台
2	SIMBAD	法国斯特拉斯堡天文数据中心
3	NED	美国喷气推进实验室 & 加州理工学院
4	ESASky	欧洲空间天文中心科学数据中心
5	NADC	中国国家天文科学数据中心

知识实体相关的文献，例如 ADS。尽管这些平台在一定程度上可以完成检索任务，然而它们主要依赖于关键词匹配机制。随着天文文献数量越来越庞大，这种基于关键词的检索方式逐渐暴露出其众多局限性。

1) 用户在面对庞大的文献库时，挑选出恰当的关键词变得颇有难度；例如，用户需要检索“M 31”相关的文献，然而 SIMBAD 显示它一共具有 39 个标识符[20]。

2) 用户检索某些信息并不能获取到相应的结果；例如，通过 ADS 来检索天体标识符“HD 213893”无法获取到任何文献，而事实上，包含这个天体标识符的文章的确存在：

“The data features used for this work include photometry in eight bands (r, i, z, J, H, K, W1, and W2). As an example, we show the spectra of ***HD 213893****, which is classified as M0 type and plotted in Figure 4 from optical to infrared.”* (Qu 等, 2024)

3) 即便用户成功检索到相关文献，也可能面临从大量结果中进一步筛选真正有价值文献的繁琐任务。虽然大部分文献检索平台尝试通过相关性算法对文献进行排序，以便于提供更加符合用户需要的检索结果，但实际效果往往不尽人意。例如，在检索望远镜名称“LAMOST（Large Sky Area Multi-Object Fiber Spectroscopic Telescope）”相关文献时，ADS 将返回大量文献，而其中排序靠前的文献也会出现与“LAMOST”关联程度低的情况，它们可能仅仅在文中提及或致谢 LAMOST。

“With the advent of very large data sets of stellar spectra such as the SDSS, ***LAMOST****, RAVE, and Gaia, it is no longer remotely possible for an observer to digest the contents of an entire data set visually.”*(Carbon 等, 2017)

4) 在天文学领域，由于专业术语的广泛使用和概念的多样性，经常会遇到所谓的语义重合现象，即同一个关键词可能对应多个不同的概念或实体。这种现象在文献检索时可能会出现问题，因为仅仅依赖关键词进行信息检索可能会带来大量的不相关信息。例如，如果检索望远镜名称“FAST（Five-hundred-meter

[20]M 31 星系详细信息https://simbad.cds.unistra.fr/simbad/sim-id?Ident=M31

Aperture Spherical radio Telescope)”，结果中会出现大量包含英文单词“fast”的文献，而其中大部分“fast”传达的意思仅为“快的”；不仅如此，国际天文学界中习惯使用人名来对望远镜或观测任务进行命名，如“Hubble”，这同样也给文献检索带来了挑战。

*“**Fast** radio bursts (FRBs) are very bright radio pulses with millisecond durations ... The most sensitive single-dish telescope at present, i.e., the Five-hundred-meter Aperture Spherical Telescope (**FAST**), is extremely powerful in the fine characterization of repeating FRBs.”*(Jiménez Martínez, 2023)

*“An intensive reverberation mapping campaign of the Seyfert 1 galaxy Mrk 817 using the Cosmic Origins Spectrograph on the **Hubble** Space Telescope revealed significant variations in the response of broad UV emission lines to fluctuations in the continuum emission.”*(Homayouni 等, 2024)

结构化数据库检索模式是指通过平台事先归纳并关联好的数据库进行检索，例如 SIMBAD。它们会定期对非结构化的期刊文献进行知识实体（如天体标识符）识别、抽取和归档，最终形成结构化的关联数据库，从而大幅度提升用户检索相关文献的速度。在进行提取任务时，运行团队首先采用某种特定方法来初步实现知识实体抽取；随后，为了确保抽取准确性和完整性，运行团队需要投入大量的人工来完成遗漏补充、错误信息纠正等工作。然而，随着天文学数据和文献资源的快速积累，传统的人工密集型方法已经难以满足需求。这些方法在处理大量数据和复杂任务时效率低下，无法及时完成数据的关联和分析工作。

由此可见，用户在使用传统的文献检索系统时，不可避免地需要投入大量时间与精力；而通过平台事先关联好的数据库进行检索可以帮助用户精准地获取相应数据或文献，但是庞大的数据量与文献数量使得运行团队难以招架。天文知识实体是实现天文数据和文献关联的关键纽带。因此，寻求先进的知识实体抽取方法，自动化、智能化地从非结构化的天文文献中提取出准确的实体信息，并将其整合至结构化的知识数据库中，已经成为天文领域一项重要的研究趋势和技术需求。

本文旨在探索一种高效且强大的天文知识实体抽取方法，以此来提高数据和文献检索的效率和准确性，进一步为天文学的数据分析、知识发现和科学研究提供强有力的支持。

1.2　国内外研究现状

本小节主要介绍实体抽取的相关概念与研究进展，包括相关定义、发展历程以及在不同学科领域的应用现状，特别是在天文领域的研究动态，并进一步分析该技术在处理专业天文文本中的重要性以及面临的挑战。

1.2.1 实体抽取概述

实体通常可以分为两类。第一类，诸如人名、地名、组织名、时间、数值、货币等普遍存在的实体，并不局限于特定学科领域。相对地，如果具备某个学科领域背景的实体，通常被称为知识实体（Knowledge Entity）(Chang 等, 2008; Ding 等, 2013; 温雯 等, 2018; Al-Moslmi 等, 2020; 李广建 等, 2023)。如物理学领域中的定律和原理、实验设备、理论模型等；化学领域中的化合物、化学反应、实验方法等；天文学领域中天体标识符、星表、天文现象、望远镜名称、观测技术、巡天计划实体等。表 1-2展示了不同领域知识实体的示例。

表 1-2　不同学科领域中的知识实体示例

Table 1-2　Examples of knowledge entities in different subject areas.

学科领域	知识实体类型	参考文献
材料学	材料类型、化学反应、微观结构、物理性质、化学性质、理论模型、加工技术、实验仪器等	Weston 等 (2019); 李洋 等 (2022); 韩玉民 等 (2022)
化学	元素、化合物、化学键、分子结构、试剂、化学反应、试验方法、分析技术等	Eltyeb 等 (2014); Leaman 等 (2015); 杨培 等 (2018)
医学	中医领域实体类型：疾病、症状、药物、治疗方法、人体系统、医学设备、医学检查项目、科室等	Liu 等 (2017); 杨巍 (2021); 耿飙 等 (2024); 吉旭瑞 等 (2024)
生物医药	药物、试剂、基因、疾病、病毒、生物分子、化学成分、机理、适应症、不良反应症状等	Zhang 等 (2013); Perera 等 (2020); 孙聪 (2021)
历史学	历史事件、历史人物、历史时期、政治制度、文化与宗教、法律与规章、经济模式、地理环境、文化遗产等	Byrne (2007); 曹树金 等 (2022); 崔鑫 等 (2023); Ehrmann 等 (2023)
计算机科学	计算机体系结构、数据库系统、编程语言、算法与数据结构、应用软件、技术领域、数据集、参数、评价指标等	彭嘉毅 等 (2019); D'Souza 等 (2022); 陈祥 等 (2023)
天文学	天体、望远镜、星表、天文现象、观测技术、巡天计划、理论模型、天文学分支等	Murphy 等 (2006); Grezes 等 (2021); Shao 等 (2023); Sotnikov 等 (2023)

实体抽取（Entity Extraction）是自然语言处理（Natural Language Processing, NLP）领域的一项关键技术，它旨在从非结构化文本中自动识别和提取具有特定意义的实体信息 (Sundheim, 1995; 郭喜跃 等, 2015; 刘浏 等, 2018; 刘春丽 等, 2023)。实体抽取作为信息抽取（实体抽取、关系抽取、事件抽取）的重要子任务，是文本挖掘、文献检索、信息关联与推荐、知识图谱构建等任务的基础工作

(Chinchor 等, 1998; Jiang, 2012; Piskorski 等, 2013; 陈基, 2016; 昝红英 等, 2020)。

实体抽取技术的研究发展史可以追溯到 20 世纪中叶，当时主要的研究工作集中在从自然语言文本中获取结构化实体信息，这标志着实体抽取技术的初步探索阶段 (刘峤 等, 2016; 刘胜宇, 2016; Zong 等, 2021)。在这一时期，两个具有代表性的研究项目出现在了大众视野。20 世纪 60 年代中期，在美国国家科学基金会（National Science Foundation, NSF）[21]下属的科学信息服务办公室（Office of Science Information Services, OSIS）[22]资助下，纽约大学启动了 Linguistic String 项目 (Sager, 1981, 1990)，旨在为科学家提供快速获取科学文献中信息的服务。他们探索的途径之一就是通过计算机分析语言，构建大规模的英语计算语法，从而完成精确检索所需信息的任务[23]。同期，耶鲁大学的 FRUMP 系统利用故事脚本理论，致力于从新闻报道中抽取相关实体信息，内容包括环境、社会热点等众多场景或领域 (DeJong, 1982)。这些项目成为了实体抽取技术早期研究的典范。

进入 20 世纪 80 年代末，实体抽取研究开始蓬勃发展，这主要得益于两个因素：首先是可获取的在线和离线文本资源的爆发式增长；其次是消息理解系列会议（Message Understanding Conference, MUC）的举办 (Sundheim 等, 1993)。该会议得到了美国国防部高级研究计划局（Defense Advanced Research Projects Agency, DARPA）[24]的资助，并在 1987 至 1998 年间成功举办了七届；其中，MUC-6 引入了新的信息提取任务，包括命名实体识别（Named Entity Recognition）、共指消歧（Coreference Resolution）和关系抽取（Relation Extraction）等 (Office, 1995; Grishman 等, 1996)。这些任务成为了往后信息抽取研究中最基本、最重要的方向。MUC 会议建立了一套详细的任务定义和严格的评估标准，为该领域的研究提供了明确的方向和评价依据，极大促进了实体提取技术的研究与发展。

进入 21 世纪，自动内容抽取 (Automatic Content Extraction, ACE) 国际评测会议接棒 MUC。会议的目标是开发先进的信息抽取和文本分析技术，实现对多语言文本的高效自动化处理 (Doddington 等, 2004)。ACE 评测会议不仅关注特定领域文本中实体的处理，还强调了跨文档处理能力和综合评价体系的建立，这标志着实体抽取技术研究的深化和完善。

随着各学科领域的发展，专业文本资源也在急剧增加。面对这些庞大的文本资源库，从中提取专业知识实体的需求也日益增长。因此，研究人员基于这些早期的实体抽取研究，针对各自领域的特点展开了深入的探索，希望通过实体抽取技术从非结构化的科学文本中快速获取专业知识实体。

1.2.2　其它领域知识实体抽取研究现状

知识实体抽取方法当中，人工标注方法是一种最为传统的手段。这种方法依赖于专家对文本中隐藏的知识实体进行详尽的识别和标注，以此构建特定的数

[21]美国国家科学基金会https://www.nsf.gov/

[22]科学信息服务办公室https://www.ncbi.nlm.nih.gov/pmc/articles/PMC200417/

[23]Linguistic String 项目https://cs.nyu.edu/cs/projects/lsp/LSP_intro.html

[24]美国国防部高级研究计划局https://www.darpa.mil/

据集或语料库 (QasemiZadeh 等, 2016; 章成志 等, 2021)。例如，Howison 等 (2016) 对 90 篇生物学文章的随机样本进行人工标注，来筛选出其中的软件名称；Wang 等 (2020) 研究聚焦于 NLP 领域的算法实体，通过手工标注方法建立了算法词汇库。尽管人工标注精确度高、灵活性强，但它也存在明显的缺点，如人工成本高昂，工作效率较低等。此外，人工标注的质量受限于标注者的专业水平和主观判断，可能导致标注结果的不一致性。尽管如此，由于人工标注的数据集对于训练其他自动化实体抽取模型至关重要，因此在一些复杂或敏感领域，如临床医学、生物医药、古籍文学等，人工标注方法仍然发挥着不可替代的作用。随着技术的发展，人工标注方法也在不断地与其它先进技术相结合，优劣互补以提高实体抽取的效率和准确性。

基于规则的实体抽取方法依赖于预先定义的规则集，这些规则通常基于特定领域的知识库和词典，通过与文本中文字或字符等信息进行匹配来实现对知识实体的自动化抽取。Friedman 等 (1994) 尝试将字典、模式匹配等多种方法融合至一个系统，以最大程度地提升医疗领域知识实体抽取任务的效果。Puccetti 等 (2023) 构建了专利关键字词典，并且使用正则表达式模型在专利文本中提取有价值的知识实体；化柏林 (2013) 运用词表与规则相合的方法从文献中抽取方法术语；许华 等 (2015) 采用基于规则的方法抽取药品说明书中的疾病、症状和致病菌三种知识实体。然而，这种方法的缺点在于其灵活性不足，对于领域外的文本或新出现的知识实体可能难以适应。因此，规则库往往需要定期更新，以适应领域知识的变化和扩展。此外，规则库通常适应能力较差，这使得它们很难被应用到其他数据集或领域中去执行知识实体的抽取任务。

传统的机器学习方法在知识实体抽取领域扮演了重要角色。这种方法通常将实体识别任务视为分类或序列标注问题，通过特征提取和模型训练来实现 (史永刚, 2006)。首先，文本被分词并转换为特征向量，这些向量捕捉了词法和句法信息。然后，利用如最大熵模型（Maximum Entropy Model, MaxEnt）(Jaynes, 1982)、隐马尔可夫模型（Hidden Markov Model, HMM）(Baum 等, 1966, 1970)、条件随机场（Conditional Random Field, CRF）(Lafferty 等, 2001)、支持向量机（Support Vector Machine, SVM）(Cortes 等, 1995) 等，对这些特征进行分析，以识别和分类实体。吴阳 (2015) 利用条件随机场模型开发了一套提取财经文章中股票名称、股票代码等实体的系统；Ju 等 (2011) 尝试使用支持向量机在生物医学文献中抽取蛋白质、基因和 DNA 等专业实体；Sobhana 等 (2010) 基于条件随机场方法在地质学文本中提取河流、山脉、岛屿、村庄、岩石等地质类实体信息。尽管这种方法提高了知识实体抽取的准确性并减少了对人工标注的依赖，但特征工程的复杂性、对高质量标注数据的需求以及较弱的泛化能力仍是其主要困境。

为了减少特征工程的工作量，科研人员引入深度学习方法抽取知识实体。深度学习方法通常不需要人工来标记特征，它们能够自动将文本转换成嵌入（embedding）信息，然后从这些信息中学习文本数据的深层特征 (Li 等, 2020; LeCun 等, 2015)。这些特征反映了词汇的语义信息以及它们在文本中的上下文关

系。这些特征最终会被送入神经网络中，如卷积神经网络（Convolutional Neural Networks, CNN）(LeCun 等, 1995)、深度神经网络（Deep Neural Networks, DNN）(Hinton 等, 2006) 和长短时记忆网络（Long-Short Term Memory, LSTM）(Hochreiter 等, 1997) 等。随后，神经网络需要依据这些特征来识别实体的存在，并确定每个实体的类别和范围。为了精确地界定实体的边界，通常会采用一系列规则或算法，如 SVM、CRF 等一些传统的机器学习方法，对文本中的每个词进行分类决策，以识别和标记出特定实体。图 1-1展示了深度学习方法实现实体抽取的主要步骤。例如，Huang 等 (2015) 尝试将双向的 LSTM 与 CRF 相结合提出了经典的 BI-LSTM-CRF 模型；Qin 等 (2018) 利用此模型在非结构化的电子病例识别临床医学实体信息；Luo 等 (2018) 在这个模型基础上加入注意力机制，提高在文档中提取化学品名称的能力；马娜 等 (2020) 则使用 BiLSTM-CNN-CRF 模型有效地探索了术语型引用对象自动化识别方法。诸如此类的研究足以说明基于深度学习的方法在实体抽取任务中成效显著，但是它们仍然面临着对大规模标注数据（非标注特征）依赖性强和计算资源需求高等问题。为了克服这些挑战，研究者们正在探索如何利用迁移学习、数据增强等技术来提高模型的泛化能力和效率。

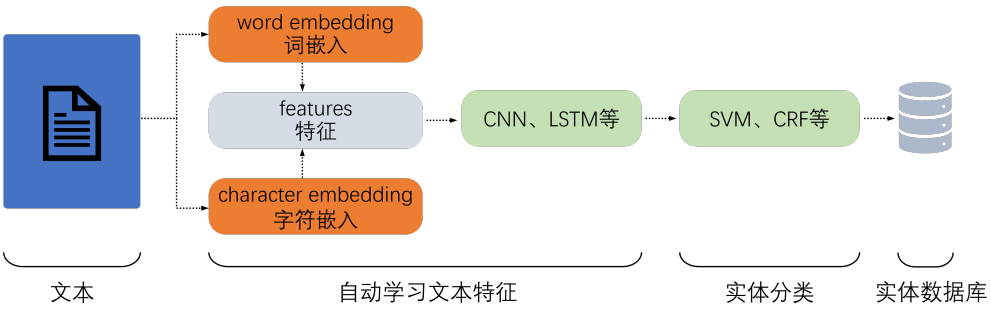

图 1-1　深度学习方法抽取实体的主要步骤

Figure 1-1　The main steps of extracting entities in deep learning methods.

深度学习方法在处理文本数据时，往往依赖于固定尺寸的窗口或卷积核，这使得其可能难以识别出跨越距离较大或结构复杂的实体。此外，深度学习模型的训练往往需要大量的标注数据，但在某些领域，可标注数据可能相对有限，这会显著影响模型的性能和泛化能力。然而，近些年，基于 Transformer 模型架构的预训练语言模型（Pre-training Language Model, PLM）发展迅速，它们的出现为这些问题提供了似乎更加有效的解决方案 (Thirunavukarasu 等, 2023)。自Devlin 等 (2018) 提出 BERT（Bidirectional Encoder Representations from Transformers）模型以来，基于 BERT 模型的各种专业领域模型爆发式增长。例如，生物医学领域的 BioBERT(Lee 等, 2020) 和 PubMedBERT(Gu 等, 2021)，临床医学的 ClinicalBERT(Huang 等, 2019a)，金融领域的 FinBERT(Liu 等, 2021) 等。

同时，研究人员发现，通过扩大参数规模，可以显著提升预训练语言模型的性能，还展现出了 BERT 等小型模型所不具备的特殊能力。为了便于区分不同参

数规模的模型，研究人员提出了“大语言模型（Large Language Model, LLM）”这一术语，用来指代那些具备数百亿甚至数千亿参数的预训练语言模型，如 GPT、Llama、Claude 系列模型等。这些模型基于海量的文本数据进行预训练，能够捕捉丰富的上下文信息，并且能够利用这些信息推断实体的界限和类别。这种方法显著减少了对大规模特定领域标注数据的依赖，增强了模型在抽取知识实体的能力，尤其在识别复杂实体时表现出色。例如，Bian 等 (2023) 将实体抽取任务分解为实体跨度提取和实体类型确定两项子任务，利用未经微调的大语言模型分步解决生物医学复杂实体识别和抽取问题；González-Gallardo 等 (2023) 利用 ChatGPT 在历史文献中识别有意义的实体；时宗彬 等 (2023) 使用 GPT-3.5 等多种大语言模型在相关研究论文中识别有机电池材料实体。通过利用大语言模型的上下文理解能力，即使在数据稀缺的情况下，研究者们也能够更加有效地完成知识实体抽取任务。

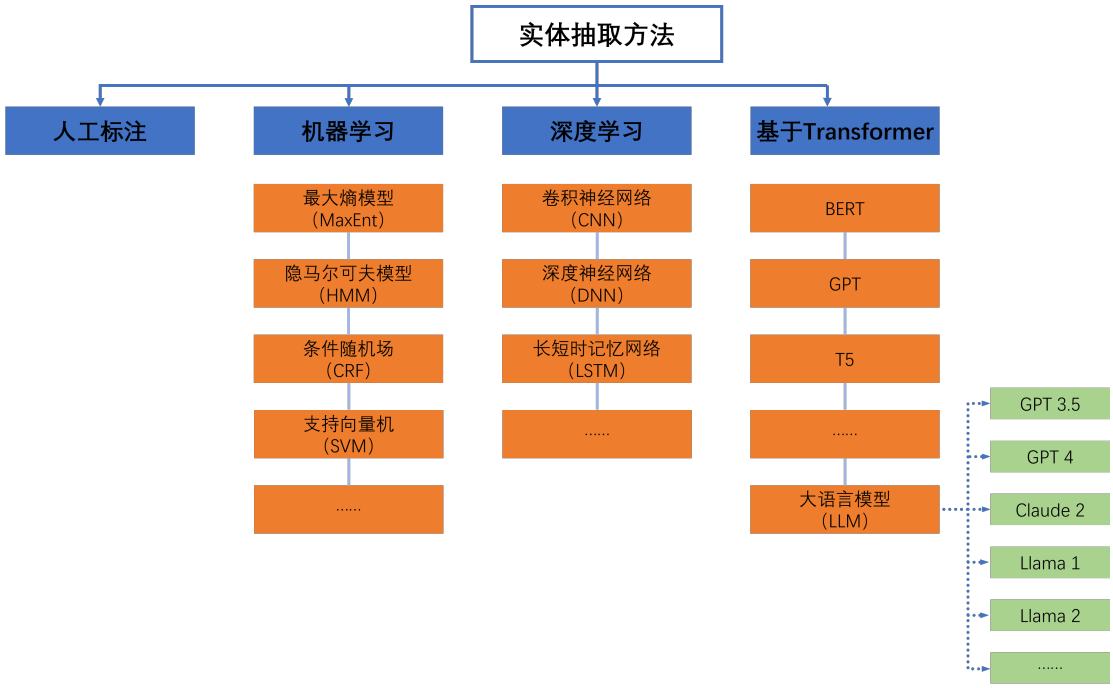

图 1-2　实体抽取的主要方法或模型

Figure 1-2　The main method or model of entity extraction.

1.2.3　天文领域知识实体抽取研究现状

天文学领域的知识实体抽取任务是指从非结构化的文本资源中准确地识别、提取并分类具有天文学意义的实体。与其他领域的实体相比，天文知识实体更为多样和复杂。表1-3展示了不同文献中包含的天体标识符和望远镜名称两种知识实体。通过观察这两类天文知识实体的特点，可以大致总结出天文学领域的知识实体提取存在以下挑战：

1）存在大量简写形式出现的天文知识实体。例如，在首次使用天体标识符“LAMOST J151003.74 + 305407.3”之后，文章会使用“J151”来作为替代；使用

“SMC” 来表示 “Small Magellanic Cloud”；使用 “Fermi-LAT” 表示 “Fermi Large Area Telescope”。

2) 命名方式差异性大。例如，有的天体会以 “望远镜名称 JHHMMSS.ss ± DDMMSS.ss” 结构化的命名形式出现，如 LAMOST J151003.74 + 305407.3；而有的天体则是以若干个字母和数字组合的非结构化命名形式出现，如 AS 245、RT Cru、354.98-02.87；还有的天体是以人们习惯的称呼出现的，如 Aldebaran、the Crab；甚至还有的天体名称中带有特殊字符，如 α^2 CVn、SgrA*。望远镜的命名方式也与之类似。

3) 实体边界难以确定。例如，天体标识符 PN H 2-5，SMP LMC 88，Large Magellanic Cloud，望远镜名称 Tsinghua-NAOC 0.8 m telescope，RATAN-600 等。

4) 陌生的甚至新的天文知识实体难以被识别。例如，天体标识符 “J091” 和望远镜名称 “Nanshan” 在缺乏前文背景介绍的情况下难以通过常规手段进行提取。

尽管抽取天文知识实体存在诸多挑战，但是随着天文领域对文本中实体信息的需求日益增长，研究人员也进行了一些探索性的研究。

与其它领域的知识实体抽取技术发展类似，天文领域早期工作主要依赖基于词典和规则的方法。DJIN（Journal of Identifiers and Names Detection）系统就是一个典型的例子。截止到 2010 年，工作团队根据《天体命名词典》(Lortet 等, 1994) 在系统中已经设计了超过 50000 个正则表达式，用来识别文章中的天体标识符 (Lesteven 等, 2010)。该系统在斯特拉斯堡天文数据中心中被成功应用，它与文献检索、天文星表检索无缝集成，构建了 SIMBAD 的一项特色服务。这种集成使得用户在文献搜索过程中能够轻松地浏览和访问文中提到的天体详细信息；同样也可以通过天体标识符实现对文献的检索。这项服务不仅优化了天文学家获取知识的过程，还突显了知识实体提取在融合天文学领域内各种信息资源中的关键作用。

在天文学实体识别和抽取的发展过程中，Murphy 等 (2006) 的研究代表了从传统方法向自动化、统计学习方法转变的重要一步。他们开发一个基于最大熵模型的天文实体提取系统，其能够应对的较为复杂语境，并准确识别出诸如源类型、源名称和观测设备等关键实体。通过利用大量标注数据，该系统不仅提高了实体识别的准确性，也为天文学界的信息提取和知识管理提供了新的工具。

基于 BERT 预训练语言模型，Grezes 等 (2021) 通过 395499 篇天文学研究论文的语料库训练出专门面向天文学领域的 astroBERT 模型。这个模型被用于开发 ADS 的命名实体识别工具，用于抽取其文献库中的天文组织、项目、术语等知识实体。并且在评估结果中，astroBERT 在天文知识实体抽取任务上的表现优于标准 BERT 模型。继这项工作取得成功之后，在 2022 年亚洲语言与计算国际会议（AACL-IJCNLP, Asian Association for Computational Linguistics and International Joint Conference on Natural Language Processing）的第一届科学出版物信息提取研讨会（First Workshop on Information Extraction from Scientific Publications, WIESP）

表 1-3　不同文献段落中天体标识符和望远镜名称两种天文知识实体示例

Table 1-3　Examples of two kinds of astronomical knowledge entities, celestial identifier and telescope name, in different literature passages.

实体类型	文献段落	参考文献
天体标识符	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 等, 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 等, 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 等, 2022)
	These data cover ∼30 deg^2 in the **SMC**, including the main bar of the galaxy, the wing extending to the east, and the tail extending further in that direction toward the **Large Magellanic Cloud**.	(Kuchar 等, 2024)
望远镜名称	…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 等, 2023)
	Gaia measurements of G29-38 will build on existing observations with **Keck**, the **Hubble Space Telescope**, **Herschel**, and **ALMA** …	(Sanderson 等, 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 等, 2022)
	We present the **Fermi-LAT** photon flux at the top of Figure 5, along with the **Nanshan** and **RATAN-600** radio flux density curves, the integrated and core 8.6 GHz flux densities obtained from VLBI observations in the bottom panel of Figure 5.	(Kun 等, 2022)

* 为了方便说明，这份表格默认将巡天计划名称、观测项目以及其它观测设备等归类为望远镜名称。

中提出了天文文献实体检测（Detection Entities in Astrophysics Literature, DEAL）

共享任务[25](Grezes 等, 2022)。DEAL 共享任务要求参与者构建能够自动提取天文学命名实体的系统，其中一些研究人员尝试训练或改进例如 mT5(Ghosh 等, 2022) 和 BERT(Alkan 等, 2022) 这样的预训练语言模型来从文本中提取天文知识实体，并取得了显著的成果。

最近，大语言模型在众多领域任务中展现出了卓越的零样本和少样本学习能力，这使得它们能够快速适应新的领域。因此，大语言模型也被研究人员积极应用于天文知识实体抽取任务。Sotnikov 等 (2023) 利用了诸如 InstructGPT-3 (Ouyang 等, 2022) 和 Flan-T5-XXL (Chung 等, 2022) 等大语言模型来从 Astronomer's Telegram 中的电报和 GCN Circulars 中的警报里提取天文知识实体，包括事件 ID 和天体名称等。他们探索了包括提示工程（Prompt Engineering）和模型微调（Model Fine-tuning）在内的各种方法来增强大语言模型的能力。他们的研究突出了大语言模型在天文学领域内实体抽取任务中的潜力。

1.3　主要研究内容

由于文献中天文知识实体的日益专业化和多样化，为每种类型的实体标注大量训练数据来开发模型显然是低效且不可持续的。因此，本文尝试探索大语言模型在天文文献中执行天文知识实体抽取（Knowledge Entity Extraction, KEE）任务的潜力。

本文选择了目前四种具有代表性的大语言模型，即 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2，并提出了一种名为 Prompt-KEE 的提示策略。Prompt-KEE 提示策略中包含 5 种提示要素，依据这些提示要素，本文设计了八种组合提示用以指导大语言模型从天文文献中提取天体标识符和望远镜名称两种最典型的天文知识实体。

为了适应这些大语言模型的 token 限制，本文构建了两种数据集，分别是 30 篇文献的全文文本数据集和这些文献的段落文本集合数据集。结合八种组合提示，本文使用 GPT-4 和 Claude 2 在全文文本数据集中进行测试，使用所有大语言模型在段落文本集合数据集中进行测试。结合实验结果，本文尝试分析影响大语言模型抽取天文知识实体性能的重要因素，并以此对未来天文文献的知识实体抽取任务提出了建议。

此外，本文分别采用基于规则的、基于机器学习的和基于小型预训练语言模型的方法进行了天文知识实体抽取实验，并且详细对比了大语言模型与它们的实验结果，在性能、工作模式、更新和维护方面展示大语言模型的特点与优势。

1.4　文章结构

本文共分为五个章节：

[25]DEAL 共享任务https://ui.adsabs.harvard.edu/WIESP/2022/SharedTasks

第一章　引言。首先介绍了论文的研究背景及意义。然后针对本文的研究内容介绍了国内外研究现状，主要包括实体抽取概述、其它领域的知识实体抽取研究现状、天文领域的知识实体抽取研究现状三个方面内容。最后给出了本文的主要研究内容与文章结构。

第二章　大语言模型与相关技术原理简介。首先介绍了大语言模型的定义和特点。接着，追溯了大语言模型的发展历程。此外，本章还详细讨论了大语言模型中的一些关键技术概念。技术原理部分深入解释了自注意力机制和 Transformer 架构。

第三章　基于大语言模型的天文知识实体抽取。利用大语言模型开展了一套完整的天文知识实体抽取实验，包括组合提示构建、实验天文知识实体与大语言模型选取、数据集构建、实验设置、实验评价和结果分析等。

第四章　其它实体抽取方法与大语言模型方法对比。对基于规则、基于机器学习、基于小规模语言模型的三种方法进行了探讨，并与 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 四种大语言模型的实验结果进行了详细的对比，说明了大语言模型在执行天文知识实体抽取任务中具有多方面优势。

第五章　总结与展望。对本文的工作内容做了总结，并阐述了本文研究的局限性以及可以进一步研究的内容。

第 2 章　大语言模型及其相关原理简介

大语言模型正在被广泛运用于解决各种实际任务，从自然语言处理到智能对话系统，再到智能辅助写作工具，大语言模型正在逐渐改变人类与数字世界的交互方式。大语言模型强大的语言理解和生成能力使其在诸多领域都发挥着重要作用，为用户提供更高效、更智能的解决方案。本章节将围绕利用大语言模型抽取天文知识实体的研究内容，对相关的大语言模型知识和原理进行介绍。

2.1　定义和特点

在人工智能领域，预训练（Pre-training）是指在小规模的、特定任务的数据集上进行微调（Fine-tuning）之前，在大规模数据集上训练模型的过程 (Qiu 等, 2020; Han 等, 2021)。

如图 2-1所示，大语言模型其实是一类先进的预训练人工智能系统，它们利用深度学习技术，尤其是基于 Transformer 架构的神经网络，在大规模文本数据集上进行预训练，来学习语言复杂的模式和特征，从而具备更强大的泛化能力，使其能够更好地理解和处理各种陌生的自然语言文本 (Zhao 等, 2023)。

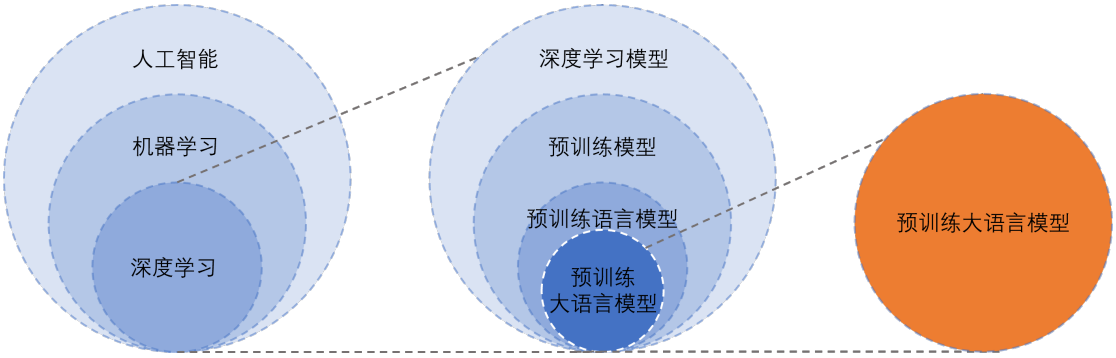

图 2-1　人工智能包含的研究方向

Figure 2-1　Research areas encompassed by artificial intelligence.

本质上，大语言模型是从小规模预训练语言模型演化而来的，但是与之相比，大语言模型具备以下几项显著的特点：

- **大规模数据训练**：大语言模型通常在 TB 以上甚至 PB 级别的大规模文本数据集上进行训练，这些数据集可能包括书籍、网页、报告、期刊论文等多种类型的文本资源。
- **涌现能力**：随着规模的增大，大语言模型有时会表现出一些小模型所不具备的能力，如简单的推理、理解和解决复杂问题，这些能力并非直接编程实现，而是随着模型复杂度的增加而自然出现的。
- **超大规模参数量**：参数量是指模型中可调整的参数的数量。大语言模型通常具有海量的涌现能力参数，这些参数在模型训练过程中不断调整，以更好地

捕捉和表示语言的复杂性。例如，GPT-3 的参数量就高达 1750 亿个 (Brown 等, 2020)。

- **强大的上下文理解能力**：通过预训练，大语言模型能够捕捉到语言的细微差别和上下文依赖性，从而在生成文本时考虑到上下文信息，提供更加准确合理的输出。
- **多任务学习和泛化能力**：大语言模型不仅能够执行特定的自然语言处理任务，如文本分类、实体抽取、情感分析等，还能够较好地适应新的领域或任务，展现出良好的泛化能力。
- **自动化训练和自我监督学习**：大语言模型的训练过程通常是自主的，它们通过自我监督学习，如预测文本序列中的下一个单词或遮蔽词汇，来提高语言理解和生成的能力。
- **巨大的计算资源消耗**：由于模型的结构复杂，以及参数和需要处理的数据规模庞大，大语言模型在预训练阶段需要消耗大量的算力。例如，训练像 GPT-3 这样的模型需要执行数千万亿亿次的浮点运算，这通常需要使用高性能的 GPU 或 TPU 集群来完成 (Anthony 等, 2020)。这些计算资源的消耗不仅体现在处理速度上，还包括能源消耗和成本投入。

正是这些特点才使得大语言模型在处理语言的深度和灵活性上与小规模语言模型有了显著区别，同时这些特点也赋予了它们在自然语言处理领域的先进性能和广泛应用前景。

2.2 发展历程

目前，大语言模型已经成为了人工智能领域重要的研究分支之一，其技术的进步被视作人工智能发展的风向标。然而，似乎是在 ChatGPT [1]发布之后，大语言模型才受到广泛关注。其实，在自然语言处理领域已经有了一系列研究成果和技术积累，这对于大语言模型的出现至关重要。

语言模型的研究最早可追溯至统计学习方法，这些模型通过统计单词或短语在文本中出现的概率来预测语言结构。尽管取得了一定的成效，但受限于数据和计算能力，这些早期语言模型的规模和性能十分有限。它们通常基于简单的 n-gram 模型 (Shannon, 1948; Brants 等, 2007)，但无法捕捉长距离的依赖关系。

随着深度学习技术的发展，神经网络开始被引入到语言模型中，词嵌入（Word Embedding）技术使得每个单词或短语能够被映射为高维空间中的向量，从而更好地捕捉词与词之间的关系。这一阶段的模型，如 Word2Vec (Mikolov 等, 2013b,a) 和 GloVe (Pennington 等, 2014)，为后续的语言模型发展奠定了基础。

2017 年，Transformer 模型 (Vaswani 等, 2017) 的提出为语言模型的发展带来了革命性的变化。其基于自注意力机制的设计，有效解决了长期依赖问题，并为后续大模型的发展奠定了基础。Transformer 的提出，标志着自然语言处理领域

[1]ChatGPT 官网https://chat.openai.com/

从循环神经网络向注意力机制的转变。

2018 年，BERT 模型 (Devlin 等, 2018) 的发布标志着预训练语言模型时代的开始。BERT 通过大规模文本数据的预训练，展现了双向上下文理解的能力，并在多项自然语言处理任务上取得了前所未有的性能。BERT 的成功，引发了预训练模型的热潮，促使研究者开始探索更大规模的模型。

研究人员发现，随着模型参数规模的增加，模型的性能也能随之提升，这促进了人们对更大模型的追求。GPT 系列模型的发布，尤其是 ChatGPT 的问世，成为了大语言模型发展的里程碑。直到如今，诸如 GPT-4、Llama 2、Claude 2 等各种大语言模型层出不穷，深入应用到各个领域。

总体而言，如图 2-2所示，语言模型演变成如今的大语言模型共经历了四个关键阶段。

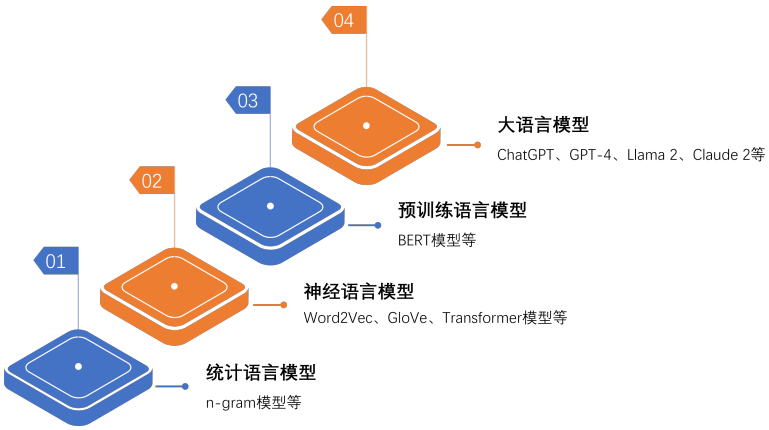

图 2-2　语言模型的四个关键发展阶段

Figure 2-2　Four key stages of development of language models.

2.3　相关知识

2.3.1　词向量与词嵌入

词向量（Word Vector）是指某个单词或短语在连续向量空间上的一组特定的数值映射，而词嵌入（Word Embedding）是指将单词或短语映射到实数向量的过程 (Mikolov 等, 2013b,a)。

在向量空间中，每个单词都被表示为一个向量，并且这些向量捕捉了单词的语义特征，使得具有相似含义的单词在向量空间中的位置相近。因此，在复杂模型中向量空间往往是高维的。

如图 2-3展示了二维向量空间上的词向量示例。其中“人工智能”、“自然语言处理”和“大语言模型”的向量更接近，而“天文”、“宇宙”的向量更接近。

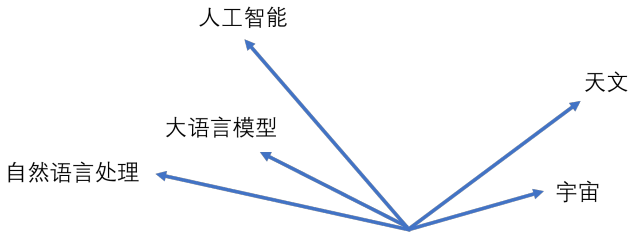

图 2-3　二维空间上的词向量示例

Figure 2-3　Example of word vectors in a two-dimensional space.

2.3.2　Token

在大语言模型领域，token 是对文本进行分割和编码时的最小单位，是模型处理文本数据的基础，也是自然语言转换为机器可解析形式的关键步骤 (Borgeaud 等, 2022; Zhao 等, 2024)。它具有以下几方面的特点：

- **形式多样**：它可以是单词、词缀、字符、标点或其他形式的文本片段。例如单词 “face” 可能直接是一个 token，而 “transformation” 则可能被分成 “trans-”，“forma-”，“-tion” 三个 tokens。
- **存在限制**：目前大语言模型的输入输出 token 数量均存在严格的 token 限制 (Hoffmann 等, 2022)，例如 Llama 2、GPT-3.5 和 GPT-4-32K 分别最多能够处理大约 4096、4096 和 32768 个 tokens。这是因为 Transformer 架构中使用的自注意力机制在处理每个 token 时都需要考虑整个输入序列，这导致计算量随序列长度的平方增长。另外也存在计算资源、内存容量以及使用成本等方面的限制。因此，为了保持模型的高效运行，需要限制序列长度。
- **转换处理**：Token 在大语言模型的高维映射空间中是以向量形式存在。因此，如图 2-4所示，大语言模型在处理自然语言文本时通常将其转换成 token 序列，并输入至神经网络中，通过一系列操作后，再将新的 token 序列以自然语言的形式从模型输出。

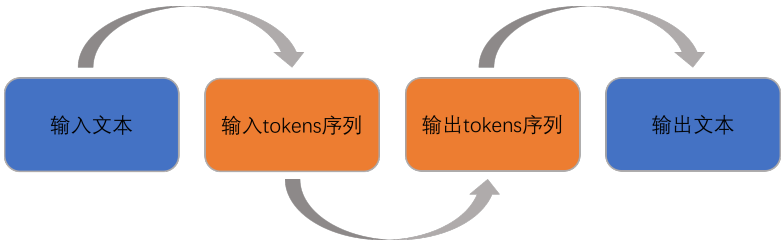

图 2-4　大语言模型处理文本的 token 转换流程

Figure 2-4　The token conversion process of a large language model for text processing.

2.3.3　提示工程

“Prompt engineering is the art of communicating with a generative AI model.”[2]

[2]提示工程和 LLMs 指南https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/

提示工程（Prompt Engineering）是指设计和优化输入提示（Prompt）来引导模型生成特定输出的过程。在大语言模型中，提示是用户提供给模型的文本输入，它可以是一个问题、一个主题描述、一段代码或者其它任何形式的文本，目的是激发模型产生符合预期的响应或行为。

提示工程的重要性在于，大语言模型通常是基于上下文来生成文本的，这意味着模型的输出质量很大程度上取决于输入提示的质量 (Amatriain, 2024)。良好的提示可以更有效地引导模型理解任务需求，从而生成更准确、更相关、更有创造性的输出 (Bsharat 等, 2023)。因此，OpenAI 和 Meta 等官方开发商为用户提供了丰富的提示工程指南[3,4]。以下是这些指南提供的一些关键性准则：

- **分配角色**：在提示中添加角色可以使大语言模型的输出更加多样。例如，当要求模型扮演某领域的专家，它会提供更加专业和深刻的见解。
- **分解任务**：将复杂的任务分解为一系列更简单的子任务分别进行提示。
- **定义术语**：当执行专业领域的任务时，准确的领域术语定义可以帮助模型深入理解。
- **提供示例**：当要求模型执行某类任务时，高质量任务示例提供的少样本提示对模型输出的准确性至关重要。
- **重复重点**：在提示中多次重复特定单词或任务等信息可以增强模型的注意力。
- **明确规则**：明确表述模型必须遵循某些规则来输出内容。

根据以上提示工程的内容可以进一步引出两个相关的概念，即上下文学习和思维链。

2.3.4　上下文学习与思维链

上下文学习（In-Context Learning）是指大语言模型能够根据给定的上下文信息来理解和生成响应的能力 (Dong 等, 2022)。这种能力允许模型在没有针对特定任务进行额外训练的情况下，通过观察一些示例（示范性输入和输出）来学习如何执行新任务 (Liu 等, 2022; Zhang 等, 2022a)。上下文学习体现了大语言模型的一种涌现能力。

思维链（Chain of Thought，CoT）是一种用于提升大语言模型在解决复杂推理任务上的方法 (Wei 等, 2022; Lyu 等, 2023)。它通过向模型展示一些少量的示例，在示例中解释推理过程，引导模型在回答问题时也显示出推理过程。这种技术可以帮助模型更好地处理需要多步骤逻辑推理的问题，如数学问题、常识推理等 (Fu 等, 2022)。思维链提示通常涉及将问题分解为多个中间步骤，并鼓励模型逐步生成这些步骤，从而提高最终结果的准确性 (Zhang 等, 2022b)。

图 2-5展示了大语言模型在执行任务时的上下文学习提示和思维链提示的对比示例。上下文学习提示更侧重于使用自然语言描述、样例来提示大语言模型，

[3]OpenAI 提示工程指南https://platform.openai.com/docs/guides/prompt-engineering

[4]Meta 提示工程指南https://github.com/meta-llama/llama-recipes

而思维链提示则关注引导大语言模型进行连续的推理。

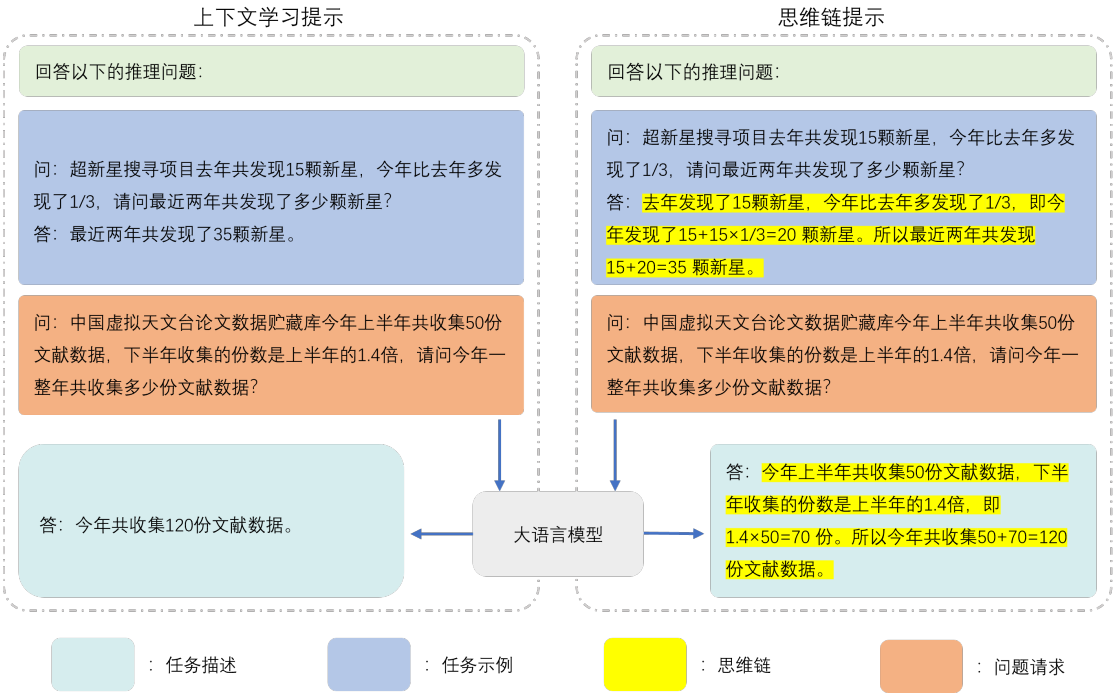

图 2-5　上下文学习提示和思维链提示的对比示例

Figure 2-5　Example of a comparison between a contextual learning prompt and a thought chain prompt.

2.3.5　幻觉

大语言模型的幻觉（也称为“幻觉效应”或“幻觉现象”）是指模型生成的信息在真实世界中并不存在，或者与现实情况不相符，又或者与用户的期望不相符，但模型却以一种非常确信的方式进行表达。这种现象在大语言模型中尤为常见，因为它们在生成文本时可能会编造事实、引用不存在的来源或者提供不准确的信息 (Li 等, 2024)。幻觉可分为两种类型，即事实性幻觉和忠实性幻觉 (Huang 等, 2023)。

事实性幻觉是指模型生成的信息不符合真实世界的事实。它包括捏造事实和与事实不一致两种类型。

- **捏造事实**：模型生成了真实世界不存在的信息。
- **与事实不一致**：模型生成了不符合真实情况的信息。

忠实性幻觉是指模型生成的内容与给定的上下文或指令不符，导致生成的文本与预期目标不一致或包含矛盾信息的现象，大致可以分为指令不一致、上下文不一致、逻辑不一致三种类型。

- **与指令不一致**：当语言模型被要求执行特定指令，如提取文本中的天体标识符知识实体，但模型输出的是望远镜名称，这就产生了指令与答案之间的不一致。

- **上下文不一致**：模型生成的文本与输入的文本在内容或风格上不匹配，例如，文本中的内容是苏梅克-列维九号彗星撞击了木星，但是输出的内容是苏梅克-列维九号彗星与地球相撞。
- **逻辑不一致**：生成的推理过程与最终结果不符合，比如，用户问题是：地球在哪个星系中？模型推理过程是：地球在太阳系中，太阳系在银河系中，大犬座矮星系是银河系的卫星星系；但是最终结果是：地球在大犬座矮星系中。

尽管幻觉现象可以通过改善训练数据质量、完善模型架构、优化提示策略等方式得以缓解，但是目前阶段大语言模型的幻觉现象无法完全避免 (Dhuliawala 等, 2023; Ji 等, 2023; Xu 等, 2024)。

2.4　关键技术原理

大语言模型通过学习大量文本数据，训练出能够掌握语言规律的神经网络。其本质在于利用学习到的语言模式和关联，来预测后续的词或文本片段，如图 2-6所示。通过这种方式，模型可以根据给定场景或文本作出适当响应，并生成自然流畅的文本。而这一系列工作的背后依赖于诸多关键技术与核心步骤。

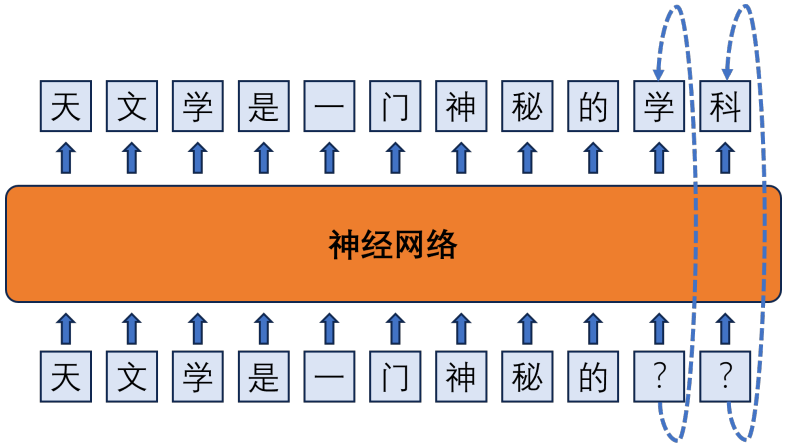

图 2-6　大语言模型预测生成文本示意图

Figure 2-6　Illustration of text generation by large language models.

2.4.1　自注意力机制

自注意力机制（Self-Attention Mechanism）是基于 Transformer 架构的大语言模型的核心组成部分 (Mnih 等, 2014; Bahdanau 等, 2014; Vaswani 等, 2017)。它是一种用于处理序列数据的机制，使得模型能够在学习和理解输入序列时将注意力集中在不同位置的信息上。这种机制允许模型在处理长序列时保持高效性，同时捕捉到序列中不同位置的重要关系。

在自注意力机制中，输入序列中的每个元素都被用来计算一组权重，这些权重表示了与该元素相关的其他元素的重要性。这些权重由查询（Query）、键（Key）和值（Value）的组合计算得出，其中查询用来衡量当前位置的重要性，键

用来衡量其他位置的相关性，而值则是被赋予权重的元素。自注意力机制通过将查询与键之间的相似度转换为权重，然后将这些权重与相应的值相乘并求和，来生成最终的输出。其注意力权重计算过程可以通过以下公式表示：

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2-1}$$

其中 Q, K, V 分别代表查询、键、值，d_k 是键向量的维度（用于缩放点积，避免过大的梯度），$\sqrt{d_k}$ 是其平方根，用于缩放点积的结果以稳定梯度。*softmax* 函数则将点积的结果转换为概率分布。

2.4.2 Transformer 架构

目前，绝大多数大语言模型采用的是 Transformer 架构，其核心思想是完全基于自注意力机制来处理序列数据。Transformer 架构主要包括以下几个组成部分：

- **自注意力层**（Self-Attention Layer）。自注意力机制允许模型在处理序列数据时，能够根据序列中其他位置的信息动态地对每个位置进行加权汇聚。这种机制允许模型在序列的不同位置计算注意力，这有助于捕捉序列内部的长距离依赖关系。
- **多头自注意力**（Multi-Head Attention）。它允许模型同时关注序列中不同位置的上下文。通过并行地执行多个注意力头，能够让这些注意力头学习到不同的表示。其公式为：

$$MultiHeadAttention(Q, K, V) = Concat(head_1, head_2, ..., head_h)W^O \tag{2-2}$$

其中，$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$，$W_i^Q$，$W_i^K$，$W_i^V$ 是第 i 个注意力头的查询、键、值矩阵权重，W^O 是多头注意力的输出权重，h 是注意力头的数量。

- **前馈神经网络**（Feed-Forward Neural Network，FNN）。在自注意力之后，每个位置的输出将通过一个前馈神经网络进行进一步的处理。其公式为：

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{2-3}$$

其中，W_1，b_1 和 W_2，b_2 是前馈网络的权重和偏置。

- **层归一化**（Layer Normalization）。在自注意力和前馈网络的输出上应用归一化，有助于稳定训练过程。其公式为：

$$LN(x) = \gamma\left(\frac{x - \mu}{\sigma_x^2 + \epsilon}\right) + \beta \tag{2-4}$$

其中，x 是输入向量，μ 是 x 的均值，σ_x^2 是 x 的方差，ϵ 是一个小常数以避免除以零，γ 和 β 是可学习的参数。

- **残差连接**（Residual Connection）。每个子层（自注意力和前馈网络）的输出都会加上其输入，然后进行层归一化，这有助于避免深层网络中梯度消失的问题。其公式为：

$$Output = LayerNorm\left(FFN\left(Attention(x)\right) + x\right) \tag{2-5}$$

其中，x 是层的输入。

总体上，Transformer 模型训练过程大致可以描述为：

$$P(y) = softmax\left(\frac{W^{O}Output(x)}{T}\right) \tag{2-6}$$

其中，$P(y)$ 是预测的概率分布，W^{O} 是输出权重矩阵，$Output(x)$ 是模型的输出，T 是温度参数，用于控制输出分布的熵。

基于以上的核心原理，Transformer 模型在执行过程种大致可以被划分为三个互相联系的层次，如图 2-7所示。

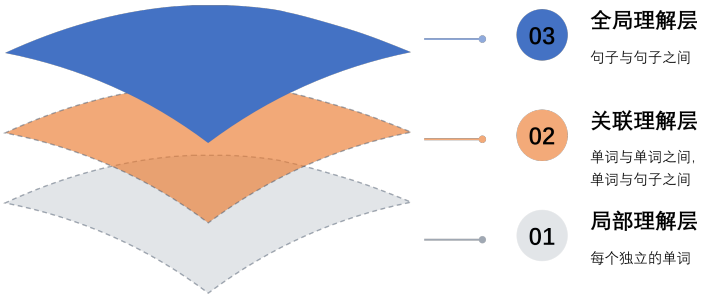

图 2-7　Transformer 模型执行过程包含的层次

Figure 2-7　Layers included in the Transformer model execution process.

- **局部理解层**：首先，Transformer 模型通过一个词嵌入层将文本中的每个单词转换为密集的向量表示。这一步骤是语言理解的基础，它允许模型捕捉到每个单词的独立语义信息。
- **关联理解层**：紧接着，在掌握了单词的独立含义之后，Transformer 模型的自注意力机制进一步分析单词之间的关联性。通过计算单词对之间的注意力权重，模型能够理解它们在特定上下文中的相互作用。这一层面的处理使得模型能够预测句子中接下来可能出现的单词，如“天文学是一门神秘的”之后很可能出现“学科”。
- **全局理解层**：最后，Transformer 模型的解码器部分利用编码器的输出来回忆和理解前面的内容，这对于生成或预测后续句子至关重要。模型不仅考虑单个句子内部的关联，还能够回忆之前句子中的信息，从而在生成回答或续写文本时考虑到更长距离的上下文关系。

2.5 本章小结

本章围绕利用大语言模型执行天文知识实体抽取任务的主题，首先介绍了大语言模型的定义和特点，包括其大规模数据训练的能力、超大规模参数、涌现能力、强大的上下文理解能力以及多任务学习和泛化能力。接着，本章追溯了大语言模型的发展历程，从早期的统计方法到现代的 Transformer 架构，说明了语言模型是如何逐步演化到如今强大而复杂的大语言模型。此外，本章还详细讨论了大语言模型中的一些关键技术概念，如词向量与词嵌入、token、提示工程、上下文学习和思维链，这些都是理解和运用大语言模型进行自然语言处理的基础。技术原理部分深入解释了自注意力机制和 Transformer 架构，这两者是大语言模型能够高效处理序列数据并捕捉长距离依赖关系的核心。

第 3 章　基于大语言模型的天文知识实体抽取

天文学期刊中包含大量的专业天文知识实体，如天体标识符和望远镜名称，这些实体通常来说难以理解和分辨。预训练的大语言模型虽然在通用语言任务上表现出色，但它们在面对特定领域时，尤其是天文学这种复杂领域，往往需要进行调整才能达到最佳性能。因此，设计专门的提示策略可以帮助模型更好地理解和适应天文领域知识实体抽取的特殊需求。受2.3.3小节提示工程的启发，针对过去研究者设计提示策略单一、简略且不明确的问题，本文根据天文领域文献的具体特征专门设计了一种名为 Prompt-KEE 的提示策略，旨在通过激活大语言模型在天文领域的潜在能力，减少对大规模标注数据的依赖，从而更准确和高效地识别和抽取天文知识实体。围绕 Prompt-KEE 提示策略，本章节利用大语言模型开展了一套完整的天文知识实体抽取实验，包括组合提示构建、实验天文知识实体与大语言模型选取、数据集构建、实验设置、实验评价和结果分析等。

3.1　Prompt-KEE 提示策略

Prompt-KEE 提示策略包含两个阶段的对话提示。在第一阶段，提示包括 4 个要素: 任务描述（Task Descriptions）、实体定义（Entity Definitions）、任务重点（Task Emphasis）和任务示例（Task Examples）。在第二阶段，提示包含 1 个要素，即二轮对话（Second Conversation），利用任务重点中的部分重要提示专门用于大语言模型的自我验证。以抽取文献中天体标识符和望远镜名称为例，按照 Prompt-KEE 策略本文进一步设计出了一组具体的提示模版，如图3-1所示。

3.1.1　任务描述要素

本文对这一部分进行了总体性的设计，以满足后续的对比实验。具体来说，首先，为了唤醒大语言模型所掌握的天文学知识，充分发挥其在天文学领域内的知识理解能力，包括对天文术语、概念和实体的深入把握，任务描述要求它们扮演一位有经验的天文学家角色，并告知它们需要掌握的工作能力 (Kong 等, 2023; Chung 等, 2024)。

- *You are an experienced astronomer, capable of easily recognizing knowledge entities (" celestial object identifiers" and "telescope names ") in a paragraph of astrophysics paper. Specifically, your task is to perform Knowledge Entity Extraction (KEE) task.*
- 您是一位经验丰富的天文学家，能够轻松识别天体物理论文段落中的知识实体（天体标识符和望远镜名称）。具体来说，你需要执行天文知识实体抽取任务。

其次，为了确保模型输出的信息便于后续的收集和实验分析，任务描述

<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 celestial object names in forms such as "several letters or numbers + constellation abbreviation" or "abbreviation of telescope name + coordinate". 2) The paragraph may also contain some telescope names represented by their own characteristics, a person's name, aperture length and address information. 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 object identifiers 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 celestial object names in forms such as "several letters or numbers + constellation abbreviation" or "abbreviation of telescope name + coordinate". 2) The paragraph may also contain some telescope names represented by their own characteristics, a person's name, aperture length and address information. 3) After outputting in JSON format, please provide your reasons for selecting these celestial objects and telescope names.

图 3-1 遵循 Prompt-KEE 提示策略用于提取天体标识符和望远镜两类天文知识实体的一组具体提示

Figure 3-1 A set of specific prompts that follow the Prompt-KEE prompting strategy for extracting two types of astronomical knowledge entities: celestial object identifiers and telescopes.

特别强调了模型需要以一种结构化的格式输出这些实体信息，即 JSON 格式 (Munnangi 等, 2024)。

- *The output should be provided in JSON format. JSON format example: "Celestial objects": ["XXX"], "Telescopes": ["XXX"].*
- *以 JSON 格式输出结果。JSON 格式示例:"Celestial objects": ["XXX"], "Telescopes": ["XXX"]*。

此外，考虑到天文学文献中经常会同时出现实体的全称和缩写形式，任务描述需要指导模型优先识别和提取实体的缩写形式，这是因为在学术写作中，天文学者更倾向于使用缩写来提高文本的紧凑性和可读性。

- *Knowledge entities in the paragraph may have both full names and abbreviations, please prioritize abbreviations based on the semantic context.*
- 段落中的知识实体可能同时具有全名和缩写形式，请根据语义上下文优先考虑缩写形式。

最后，由于目前多数大语言模型不支持直接处理以 PDF 形式呈现的图和表格数据 (Bisercic 等, 2023)，所以为了避免图和表格数据的干扰，任务描述选择让模型集中注意力在文本中抽取知识实体，确保所有大语言模型提取范围一致性。

- *Please do not extract data from the figures and tables, focus on the unstructured textual data.*
- 请不要从图表中提取数据，重点关注非结构化的文本数据。

3.1.2　实体定义要素

预训练的大语言模型在分辨高度专业化的天文术语方面可能会存在挑战，因为许多相似的术语会分散大语言模型的注意力，从而影响最终的表现。因此，实体定义为大语言模型提供了天文知识实体的准确定义，帮助模型更好地把握它们的概念。

首先，这里参考了维基百科（Wikipedia），对“天体（celestial object）”和“望远镜（telescope）”进行了定义，即：

- ***Celestial Object***：*A celestial object is a naturally occurring physical entity, association, or structure that exists within the observable universe.* [1]
- **天体**：天体是指存在于可观测宇宙中的自然存在的物理实体、关联体或结构。
- ***Telescope***：*A telescope is a device used to observe distant objects by their emission, absorption, or reflection of electromagnetic radiation.* [2]
- **望远镜**：望远镜是一种通过电磁辐射的发射、吸收或反射来观察远处物体的设备。

[1] 天体的定义https://en.wikipedia.org/wiki/Astronomical_object

[2] 望远镜的定义https://en.wikipedia.org/wiki/Telescope

接着，定义了“天体标识符（celestial object identifier）”以及“望远镜名称”（telescope names），即：

- ***Celestial Object Identifier***：*A celestial identifier is a unique tag or code used to identify and classify celestial objects.*
- **天体标识符**：天体标识符是用来唯一标识天文对象的名称或编号。
- ***Telescope Name***：*Telescope name refers to the unique designation given to a specific telescope.*
- **望远镜名称**：是指给特定望远镜分配的唯一标识。

需要注意的是，由于不同的命名习惯或标准，天体的表示可能有不同的术语，例如 Sun、Vega 这种形式通常被称为天体名称，而 LAMOST J004936.62+375022.8、AS Ser 等又通常被称为天体标识符。为了方便大语言模型理解，因此实体定义使用天体标识符作为通用术语。此外，本文没有严格区分望远镜名称、其他观测设施名称和巡天计划或项目名称，例如大天区面积多目标光纤光谱天文望远镜（LAMOST, Large Sky Area Multi-Object Fiber Spectroscopic Telescope）、全球天体测量干涉仪（Gaia, Global Astrometric Interferometer for Astrophysics）、斯隆数字化巡天（SDSS, Sloan Digital Sky Survey）；事实上，在科学研究中，天文学家通常不会强调它们之间的差异，因此，在天文文献中这些术语往往以并列的关系出现。据此，为了方便实体抽取实验，本文在实体定义要素中将它们统一归纳到“望远镜名称”中。

3.1.3 任务重点要素

在这一部分，任务重点详细阐述了大语言模型在执行天文知识实体抽取任务时需要特别关注的领域知识和注意事项。这部分内容的主要目的是激活和提高模型在执行任务时自我改进的能力，确保模型能够更准确地识别和抽取特定的实体。

首先，任务重点指明天体会有众多标识格式，如“若干个字母或数字 + 星座名称缩写”、“望远镜名称缩写 + 坐标”等，例如 AS Ser、LAMOST J004936.62 + 375022.8。

- *The paragraph may also contain some celestial object identifiers in forms such as ”several letters or numbers + constellation abbreviation” or ”abbreviation of telescope name + coordinate”.*
- 段落中可能还包含一些以“若干个字母或数字 + 星座名称缩写”或“望远镜名称缩写 + 坐标”等形式表示的天体标识符。

其次，望远镜通常可以通过自身特性、人名、地址信息和口径长度来命名，如大天区面积多目标光纤光谱天文望远镜（Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST）、詹姆斯·韦伯望远镜（James Webb Space Telescope, JWST）、丽江 2.4 米望远镜（Lijiang 2.4-meter Telescope）。这些强调会促使模型在处理文本时特别注意这些可能与天体标识符和望远镜名称相关的描述。

- *The paragraph may contain some telescope names represented by their own characteristics, a person's name, aperture length and address information.*
- 这段文字可能包含一些以自身特性、人名、口径长度和地址信息命名的望远镜。

另外，任务强调使用自检提示来鼓励大语言模型更深刻地理解上下文，从而纠正输出中的潜在错误 (Gero 等, 2023; Dan 等, 2023)。

- *Entities should be verified repeatedly before returning them.*
- 在输出实体之前，应对它们进行多次验证。
- *Please ensure that all entities from the paragraph have been extracted.*
- 请确保从段落中提取出了所有的实体。
- *After outputting in JSON format, please provide your reasons for selecting these celestial objects and telescope names.*
- 在以 *JSON* 格式输出后，请说明输出这些天体标识符和望远镜名称的原因。

最后，强调还包含了一些其它不可忽视的提示。大语言模型时常会出现不可控的幻觉现象，这就需要提醒模型不要创造文本中不存在的实体信息 (Wang 等, 2023)。要求提供提取相关实体的原因，能够帮助大语言模型完善思维链和梳理推理过程，增强其输出效果 (Zhang 等, 2022b; Huang 等, 2023)。

- *Don't create entities that's not in the given paragraph.*
- 不要创造给定段落中不存在的实体。
- *The given paragraph may not contain corresponding knowledge entities. If not exist, do not output such entities.*
- 给定的段落中可能不包含相应的知识实体。如果这样，请不要输出这些实体。
- *After outputting in JSON format, please provide your reasons for selecting these celestial objects and telescope names.*
- 在以 *JSON* 格式输出后，请提供选择这些天体标识符和望远镜名称的原因。

3.1.4　任务示例要素

为了让大语言模型学习天文知识实体抽取任务从输入到输出的映射，任务示例提供了一系列具体的示例来指导模型。这些示例旨在模拟真实的任务场景，帮助模型更加具体地理解天文知识实体的命名规则以及它们在上下文的关系，从而提高抽取的准确率 (Levy 等, 2022; Su 等, 2022; Ye 等, 2022)。

其中，第一个是综合示例，既包含了天体标识符，也包含了望远镜名称。其中，天体标识符是由星座名称简写和具体编号组成，而望远镜名称主要强调了它们可能与其位置和口径等信息相关。这三个示例句子分别来自Han 等 (2018), Zhang 等 (2020b)，和 Zhang 等 (2020a) 的论文。

- ***Example 1***:

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"];

第二个示例则针对天体的另一种主要的命名方式，这些标识符通常由望远镜名称的缩写和与之相关的坐标信息组成。

- ***Example 2***:

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"];

为了加深模型对任务的理解，第三个示例通过提供一段不包含望远镜名称的输入文本，指导模型不输出望远镜名称。这种设计有助于模型学习如何根据上下文信息区分和处理不同类型的天文知识实体，即使在某些实体的信息可能并未被直接提及的情况下。

- ***Example 3***:

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": []

3.1.5 二轮对话要素

大语言模型的研究者们意识到，在第一次抽取过程中可能存在实体识别遗漏或不准确的情况 (Ji, 2023)，因此在提示策略中设计了一轮额外的对话。

这一阶段的对话提示目的是引导大语言模型对先前的抽取结果进行验证和复查，纠正可能存在的错误并补充遗漏的信息。通过这种方式，模型可以实现信息校正与检索增强生成（Retrieval Augmented Generation，RAG）(Huang 等, 2023; Gao 等, 2023)，从而提高抽取天文知识实体的精确度。

- *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 ex-*

perience.

- 你提取的知识实体可能不完整或不准确，请结合前一阶段的提取结果和经验重新提取。

3.2　应用的大语言模型介绍

目前市面上大语言模型的种类繁多，选择合适的大语言模型应用在天文领域的知识实体抽取研究至关重要。结合前期对各种大语言模型的调研，本文选取了四种目前被广泛接受的大语言模型，分别是 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2，并且针对它们的架构特点、训练机制以及在天文学文本分析中的应用潜力进行了介绍。

3.2.1　Llama-2-70B

Llama-2-70B[3]是 Meta 公司[4]开发的一款先进的大语言模型，它是 Llama 2 系列模型（包括 7B、13B 和 70B 三种版本）的其中之一，拥有大约 700 亿个参数。Llama-2-70B 通过精心设计的预训练和微调过程，能够在多种基准测试中取得优异的成绩。具体来说，Llama-2-70B 在编码、数学和知识密集型任务上表现出色，同时也在多回合对话和多文档搜索查询等真实世界的语言任务中展露出强大的能力。最重要的是，此模型在与人类知识交互方面也表现优异，特别是在对话生成和理解方面 (Touvron 等, 2023)。

Llama-2-70B 这些强大的能力得益于其自身内在特性。首先，它采用了分组查询注意力（Grouped-Query Attention）机制，这种技术有助于模型更高效地处理大量信息，提高了推理的可扩展性。其次，Llama-2-70B 具有较大的上下文窗口，能够处理长达 4096 个 tokens 的输入，这显著提升了其处理长文本的能力。此外，该模型在预训练阶段使用了大量来自公开来源的数据，经过对这些数据的梳理与优化，在提升了训练数据的质量的同时，增强了泛化能力，减少了潜在的知识偏见。

Llama-2-70B 具备的这些内在优势使其成为天文学领域内一个极具潜力的工具 (Nguyen 等, 2023; Perkowski 等, 2024)。随着天文文献数据的不断增长和深入挖掘，Llama-2-70B 有望在推动天文学知识发现和传播方面发挥更加重要的作用。

3.2.2　GPT-3.5

2022 年 11 月 30 日，OpenAI[5]发布了 GPT-3.5，与之一同发布的还有 ChatGPT。ChatGPT 作为 GPT-3.5 的微调版本，是一款面向公众的智能聊天机器人，一经发布就受到了全社会的广泛关注，成为了先进 AI 的代名词。GPT-3.5 是 GPT-3 系

[3]Llama 2https://llama.meta.com/llama2/

[4]Meta 公司官网https://llama.meta.com/

[5]OpenAI 公司官网https://openai.com/

列[6]的进阶版本，它在模型性能和效率方面得到了显著的优化和提升。该模型采用了先进的 Transformer 架构，使用大规模的网页、书籍、维基百科等文本数据集进行预训练，掌握了海量的语言模式和规律，同时也使其具备丰富的领域知识背景。

GPT-3.5 模型的训练过程分为两个阶段：预训练和微调。在预训练阶段，模型通过无监督学习从大量文本中吸取知识，而微调阶段则通过有监督学习进一步优化模型，使其更好地适应特定的应用场景。这一灵活的训练策略使得 GPT-3.5 具备更广泛的用途，包括文本生成、问答、翻译、摘要生成等多种任务。同时，这也造就了 GPT-3.5 拥有包括 ChatGPT、GPT-3.5-Turbo、code-davinci-002 在内的众多版本。

此外，GPT-3.5 也具备不同上下文窗口的版本，即 GPT-3.5-4K、GPT-3.5-Turbo-16K，分别能够处理 4096 和 16385 个 tokens（约为 3072 和 12289 个单词，100 tokens ≒ 75 words[7]）的输入[8]。较长的上下文窗口使得模型在处理长文本方面的能力尤为突出，更好地理解复杂的语言结构和上下文关系。

GPT-3.5 作为 GPT 系列的一个里程碑，不仅在技术层面取得了显著进步，而且在实际应用中也同样展现了强大的适应能力。它的出现为自然语言处理领域带来了新的研究方向，同时也为诸如天文学等复杂研究领域的文本处理工作提供了机遇。

尽管 GPT-3.5 的功能强大，但它的使用成本也相对较高。根据 OpenAI 的定价方案，GPT-3.5 的使用费用会根据不同模型和处理的 token 数量进行计费（最新计价信息可查阅 OpenAI 官网[9,10]）。

3.2.3 GPT-4

GPT-4[11]是 OpenAI 继发布 GPT-3.5 之后的又一里程碑式的成果，代表着 AI 技术的最新进展。尽管 GPT-4 是一款大型多模态模型，能够通过理解文本来生成相关图像，甚至能够理解图像表达的含义 (Achiam 等, 2023)，但用户通常习惯称之为大语言模型，这主要归因于其具有更加强大的文本分析和推理能力。虽然在现实世界的许多场景中，GPT-4 的能力可能还无法与人类相媲美，但在许多专业和学术基准测试中，它已经表现出接近人类的水平，例如在模拟律师资格考试中，GPT-4 的成绩排名能够进入前 10%(Katz 等, 2024)。

与 GPT-3.5 类似，GPT-4 同样拥有众多版本和不同的上下文窗口配置。GPT-4 的推出，特别是其 GPT-4-32K 版本（支持处理大约 32768 个 tokens，约 24576 个单词），标志着大语言模型在处理大规模文本数据方面的能力有了显著的提升。

[6]GPT-3 官网https://openai.com/blog/gpt-3-apps

[7]Token 及其计算https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

[8]GPT-3.5 详细信息https://platform.openai.com/docs/models/gpt-3-5-turbo

[9]OpenAI 产品计价方案https://openai.com/pricing

[10]Token 计算方式https://platform.openai.com/tokenizer

[11]GPT-4 官网https://openai.com/gpt-4

GPT-4 Turbo 模型的 128K 上下文窗口（支持处理大约 128000 个 tokens，约 96000 个单词），使得它能够处理相当于 300 多页文本的内容，这对于需要深入理解和挖掘长篇文本信息的任务尤为重要[12]。

总体来看，GPT-4 在处理长文本方面具有更强的能力，使其能够处理更长的天文学研究论文，这对于通过广泛的文本上下文来深入理解复杂的天文现象和理论至关重要。此外，GPT-4 先进的推理能力也有助于识别天文文献中复杂且晦涩的实体信息。然而，相比于 GPT-3.5，GPT-4 的 API 价格更加昂贵。

3.2.4　Claude 2

继 GPT-4 在诸多方面实现技术突破之后，人工智能领域再次迎来了一款备受注目的大语言模型——Claude 2[13]。该模型由 Anthropic 公司开发，在诸多方面展现了其卓越的性能，特别是在文本分析、编码、问题推理等方面表现尤为出色 (Caruccio 等, 2024)。

Claude 2 的训练语料库有多种类型的数据，包括大规模的互联网对话数据集、新闻、书籍、结构化表格数据以及大量的跨学科领域报告和代码等，增强了其在不同领域的理解和生成能力。

与 GPT-4 类似，Claude 2 也能够处理超长文本，并且在这方面甚至超越了 GPT-4。Claude 2 的上下文窗口从以前的 9K tokens 扩展到了现在的 200K tokens[14]，尽管目前发布的版本仅支持 100K tokens，但这一能力足以使其成为长文本处理的佼佼者。

这些特点或许使得 Claude 2 能够成为天文领域内的实用性工具，特别是在长篇幅天文文献的知识实体抽取方面。

综上，考虑到这些模型的各版本使用渠道的难易程度以及本文实验的实际情况，最终选择了 Llama-2-70B、GPT-3.5-4K、GPT-4-128K 和 Claude 2-100K。需要注意的是，为了叙述的简洁性，本文将统一使用 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 来指代这些模型的特定版本。

3.3　**实验测试**

本小节通过精心设计的实验来深入探索大语言模型在天文学知识实体提取任务中的潜力。实验选取了 30 篇天文学文献，构建了全文文本和段落文本集合两种数据集，以适应各大语言模型的 token 限制。实验采用四种主流大语言模型，并结合多种组合提示，旨在评估它们在提取天文领域特定实体方面的精确度和全面程度。基于这些实验的结果，本文将尝试说明大语言模型在处理专业天文文献时的优势和挑战，为未来的研究和应用提供指导。

[12]GPT-4 详细信息https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models

[13]Claude 2 官网https://claude.ai/chats

[14]Claude 2 详细信息https://www.anthropic.com/claude

3.3.1 数据集

为了全面评估大语言模型执行实体抽取任务的能力，本文依据如下标准筛选出了一批天文文献。其中，重点是挑选那些包含丰富天文知识实体的文章，例如天体标识符和望远镜名称，以便在实验中提供多样的实体样本。另外，需要确保选择的文章研究主题尽可能多地覆盖天文学各个研究子领域，包括星系、恒星、行星等；同时也需要包含光学、射电和 X 射线等众多观测波段，保证数据集能够充分体现天文学研究的多样性与复杂性。此外，所选文章需要结构、逻辑清晰，内容详尽，这对于模型理解文本语境、准确识别和提取天文知识实体至关重要。通过这些严格的筛选标准，本文旨在构建一个数据集以期能够真正展现大语言模型在天文学文献中执行知识实体抽取任务潜力。

本文从天文领域的多个权威期刊中精心挑选了近十年来的 30 篇文献，这些期刊包括 Astrophysical Journal (ApJ) [15]、Astrophysical Journal Supplement Series（ApJS）[16]、Astronomy & Astrophysics (A&A) [17]、The Astronomical Journal(AJ)[18]、Monthly Notices of the Royal Astronomical Society (MNRAS) [19]和 Research in Astronomy and Astrophysics (RAA) [20]。

然而，尽管 GPT-4 和 Claude 2 具有超长上下文处理能力，它们能够接受的 token 数目足以一次性解析天文文献级别长度的文本，但 Llam2-70B 和 GPT-3.5 目前支持的 token 数目显然无法支持。按照模型支持的最大 token 数目来直接划分文献文本会破坏上下文语义信息，这会限制大语言模型的理解和推理能力。考虑到 Llama-2-70B 和 GPT-3.5 支持的最大 token 数目足以适应文献中的段落长度，因此，本文按照段落的顺序对每篇文献进行了分割，最大程度保持上下文语义的连贯和完整。经过处理，这些文献被分成 20 到 100 个不等的段落，形成 30 个段落文本集合。

综上，本文共构建了两种数据集，即 30 篇文献的全文文本数据集和这些文献的段落文本集合数据集。按照优先提取缩写的原则，本文对其中包含的天体标识符和望远镜名称进行了标注，分别获得 446 项天体标识符和 224 项望远镜名称。这些文献的 DOI 以及具体的标注实体已经上传至中国虚拟天文台论文数据贮藏库，网址为https://doi.org/10.12149/101357，图 3-2展示了其中一个标注示例。

3.3.2 实验设置

本小节从以下几个方面设计了对比实验。图3-3说明了实验流程。图中的“Descriptions（描述）”、“Definitions（定义）”、“Emphasis（重点）”、“Examples（示例）”和“Second Conversation（二轮对话）”分别表示 Prompt-KEE 提示策略中的

[15]ApJ 官网https://iopscience.iop.org/journal/0004-637X

[16]ApJS 官网https://iopscience.iop.org/journal/0067-0049

[17]A&A 官网https://www.aanda.org/

[18]AJ 官网https://iopscience.iop.org/journal/1538-3881

[19]MNRAS 官网https://academic.oup.com/mnras/

[20]RAA 官网https://www.raa-journal.org/

```
“paper 3”: {
        “DOI”: “10.1051/0004-6361/202244417”,
        “entity”: {
            “Telescope name”: [
                “SDSS”, “ROSAT”, “INTEGRAL”, “Swift”, “2MASS”, “Hubble Space Telescope”,
                “James Webb Space Telescope”
            ],
            “Celestial object identifier”: [
                “SDSS J110511.15+530806.5”, “small magellanic cloud”, “RX J0439.6-5311”,
                “IRAS 14026+4341”, “Milky Way”, “PG 1211+143”, “PG 1048+213”, “Mrk 509”,
                “RX J0439.6-5311”, “PG 1115+407”, “Ark 120”, “PG 1448+273”, “NGC 5548”,
                “APM 08279+5255”, “PG 2112+059”, “HS 1603+3820”
            ]
        }
    }
```


图 3-2　文献的天文知识实体标注示例

Figure 3-2　Example of astronomical knowledge entity annotation in literature.

要素“Task Descriptions（任务描述）”、“Entity Definitions（实体定义）”、“Task Emphasis（任务重点）”、“Task Examples（任务示例）”和“Second Conversation（二轮对话）”。

首先，由于任务描述是总体性的，需要将任务描述与其他提示要素相结合，探讨每种提示要素对大语言模型抽取知识实体的影响。这些组合包括：

1）Des_Only：仅总体性的“任务描述”；

2）Des_Def：“任务描述”与“实体定义”的组合；

3）Des_Emp：“任务描述”与“任务重点”的组合；

4）Des_Exa：“任务描述”与“任务示例”的组合；

5）Des_Def_Emp：“任务描述”、“实体定义”和“任务重点”的组合；

6）Des_Def_Emp_Exa：“任务描述”、“实体定义”、“任务重点”和“任务示例”的组合；

7）Des_Def_Emp_Con：“任务描述”、“实体定义”、“任务重点”和“二轮对话”的组合；

8）All：“任务描述”、“实体定义”、“任务重点”、“任务示例”和“二轮对话”的组合。

根据实践经验，研究人员观察到对实体的定义和任务的强调往往对大语言模型的输出有正向且稳定的影响。然而，当包含任务的示例时，目前部分大语言模型的性能可能会表现出许多不确定性 (Zhao 等, 2021; McKenna 等, 2023; Shi 等, 2023)；因此，本文在组合 4）、6）和 8）中分别加入了“任务示例”来验证这种可能性。此外，二轮对话用来再次强调部分的“任务重点”，所以相应地构建组合 7）和 8）。

结合以上八种不同的提示组合，将 30 篇文献的全文文本输入给 GPT-4 和 Claude 2 进行实验；对于 30 篇文献的段落文本集合数据集，同样结合这些组合提示将它们输入到 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 中。需要注意的是，对于 Llama-2-70B 和 GPT-3.5，从段落文本集合数据集中提取的知识实体需要经

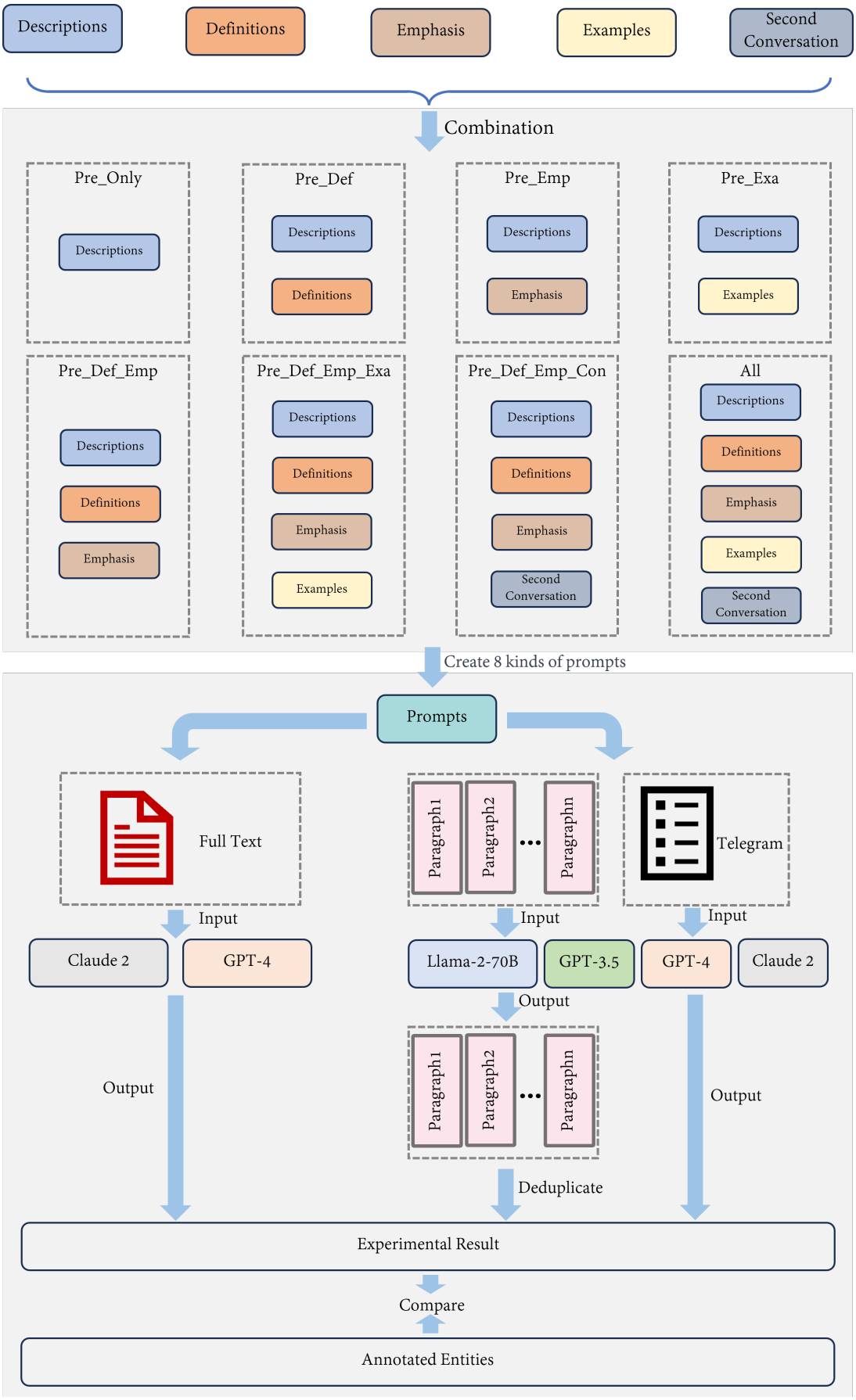

图 3-3　实验流程图

Figure 3-3　Experimental pipeline.

过一系列后处理，包括合并结果、删除重复实体。

最后，将后处理过的实验结果与对应文献中标注的知识实体进行比较。

3.3.3　评价指标

在知识实体抽取任务中，精确率（Precision）、召回率（Recall）和 F1 Score 是确定模型从文本中识别和提取相关实体有效性的三种关键评估指标。

精确率：此标准用来评估抽取正确实体的准确度，也称之为“查准率”。在知识实体抽取任务中，精确率定义为模型准确识别的实体（TP）占提取实体总数（TP + FP）的比例。精确率的提高表明正确识别的实体识别的比例很大，这意味着错误识别的实体数量减少。计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{3-1}$$

召回率：此标准衡量了模型能够抽取所有正确实体的能力，也称为“查全率”。具体来说，它是模型成功提取的实体（TP）与实际存在于文本中应被提取的实体总数（TP + FN）之间的比例。表现出高召回率的模型表明它能够全面地提取大部分应该提取实体，最大限度地减少遗漏识别的可能性。计算公式为：

$$Recall = \frac{TP}{TP + FN} \tag{3-2}$$

F1 Score：此标准是**精确率**和**召回率**的调和平均值。作为一个综合的评价指标，它同时考虑了模型正确识别实体的能力和模型识别到所有相关实体的能力。在理想情况下，希望模型在精确度和召回率两个性能指标上都能表现出色，即同时具有高精确度和高召回率。然而，在实际应用中，往往需要在这两个指标之间做出权衡。这是因为在提高一个指标的同时，可能会降低另一个指标的性能。计算公式为：

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3-3}$$

3.4　结果与分析

本小节结合实验结果全面分析 Prompt-KEE 提示策略的有效性和大语言模型提取天文知识实体能力。本小节分别对文献全文文本数据集和段落文本集合数据集两个部分的实验结果进行分析。

3.4.1　全文文本数据集实验结果与分析

表3-1展示了 GPT-4 和 Claude 2 在八种不同的组合提示下，从 30 篇文章的全文文本中提取天体标识符和望远镜名称两种知识实体的结果。此外，在图3-4中

表 3-1　GPT-4 和 Claude 2 分别使用八种组合提示从 30 篇文献的全文中提取天体标识符和望远镜名称两种知识实体的结果

Table 3-1　The results of GPT-4 and Claude 2 in extracting two kinds of knowledge entities, celestial object identifier and telescope name, 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

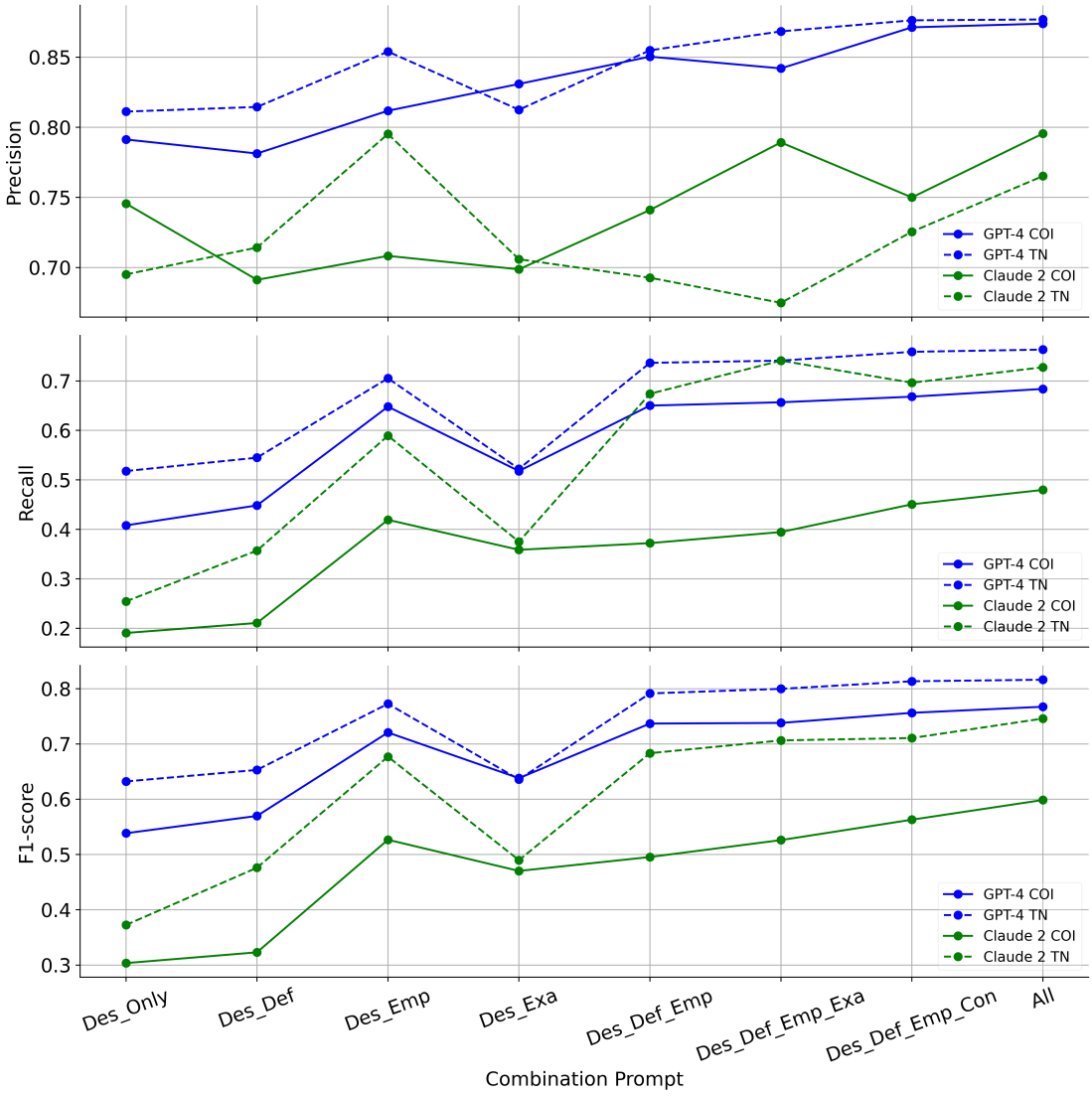

图 3-4　GPT-4 与 Claude 2 在全文文本数据集中提取天体标识符和望远镜名称两种天文知识实体的精确率、召回率和 F1 Score 的比较

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

分别比较了 GPT-4 和 Claude 2 抽取的精确率、召回率和 F1 Score。在对这些结果的比较分析中，本小节给出了几个关键的结论。

首先，GPT-4 在三项评价指标上始终优于 Claude 2。特别是在精确率和 F1 Score 中，GPT-4 领先的优势更为明显。这表明 GPT-4 在提取天文文献中的天体标识符和望远镜名称时，能够更准确地识别相关的实体，并且能够在精确率和召回率之间取得较好的平衡。相比之下，尽管 Claude 2 模型在任务中也表现出了一定的能力，但它在精确率方面的表现不如 GPT-4，这意味着 Claude 2 误识别了更多的实体。

其次，观察表3-1和图3-4可以发现，在这项任务中，5 种提示要素总体上都能给大语言模型的抽取任务带来提升作用，其中“任务强调”带来的提升效果最为明显。

再者，还可以发现多样化的组合提示也能提升 GPT-4 和 Claude 2 的性能。当“任务描述”与其他提示要素相结合时，尤其是加入“任务重点”提示要素，相比于更简单的组合提示，能够显著提高模型的提取效果。GPT-4 模型在这方面表现得尤为出色，它能够有效地利用包含多种要素的提示，在“All”组合提示中同时获得最高的精确率、召回率和 F1 Score。Claude 2 虽然也能从更详细的提示中获得性能提升，并且也在“All”组合提示中三项指标表现最好，但其提升的幅度并没有 GPT-4 显著。

此外，从图3-4中可以明显地观察到，除在“Des_Exa”组合提示下，两种模型识别望远镜的效果和识别天体标识符的效果较为接近之外，其它情况下识别望远镜的效果要远高于识别天体标识符的效果。另外，比较天体标识符和望远镜名称的提取效果可以发现，GPT-4 在这两种实体上的表现比较相近，而 Claude 2 则表现出更多的差异。具体来说，GPT-4 提取望远镜名称的效果略好，但是总体上较为均衡；而 Claude 2 提取望远镜名称的效果总体上远远好于提取天体标识符的效果，尤其是召回率和 F1 Score。

最后，在使用组合提示的情况下，召回率的提升幅度远高于精确率。这表明，通过精心设计的提示，模型在不遗漏重要实体信息方面表现出的效果更为显著。

3.4.2 段落文本集合数据集实验结果与分析

表3-2展示了 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 四种大语言模型从 30 篇文献的段落文本集合数据集中抽取天体标识符和望远镜名称两种知识实体的结果。在图3-5中，分别比较了其精确率、召回率和 F1 Score 三种评价指标。在对这些结果的比较分析中，本小节给出了几个关键的结论。

Llama-2-70B 在召回率方面表现较好，尤其是在识别望远镜名称方面。然而，该模型在精确率方面表现较差，例如“DES_Only”和“Des_Exa”两项组合提示用来识别天体标识符时仅有 0.0450 和 0.0320，这导致其 F1 Score 被大幅度拉低。这表明 Llama-2-70B 虽然能够正确识别出大量的相关知识实体，但同时也包含较多被错误抽取的知识实体。而且，这种情况在所有组合提示中都存在，说明

表 3-2　Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 分别使用八种组合提示从 30 篇文章的段落集合中提取天体标识符和望远镜名称两种知识实体的结果

Table 3-2　The results of Llama-2-70B, GPT-3.5, GPT-4 and Claude 2 in extracting two kinds of knowledge entities, celestial object identifier and telescope name, 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

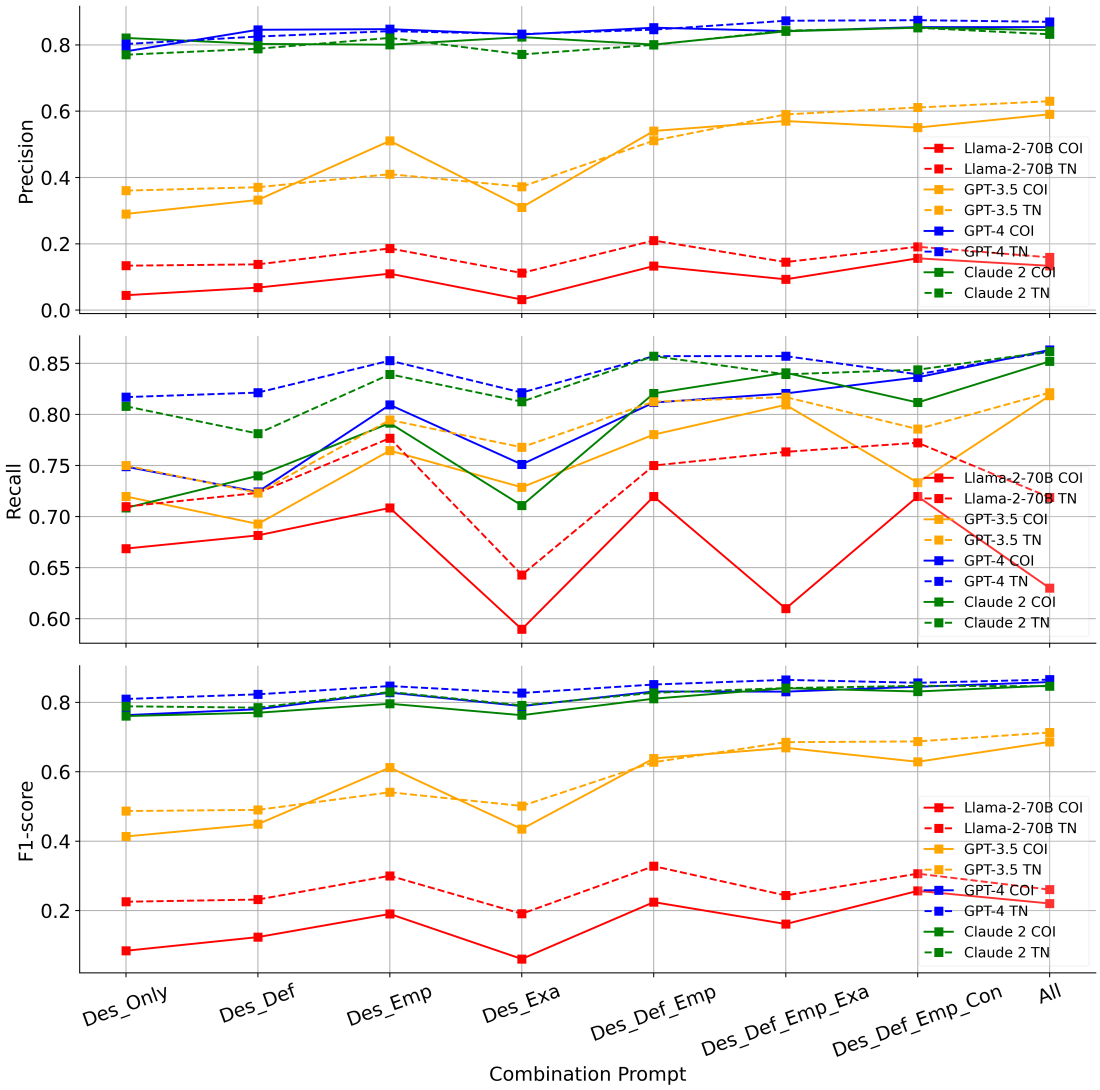

图 3-5　Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 在段落文本集合数据集中分别提取天体标识符和望远镜名称两种天文知识实体的精确率、召回率和 F1 Score 的比较

Figure 3-5　Comparison of precision, recall, and F1 Score for Llama-2-70B, GPT-3.5, GPT-4 and Claude 2 in extracting celestial object identifiers (COI) and telescope names (TN) from the paragraph collections.

Llama-2-70B 在天文知识实体提取任务上存在一定的局限性。

此外，Llama-2-70B 在使用包含“任务示例”时表现出的性能，证实了示例可能会给模型的性能带来不确定性，例如组合提示“Des_Exa”的 0.0320 和 0.1121 相比于“Des_Only”的 0.0450 和 0.1341 都较低。这种现象主要是因为带有示例的提示可能会引导模型过分依赖示例中提供的信息，导致模型在实际处理文本时，更倾向于选择示例中给出的实体作为输出，而忽略或未能充分关注文本中真实存在的实体信息。结果就是，模型可能会错误地输出一些和文本不相关的实体，或者遗漏一些实际存在的相关实体，从而影响了实体提取的整体效果。这表明在使用带有示例的提示时，需要考虑具体的使用场景，避免对模型的判断产生负面影响。

GPT-3.5 相较于 Llama-2-70B，在精确率和 F1 Score 两种指标上有着明显的进步。特别是在使用“All”组合提示时，GPT-3.5 的性能达到了最佳状态，F1 Score 分别为 0.6861 和 0.7131；这表明该模型能够有效地利用提示中的信息来完成知识实体提取任务。这也能够说明，与 Llama-2-70B 相比，GPT-3.5 在保证提取结果准确性的同时，也能兼顾提取结果的全面性。

Claude 2 虽然在总体性能上略微落后于 GPT-4，但它的表现依然出色。与 Llama-2-70B 和 GPT-3.5 相比，Claude 2 无论是识别天体标识符还是望远镜名称，在精确率和 F1 Score 上都有着更高的分数。与其它模型类似，在使用“All”组合提示时，Claude 2 的 F1 Score 分别达到了最高的 0.8520 和 0.8616，这表明它拥有理解复杂提示的出色能力。这些结果表明，Claude 2 对文本有着更深层次的理解能力，并在仔细的提示下能完成更加细致的实体提取任务。

3.4.3　实验结果总体分析

通过综合分析表3-1和表3-2的 F1 Score，可以发现 Prompt-KEE 提示策略能够显著激活了四种大语言模型识别天体标识符和望远镜名称两种天文知识实体的能力，尤其是“任务重点”提示要素的加入会大幅度提升模型性能。

此外，四种大语言模型识别望远镜名称的效果总体上要优于识别天体标识符的效果。这种差异可能归因于望远镜名称往往伴随着一些特定的、具有高辨识度的词汇出现，例如“telescope”、“survey”等，这些词汇有助于模型更准确地理解和判断这类实体。另外，望远镜名称的数量也相对有限，并且这些名称或许已经被这些模型的预训练知识库中所涵盖，因此模型能够更加精准、全面地识别和提取。相对而言，天体标识符的数量庞大，并且随着观测的不断开展还在快速积累，再加之其命名方式多样，当这些天体标识符出现在天文文献中会对大语言模型构成了巨大的挑战。因此，模型在识别和提取天体标识符时可能会遇到更多的困难，从而出现其效果不如望远镜名称抽取效果的情况。

图3-6比较了 GPT-4 和 Claude 2 在全文文本数据集和段落文本集合数据集中提取天体标识符和望远镜名称的三种评估指标。可以观察到 GPT-4 和 Claude 2 在两种不同长度的文本数据集上的表现差异较大。具体来说，两种模型在文献全文

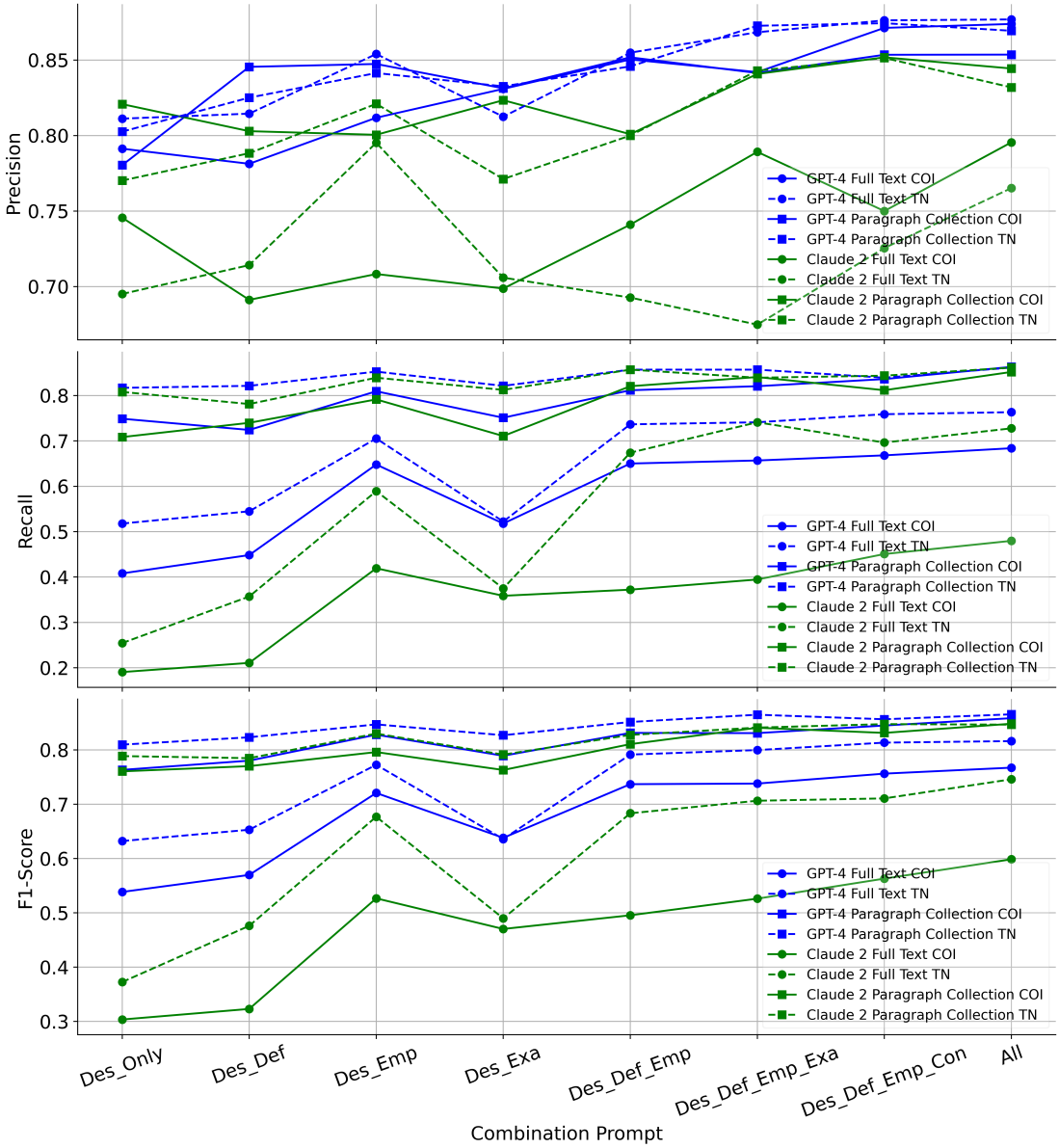

图 3-6　GPT-4 和 Claude 2 在全文文本数据集和段落文本集合数据集中分别提取天体标识符和望远镜名称两种天文知识实体的精确率、召回率和 F1 Score 的比较

Figure 3-6　Comparison of precision, recall, and F1 Score for GPT-4 and Claude 2 in extracting celestial object identifiers (COI) and telescope names (TN) from the full texts and paragraph collections.

文本数据集中一直保持较高的精确率，而召回率会随着提示的丰富程度提升幅度较大；但在段落文本集合数据集中两种模型均保持较高的精确率和召回率。例如在识别望远镜时，GPT-4 在全文文本数据集中的精确率从“Des_Only”的 0.8112 只提升到了“All”中的 0.8769，幅度为 0.0657；而其召回率由 0.5179 提升到了 0.7634，幅度为 0.2455，大约是精确率提升幅度的 4 倍。但在段落文本集合数据集中，GPT-4 的精确率从“Des_Only”的 0.8026 只提升到了“Des_Def_Emp_Con”中的 0.8744，幅度为 0.0718；其召回率由“Des_Only”的 0.8170 也只提升到了“All”中的 0.8616，幅度为 0.0446。这些差异可能主要归因于以下因素：

1) **上下文信息**：文献中完整的上下文信息对于 GPT-4 和 Claude 2 在执行天文知识实体提取任务时理解与天体标识符和望远镜名称相关的语义细节至关重要。然而，与文献全文文本相比，文献的段落文本提供的上下文范围更窄，大语言模型能够获取的上下文信息相对较少。尽管如此，由于 GPT-4 和 Claude 2 拥有巨大且多样训练数据，可能已经学习到了大量与文本主题相关的知识，所以它们仍然能够保证较高的精确率。这意味着即使在上下文信息较少的情况下，这些模型仍然能够有效地执行实体提取任务。

2) **实体分布**：文献全文文本中的知识实体分布通常较为稀疏，而在段落文本中相对更为集中。这种分布密度上差异会影响模型识别实体的能力。具体来说，在全文文本中实体分布较广并且伴随着大量冗余信息，这可能导致模型会忽略某些实体，从而降低了召回率；而在段落文本中，由于实体更为集中且冗余信息较少，模型更容易提取到潜在的实体，这有利于维持较高的召回率。

文献全文文本和段落文本的这些不同特点会较大程度影响大语言模型的处理信息的难度和注意力分配，进而导致它们在两种不同长度的文献文本中提取天文知识实体的表现有所差异。

3.5　本章小结

本章探索了大语言模型执行天文知识实体抽取任务的潜力。实验采用了四种主流大语言模型，即 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2，并构建了 30 篇天文文献的全文文本和段落文本集合两种数据集进行测试。通过设计的 Prompt-KEE 策略指导四种大语言模型从两种数据集中提取天体标识符和望远镜名称。结果表明，结合 Prompt-KEE 策略的大语言模型在知识实体抽取任务上表现出显著的潜力。最后，本章节详细分析了实验结果，并给出了合理的见解。

第 4 章　其它实体抽取方法与大语言模型方法对比

为了进一步了解大语言模型执行天文知识实体任务的优势，本章节尝试探索其它实体提取方法在实体提取过程中的效果，并将其与第3章节的实验结果进行详细比较。这些方法是1.2.3小节天文领域知识实体抽取研究现状中提到的基于规则的、基于机器学习的、以及基于小规模预训练语言模型的方法。

4.1　其它方法介绍

4.1.1　基于规则的方法

截止到 2010 年，DJIN 系统在《天体命名词典》的基础上构建了一个包含超过 50000 个正则表达式的集合，旨在从文本中提取尽可能多的天体标识符。遗憾的是，SIMBAD 并没有提供 DJIN 的访问入口以及这些正则表达式的获取方式，这意味着利用它们来完成测试实验的想法是不可行的。此外，虽然全球各类天文观测设备及在此基础上开展的的巡天计划在数量上是有限且可以被完全统计的，但目前并没有一个详尽且权威的列表可供参考。

因此，本文借鉴了Lesteven 等 (2010) 的开发经验，设计了一套包含 116 条常规天体标识符 (如 LAMOST J151003.74 + 305407.3、NGC 1866) 的提取规则，表4-1给出了相关的正则表达式匹配示例。而望远镜名称 (如 Hubble Space Telescope、Arecibo Telescope) 的提取，本文则采用字典匹配的方法，其中包含 53 条字典项。

表 4-1　正则表达式匹配天体标识符示例

Table 4-1　Regular expression matching examples for celestial object identifiers.

正则表达式	匹配字符串
LAMOST\s+J\d{6}\.\d+\s+[+-]+\d{6}\.\d+	LAMOST J075807.54 − 043205.3
[A-Za-z]+\s?\d+	NGC 1866、HZ 3、Ton 927、HD 213893
TYC\s+\d+[-]\d+[-]\d+	TYC 4895-599-1、TYC 170-779-1
[A-Za-z]+\s+\d+\s*[+-]\s*\d+	KPD 0033 + 5229、IRAS 14026 + 4341、PG 1115 + 407

4.1.2　基于机器学习的方法

本文选取了基于机器学习方法中的最大熵模型（MaxEnt）来进行实验对比，它是知识实体抽取研究领域中最为经典的方法之一。

利用该模型有效抽取天文知识实体的前提是必须提供充足的高质量、带标注的天文文本数据。因此，本文选择包含天体标识符和望远镜名称标注的 DEAL 数据集作为训练语料库。需要注意的是，在此数据集中，“Celestial Object”（包含 10359 项）实体类别对应于本文的 “celestial object identifier”，即天体标识符；

而“Telescope”（包含 5506 项）和“Survey”（包含 3430 项）实体类别对应于本文的“telescope name”，即望远镜名称。通过这种方式，确保了模型训练以及后续实验的统一。有关此数据集的详细信息可在其相关的发布[1] [2]中获取，并且也可以在 HuggingFace 网站[3]中公开获取此数据集。

本文使用 nltk（Natural Language Toolkit, 自然语言工具包）库[4]中的 MaxentClassifier 类和其它相关函数完成了模型的训练和最终的实验测试工作。

4.1.3 基于小规模语言模型的方法

受到Alkan 等 (2022) 工作的启发，本文选择了 PyTorch HuggingFace 的 SciBERT 模型 (Beltagy 等, 2019) 的 scibert_scivocab_case 版本[5]作为基于小型语言模型的方法来进行实验对比。由于 scibert_scivocab_case 版本的模型在大量科学论文的语料库上进行了专门的预训练，并且使用了针对科学领域优化的词汇表，从而能够更好地理解和捕捉科学术语和概念。因此，对此版本模型进行微调，可以使其进一步被优化，能够更加适应天文文献的知识实体抽取任务。

在微调阶段，本文继续使用 DEAL 数据集 (Grezes 等, 2022) 来完成训练工作。并且为了适应 BERT 模型的最大文本序列长度的限制（通常是 512 个 tokens），本文利用了滑动窗口策略。在训练阶段，本文设置了 30 次训练轮数，确保模型较为完整地学习 DEAL 数据集中的特征。

4.2 对比实验与结果分析

由于这些方法能够处理的文本长度均相对较短，为了保证它们的提取结果可以与四种大语言模型的提取结果进行比较，本章节统一选择使用 30 篇文献的段落文本集合数据集来对它们各自进行实验。此外，与大语言模型从段落文本集合数据集中提取知识实体的过程类似，这些方法在完成段落文本实体抽取后也需要进行合并、去重处理。

对于四种大语言模型，本文从表3-2中选取了它们各自最高 F1 Score 所在的组合提示实验结果来作为比较基准。表 4-2展示了 Llama-2-70B、GPT-3.5、GPT-4、Claude 2 以及基于规则（Rule）、最大熵模型（MaxEnt）和 SciBERT 三种方法的性能对比。图4-1和图4-2是这些方法分别抽取天体标识符和望远镜名称性能的对比。以下是对它们的性能和具体抽取表现进行了分析。

性能：从表3-2的 F1 Score 可以观察出，基于规则的方法和最大熵模型的整体性能明显低于四种大语言模型。相比之下，尽管 SciBERT 在两类实体上的召回率明显低于 Llama-2-70B，但是其精确率和 F1 Score 更具优势。具体来说，SciBERT

[1] DEAL 共享任务的标签定义https://ui.adsabs.harvard.edu/WIESP/2022/LabelDefinitions

[2] DEAL 数据集详细信息https://ui.adsabs.harvard.edu/blog/ads-models-and-datasets

[3] DEAL 数据集下载地址https://huggingface.co/datasets/adsabs/WIESP2022-NER

[4] 自然语言工具包介绍https://www.nltk.org/

[5] SciBERT 模型https://github.com/allenai/scibert

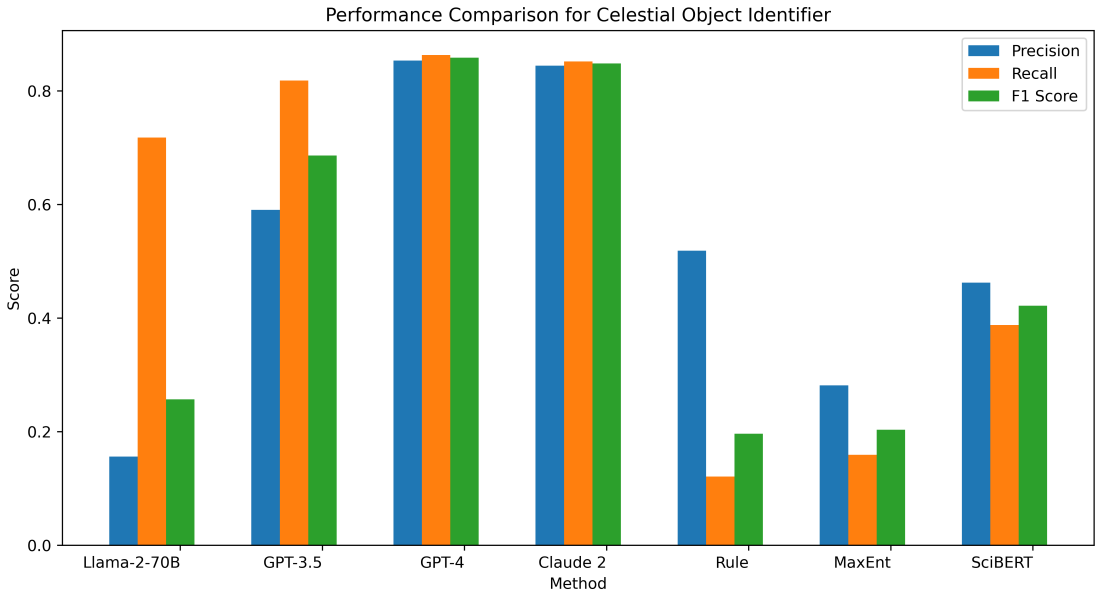

图 4-1　四种大语言模型和其他方法抽取天体标识符的性能对比

Figure 4-1　Performance comparison of four large language models and other methods for extracting celestial identifiers.

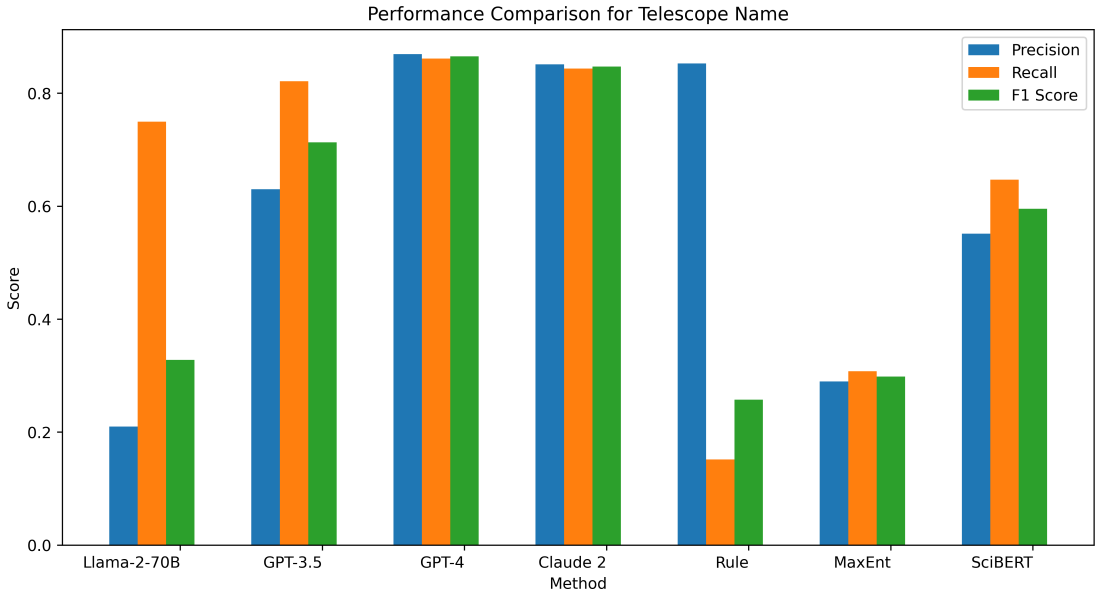

图 4-2　四种大语言模型和其他方法抽取望远镜名称的性能比较

Figure 4-2　Performance comparison of four large language models and other methods for extracting telescope names.

表 4-2　四种大语言模型和其他方法在从段落集合中提取天体标识符和望远镜名称两种天文知识实体中的性能比较

Table 4-2　Performance comparison of four large language models and other methods in extracting two kinds of astronomical knowledge entities, celestial object identifier and telescope name, from paragraph collections.

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

在提取天体标识符和望远镜名称方面分别比 Llama-2-70B 高出 0.1650 和 0.2675。在4.1节其它方法介绍中，本文说明了相应的重要训练数据集由于一些客观因素而无法获取，所以需要明确的是，这些方法抽取天文知识实体的性能可能并没有达到与 DJIN 等成熟系统相当的水平。因此，本文将不过多关注大语言模型和其它知识实体抽取方法之间的性能差异。相反，本文重点需要关注它们在抽取过程中具体的关键性差异上。这些差异如表4-3和表4-4所示。以下是对实验结果的详细分析。

1) 通过对表4-3和表4-4中的 6 项差异对比分析，可以发现大语言模型在识别天体标识符和望远镜名称两种天文知识实体的能力优于其他方法。

2) 从所有的差异中可以看出，大语言模型在处理知识实体边界（如复合名词或缩写）方面表现更好。例如，“Small Magellanic Cloud”是一个复合型实体，三个单词结合才能指代一个特定的星系。基于规则的方法需要在字典中包含相应的复合型实体才能识别出这些结构，而最大熵模型和 SciBERT 在处理它们时同样也会遇到困难。然而，即使“Small Magellanic Cloud”、“LAMOST J0037 + 4016”跨越多个单词或数字，四种大语言模型也都能将它们识别为一个整体。

3) 表4-3中的差异 2 和表4-4的差异 2 表明，大语言模型具有更好的泛化能力。尽管文本中的知识实体不常见或者完全新颖，但是大语言模型擅长通过理解上下文信息以及实体模式来判断未知实体的类别，这使得它们在面对新文本时表现得更加出色。相比之下，其他方法很难应对这种情况。

4) 大语言模型在实体消歧方面同样表现出色。在表4-3的差异 2 中，“Andromeda galaxy”和“M31”表示的是同一个星系。大语言模型能够判断出它们之间的关系，并将它们以“Andromeda galaxy (M31)”形式作为一个整体抽取。在

表 4-3　四种大语言模型和其他方法在从段落集合中抽取天体标识符关键结果比较

Table 4-3　Comparison of key results between four large language models and other methods in extracting celestial object identifiers from paragraph collections.

Number	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 等, 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	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 等, 2018)	**Expected**	[J1334]
		Llama-2-70B	[J1334]
		GPT-3.5	[J1334]
		GPT-4	[J1334]
		Claude 2	[J1334]
		Rule	[]
		MaxEnt	[]
		SciBERT	[]
3	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 等, 2019b)	**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]
		SciBERT	[LAMOST J0037+4016, Andromeda galaxy, M31]

表 4-4　四种大语言模型和其他方法在从段落集合中抽取望远镜名称的关键结果比较

Table 4-4　Comparison of key results between four large language models and other methods in extracting telescope names from paragraph collections.

Number	Sentence	Method	Output
1	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 等, 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]
2	The first time was on 2014 November 15 and 16. We used the 60 cm reflecting telescope to perform R-band photometry. The second time was on 2019 January 17 and 18. (Lu 等, 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]
		Rule	[cm reflecting telescope]
		MaxEnt	[reflecting telescope]
		SciBERT	[60 cm reflecting telescope]
3	…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 等, 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, Planck]
		MaxEnt	[Swift, LAMOST, Fermi, Planck]
		SciBERT	[XMM-Newton, Swift, Giommi, Fermi, Planck]

表4-4的差异 3 中，尽管同时存在人名（Page, Grupe 和 Giommi）和以人名命名的望远镜名称 (XMM-Newton, Swift, Fermi, Planck)，但四种大语言模型依然能够准确区分它们。更重要的是，天体标识符 “LAMOST J1131 + 3114” 中包含的望远镜名称 “LAMOST” 也被正确排除。尽管 “LAMOST” 和 “J1131 + 3114” 构成了一个整体，但是并没有传达 “望远镜” 的相关含义。然而，其他方法在这些方面的表现都不令人满意。

抽取模式：模型在实体提取上的性能差异与这些方法各自的工作模式密切相关。大语言模型已经在预训练阶段学习了大量的领域背景知识，因此它们在执行任务时只需要很少的提示信息就可以完成对天体标识符和望远镜名称的抽取。相比之下，基于规则的实体提取方法严重依赖于预定义的规则集和字典进行模式匹配，这导致其提取未知或复杂实体的能力极其有限。最大熵模型和 SciBERT 似乎对这些问题有所缓解，但它们需要大规模的高质量天文文本来增强抽取实体的能力。相对于大语言模型的简单快捷的抽取模式，这些规则设计、数据集构建等都属于劳动密集型任务，因此，其它方法在这些方面缺乏竞争力。

更新和维护：尽管本文利用这些方法完成了知识实体抽取的测试实验，但在实际应用中，方法或者模型的更新和维护是一项必不可少的任务。基于规则的方法和最大熵模型需要定期更新，以适应抽取新的天文知识实体，这个过程需要耗费大量人力。SciBERT 受益于其在已经通过广泛的科学文本完成预训练，需要更新的频率可能更低，但仍然需要阶段性的微调。相比之下，大语言模型通常由专业公司更新和维护，这相应减少甚至避免了用户维护大语言模型的需要，为天文学中的实体抽取任务提供了可持续性的解决方案。

4.3　本章小结

本章节对基于规则的、基于机器学习的、基于小规模语言模型的三种其它方法进行了探讨，并与 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 四种大语言模型的实验结果进行了对比。通过对比实验，本章发现大语言模型在天文知识实体抽取任务中表现出显著的优势，尤其是在处理复杂实体边界、实体消歧以及泛化能力等方面。此外，本文还分析了这些方法在工作模式、更新和维护方面的特点，指出了它们在实际应用中可能面临的挑战以及大语言模型具备的优势。

第 5 章　总结与展望

随着天文学的不断发展，天文数据的迅速积累也进一步促进了天文文献数量的持续攀升。这些文献中包含的天文数据以及对数据的深入理解，对于天文学研究至关重要。然而，当前天文数据与文献之间较低的关联程度给天文学研究造成了诸多不便。天文知识实体作为天文数据与文献的关键纽带，是实现天文数据与文献的关联融合的基本要素。因此，在大规模天文数据与文献的背景下，天文领域对知识实体抽取技术提出了新要求。传统的实体抽取方法在面对复杂和多样化的天文知识实体时显得力不从心，因此，探索更加强大的实体抽取技术势在必行。当前大语言模型在各种自然语言处理任务中表现出强大的能力，因此本文通过一系列的实验探索了大语言模型在天文知识实体抽取任务上的潜力。本文研究内容主要包括以下几个部分：

为了指导大语言模型在天文文献中能够更加有效地抽取相关知识实体，本文提出了 Prompt-KEE 提示策略，其中包含任务描述、实体定义、任务重点、任务示例和二轮对话五项提示要素。

通过遵循 Prompt-KEE 提示策略，本文设计出一组具体的提示要素，并将这些具体的提示要素进行重点组合，构建出了八种组合提示。为了深入研究 Prompt-KEE 提示策略的有效性，本文根据任务需求和模型特点选取了四种大语言模型进行实验，分别为 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2。此外，本文还选取了天体标识符和望远镜名称两种典型的天文知识实体作为抽取对象。在此基础之上，本文按照严格的标准挑选了 30 篇天文文献，并对其中的天体标识符和望远镜名称进行了标注。不仅如此，为了适应不同大语言模型的 token 限制，本文利用 30 篇文献构造了文献全文文本和段落文本集合两种数据集。

利用八种组合提示，本文首先使用 GPT-4 和 Claude-2 对全文文本数据集进行了测试，然后使用四种大语言模型对段落文本集合进行了测试。实验结果表明，大语言模型执行天文知识实体抽取任务的潜力显著，但在性能上存在一些差异。针对这些差异，本文给出详细的分析和见解。

最后，本文对基于规则、基于机器学习和基于小规模语言模型的三种其它实体提取方法进行了探讨，并与 Llama-2-70B、GPT-3.5、GPT-4 和 Claude 2 四种大语言模型的实验结果进行了对比。通过对比实验，本文发现大语言模型在处理复杂实体边界、实体消歧、泛化能力、工作模式、更新和维护等诸多方面表现出显著优势。

本文的研究验证了大语言模型在执行天文知识实体抽取任务的潜力和优势，同时也表明了 Prompt-KEE 提示策略指导大语言模型的有效性。基于本文的工作，未来仍可在多个方向进行深化应用。比如，研究大语言模型对于其它更复杂、更专业类型的实体抽取效果；研究如何从天文文献的表格和图片中更准确地提取天文知识实体。此外，不同模型和任务场景，可能需要不同的提示优化策略。以

下是几项可能值得未来进一步探索的研究：

1) 领域知识是增强大语言模型理解、推理和泛化能力的关键；因此，用广泛的天文学知识来训练大语言模型，可以进一步提高它们提取天文知识实体的能力。

2) 更丰富的提示信息不一定能够提升模型的性能，包含任务示例的提示可能会影响大语言模型提取实体的效果，这表明提示需要基于特定模型和场景进行仔细设计。因此，进一步探索和优化 Prompt-KEE 策略是必要的。

3) 文献全文文本数据集和段落文本集合数据集的提取结果存在明显差异，这表明数据集的特征会影响模型的注意力分配，进而影响模型的提取性能。如何引导模型在两种不同长度的文本中表现更好，是值得进一步研究的问题。

4) 天文学家习惯以结构化的表格来呈现天体标识符等专业知识实体，而现有的大多数模型缺乏从文献中提取出结构化表格数据的能力。未来可以尝试利用 OCR（Optical Character Recognition，光学字符识别）等技术专门解决这类问题。

作为中国虚拟天文台的研究成果之一，本文所尝试的天文知识实体抽取方法在未来将以具体的应用融入到国家天文科学数据中心的天文文献服务系统和天文异构数据融合系统中，为天文数据管理和知识服务领域发挥作用。

本文的研究内容未来可能不仅仅促进天文数据与文献关联检索方面的研究，还会帮助天文数据分析、知识图谱构建、知识发现、天文设施的文献产出自动化评估以及其它天文学研究更好地开展。

参考文献

刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(03): 582-600.

刘春丽, 陈爽. 科学文献中的知识实体抽取与评价研究综述 [J]. 现代情报, 2023, 43: 143-163.

刘浏, 王东波. 命名实体识别研究综述 [J]. 情报学报, 2018, 37(03): 329-340.

刘胜宇. 生物医学文本中药物信息抽取方法研究 [D]. 哈尔滨工业大学, 2016.

化柏林. 针对中文学术文献的情报方法术语抽取 [J]. 现代图书情报技术, 2013: 68-75.

史永刚. 结合机器学习方法的命名实体识别研究 [D]. 电子科技大学, 2006.

吉旭瑞, 魏德健, 张俊忠, 等. 中文电子病历信息提取方法研究综述 [J]. 计算机工程与科学, 2024, 46: 325-337.

吴阳. 财经领域命名实体识别方法的研究与系统实现 [D]. 哈尔滨工业大学, 2015.

孙聪. 面向生物医学文献的化学物蛋白质关系抽取研究 [D]. 大连理工大学, 2021.

崔鑫, 王琰, 侯小刚, 等. 基于词汇增强的典型文物命名实体识别算法 [J/OL]. 中国传媒大学学报 (自然科学版), 2023, 30: 51-55. DOI: 10.16196/j.cnki.issn.1673-4793.2023.02.001.

彭嘉毅, 方勇, 黄诚, 等. 基于深度主动学习的信息安全领域命名实体识别研究 [J]. 四川大学学报 (自然科学版), 2019, 56: 457-462.

时宗彬, 乐小虬. 基于本地大语言模型和提示工程的材料信息抽取方法研究 [J]. 数据分析与知识发现, 2023: 1-13.

咎红英, 关同峰, 张坤丽, 等. 面向医学文本的实体关系抽取研究综述 [J/OL]. 郑州大学学报(理学版), 2020, 52: 1-15. DOI: 10.13705/j.issn.1671-6841.2020190.

曹树金, 岳文玉. 基于深度学习的中共党史文献命名实体识别研究 [J]. 情报资料工作, 2022, 43: 81-88.

李广建, 袁钺. 基于深度学习的科技文献知识单元抽取研究综述 [J]. 数据分析与知识发现, 2023, 7: 1-17.

李洋, 蔡红珍, 邢林林, 等. 基于对抗迁移的复合材料检测领域命名实体识别 [J]. 科学技术与工程, 2022, 22: 13370-13377.

杨培, 杨志豪, 罗凌, 等. 基于注意机制的化学药物命名实体识别 [J]. 计算机研究与发展, 2018, 55: 1548-1556.

杨巍. 基于深度学习的病历命名实体识别研究 [D]. 北京交通大学, 2021.

温雯, 伍思杰, 蔡瑞初, 等. 面向专业文献知识实体类型的抽取和标注 [J]. 中文信息学报, 2018, 32: 102-115.

章成志, 谢雨欣, 宋云天. 学术文本中细粒度知识实体的关联分析 [J]. 图书馆论坛, 2021, 41: 12-20.

耿飙, 梁成全, 魏炜, 等. 基于深度学习的非结构化医学文本知识抽取 [J/OL]. 计算机工程与设计, 2024, 45: 177-186. DOI: 10.16208/j.issn1000-7024.2024.01.023.

许华, 刘茂福, 姜丽, 等. 基于语言规则的病症菌实体抽取 [J/OL]. 武汉大学学报 (理学版), 2015, 61: 151-155. DOI: 10.14188/j.1671-8836.2015.02.008.

郭喜跃, 何婷婷. 信息抽取研究综述 [J]. 计算机科学, 2015, 42: 14-17+38.

陈基. 命名实体识别综述 [J]. 现代计算机 (专业版), 2016: 24-26.

陈祥, 张仰森, 李尚美, 等. 面向计算机科学领域的专业实体识别 [J]. 重庆理工大学学报 (自然科学), 2023, 37: 205-212.

韩玉民, 郝晓燕. 基于子词嵌入和相对注意力的材料实体识别 [J]. 计算机应用, 2022, 42: 1862-1868.

马娜, 张智雄, 吴朋民. 基于特征融合的术语型引用对象自动识别方法研究 [J]. 数据分析与知识发现, 2020, 4: 89-98.

Accomazzi A. Decades of transformation: Evolution of the nasa astrophysics data system's infrastructure [J]. arXiv preprint arXiv:2401.09685, 2024.

Accomazzi A, Eichhorn G, Kurtz M J, et al. The nasa astrophysics data system: Architecture [J]. Astronomy and Astrophysics Supplement Series, 2000, 143(1): 85-109.

Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774, 2023.

Akras S, Guzman-Ramirez L, Leal-Ferreira M L, et al. A census of symbiotic stars in the 2mass, wise, and gaia surveys [J]. The Astrophysical Journal Supplement Series, 2019, 240(2): 21.

Al-Moslmi T, Ocaña M G, Opdahl A L, et al. Named entity extraction for knowledge graphs: A literature overview [J]. IEEE Access, 2020, 8: 32862-32881.

Alkan A K, Grouin C, Schüssler F, et al. A majority voting strategy of a scibert-based ensemble models for detecting entities in the astrophysics literature (shared task) [C]//First Workshop on Information Extraction from Scientific Publications. 2022.

Amatriain X. Prompt design and engineering: Introduction and advanced methods [J]. arXiv preprint arXiv:2401.14423, 2024.

Anthony L F W, Kanding B, Selvan R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models [J]. arXiv preprint arXiv:2007.03051, 2020.

Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.

Baines D, Giordano F, Racero E, et al. Visualization of multi-mission astronomical data with esasky [J]. Publications of the Astronomical Society of the Pacific, 2016, 129(972): 028001.

Baum L E, Petrie T. Statistical inference for probabilistic functions of finite state markov chains [J]. The annals of mathematical statistics, 1966, 37(6): 1554-1563.

Baum L E, Petrie T, Soules G, et al. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains [J]. The annals of mathematical statistics, 1970, 41 (1): 164-171.

Beltagy I, Lo K, Cohan A. Scibert: A pretrained language model for scientific text [J]. arXiv preprint arXiv:1903.10676, 2019.

Bian J, Zheng J, Zhang Y, et al. Inspire the large language model by external knowledge on biomedical named entity recognition [J]. arXiv preprint arXiv:2309.12278, 2023.

Bisercic A, Nikolic M, van der Schaar M, et al. Interpretable medical diagnostics with structured data extraction by large language models [J]. arXiv preprint arXiv:2306.05052, 2023.

Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens [C]//International conference on machine learning. PMLR, 2022: 2206-2240.

Brants T, Popat A, Xu P, et al. Large language models in machine translation [C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007: 858-867.

Brown T, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

Bsharat S M, Myrzakhan A, Shen Z. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4 [J]. arXiv preprint arXiv:2312.16171, 2023.

Byrne K. Nested named entity recognition in historical archive text [C]//International Conference on Semantic Computing (ICSC 2007). IEEE, 2007: 589-596.

Carbon D F, Henze C, Nelson B C. Exploring the sdss data set with linked scatter plots. i. emp, cemp, and cv stars [J]. The Astrophysical Journal Supplement Series, 2017, 228(2): 19.

Caruccio L, Cirillo S, Polese G, et al. Claude 2.0 large language model: tackling a real-world classification problem with a new iterative prompt engineering approach [J]. Intelligent Systems with Applications, 2024: 200336.

Chang X, Zheng Q. Knowledge element extraction for knowledge-based learning resources organization [C]//Advances in Web Based Learning–ICWL 2007: 6th International Conference Edinburgh, UK, August 15-17, 2007 Revised Papers 6. Springer, 2008: 102-113.

Chinchor N, Marsh E. Muc-7 information extraction task definition [C]//Proceeding of the seventh message understanding conference (MUC-7), Appendices. 1998: 359-367.

Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models [J]. arXiv preprint arXiv:2210.11416, 2022.

Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models [J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.

Cortes C, Vapnik V. Support-vector networks [J]. Machine learning, 1995, 20: 273-297.

Cruces M, Spitler L, Scholz P, et al. Repeating behaviour of frb 121102: Periodicity, waiting times, and energy distribution [J]. Monthly Notices of the Royal Astronomical Society, 2021, 500(1): 448-463.

Cui C, Tao Y, Li C, et al. Towards an astronomical science platform: Experiences and lessons learned from chinese virtual observatory [J]. Astronomy and Computing, 2020, 32: 100392.

Dan Y, Lei Z, Gu Y, et al. Educhat: A large-scale language model-based chatbot system for intelligent education [J]. arXiv preprint arXiv:2308.02773, 2023.

DeJong G. An overview of the frump system [J]. Strategies for natural language processing, 1982: 149-176.

Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

Dhuliawala S, Komeili M, Xu J, et al. Chain-of-verification reduces hallucination in large language models [J]. arXiv preprint arXiv:2309.11495, 2023.

Ding Y, Song M, Han J, et al. Entitymetrics: Measuring the impact of entities [J]. PloS one, 2013, 8(8): e71416.

Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program-tasks, data, and evaluation. [C]//Lrec: volume 2. Lisbon, 2004: 837-840.

Dong Q, Li L, Dai D, et al. A survey on in-context learning [J]. arXiv preprint arXiv:2301.00234, 2022.

D' Souza J, Auer S. Computer science named entity recognition in the open research knowledge graph [C]//International Conference on Asian Digital Libraries. Springer, 2022: 35-45.

Ehrmann M, Hamdi A, Pontes E L, et al. Named entity recognition and classification in historical documents: A survey [J]. ACM Computing Surveys, 2023, 56(2): 1-47.

Eichhorn G, Kurtz M J, Accomazzi A, et al. The nasa astrophysics data system: The search engine and its user interface [J]. Astronomy and Astrophysics Supplement Series, 2000, 143(1): 61-83.

Eltyeb S, Salim N. Chemical named entities recognition: a review on approaches and applications [J]. Journal of cheminformatics, 2014, 6: 1-12.

Friedman C, Alderson P O, Austin J H, et al. A general natural-language text processor for clinical radiology [J]. Journal of the American Medical Informatics Association, 1994, 1(2): 161-174.

Fu Y, Peng H, Sabharwal A, et al. Complexity-based prompting for multi-step reasoning [C]//The Eleventh International Conference on Learning Representations. 2022.

Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey [J]. arXiv preprint arXiv:2312.10997, 2023.

Gero Z, Singh C, Cheng H, et al. Self-verification improves few-shot clinical information extraction [J]. arXiv preprint arXiv:2306.00024, 2023.

Ghosh M, Santra P, Iqbal S A, et al. Astro-mt5: Entity extraction from astrophysics literature using mt5 language model [C]//Proceedings of the first Workshop on Information Extraction from Scientific Publications. 2022: 100-104.

Giordano F, Racero E, Norman H, et al. Esasky: A science-driven discovery portal for space-based astronomy missions [J]. Astronomy and Computing, 2018, 24: 97-103.

González-Gallardo C E, Boros E, Girdhar N, et al. Yes but.. can chatgpt identify entities in historical documents? [C]//2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2023: 184-189.

Grezes F, Blanco-Cuaresma S, Accomazzi A, et al. Building astrobert, a language model for astronomy & astrophysics [J]. arXiv preprint arXiv:2112.00590, 2021.

Grezes F, Blanco-Cuaresma S, Allen T, et al. Overview of the first shared task on detecting entities in the astrophysics literature (deal) [C]//Proceedings of the first Workshop on Information Extraction from Scientific Publications. 2022: 1-7.

Grishman R, Sundheim B M. Message understanding conference-6: A brief history [C]//COLING 1996 volume 1: The 16th international conference on computational linguistics. 1996.

Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing [J]. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3 (1): 1-23.

Han X L, Zhang L Y, Shi J R, et al. Cataclysmic variables based on the stellar spectral survey lamost dr3 [J]. Research in Astronomy and Astrophysics, 2018, 18(6): 068.

Han X, Zhang Z, Ding N, et al. Pre-trained models: Past, present and future [J]. AI Open, 2021, 2: 225-250.

Helou G, Madore B, Schmitz M, et al. The nasa/ipac extragalactic database [J]. Databases & On-Line Data in Astronomy, 1991: 89-106.

Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18(7): 1527-1554.

Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.

Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models [J]. arXiv preprint arXiv:2203.15556, 2022.

Homayouni Y, Kriss G A, De Rosa G, et al. Agn storm 2. v. anomalous behavior of the c iv light curve of mrk 817 [J]. The Astrophysical Journal, 2024, 963(2): 123.

Howison J, Bullard J. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature [J]. Journal of the Association for Information Science and Technology, 2016, 67(9): 2137-2155.

Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission [J]. arXiv preprint arXiv:1904.05342, 2019.

Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions [J]. arXiv preprint arXiv:2311.05232, 2023.

Huang Y, Zhang H W, Wang C, et al. A new luminous blue variable in the outskirts of the andromeda galaxy [J]. The Astrophysical Journal Letters, 2019, 884(1): L7.

Huang Z, Xu W, Yu K. Bidirectional lstm-crf models for sequence tagging [J]. arXiv preprint arXiv:1508.01991, 2015.

Jaynes E T. On the rationale of maximum-entropy methods [J]. Proceedings of the IEEE, 1982, 70 (9): 939-952.

Ji B. Vicunaner: Zero/few-shot named entity recognition using vicuna [J]. arXiv preprint arXiv:2305.03253, 2023.

Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation [J]. ACM Computing Surveys, 2023, 55(12): 1-38.

Jiang J. Information extraction from text [J]. Mining text data, 2012: 11-41.

Jiménez Martínez I. Constraining vhe and optical emission from fast radio bursts with the magic telescopes [J]. Highlights of Spanish Astrophysics XI, 2023: 141.

Ju Z, Wang J, Zhu F. Named entity recognition from biomedical text using svm [C]//2011 5th international conference on bioinformatics and biomedical engineering. IEEE, 2011: 1-4.

Katz D M, Bommarito M J, Gao S, et al. Gpt-4 passes the bar exam [J]. Philosophical Transactions of the Royal Society A, 2024, 382(2270): 20230254.

Kong A, Zhao S, Chen H, et al. Better zero-shot reasoning with role-play prompting [J]. arXiv preprint arXiv:2308.07702, 2023.

Kuchar T, Sloan G, Mizuno D, et al. Smc-last extracted photometry [J]. The Astronomical Journal, 2024, 167(4): 149.

Kun E, Jaroschewski I, Ghorbanietemad A, et al. Multimessenger picture of j1048+ 7143 [J]. The Astrophysical Journal, 2022, 940(2): 163.

Kurtz M J, Eichhorn G, Accomazzi A, et al. The nasa astrophysics data system: Overview [J]. Astronomy and astrophysics supplement series, 2000, 143(1): 41-59.

Lafferty J, McCallum A, Pereira F, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Icml: volume 1. Williamstown, MA, 2001: 3.

Leaman R, Wei C H, Lu Z. tmchem: a high performance approach for chemical named entity recognition and normalization [J]. Journal of cheminformatics, 2015, 7: 1-10.

LeCun Y, Bengio Y, et al. Convolutional networks for images, speech, and time series [J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.

LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521(7553): 436-444.

Lee J, Yoon W, Kim S, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234-1240.

Lesteven S, Bonnin C, Derriere S, et al. Djin: Detection in journals of identifiers and names [C]// Library and Information Services in Astronomy VI: 21st Century Astronomy Librarianship, From New Ideas to Action: volume 433. 2010: 317.

Levy I, Bogin B, Berant J. Diverse demonstrations improve in-context compositional generalization [J]. arXiv preprint arXiv:2212.06800, 2022.

Li J, Sun A, Han J, et al. A survey on deep learning for named entity recognition [J]. IEEE transactions on knowledge and data engineering, 2020, 34(1): 50-70.

Li J, Chen J, Ren R, et al. The dawn after the dark: An empirical study on factuality hallucination in large language models [J]. arXiv preprint arXiv:2401.03205, 2024.

Li X, Wang X, Liu J, et al. Lamost j2043+ 3413–a fast disk precession sw sextans candidate in period gap [J]. arXiv preprint arXiv:2306.07529, 2023.

Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning [J]. Advances in Neural Information Processing Systems, 2022, 35: 1950-1965.

Liu Z, Yang M, Wang X, et al. Entity recognition from clinical texts via recurrent neural network [J]. BMC medical informatics and decision making, 2017, 17: 53-61.

Liu Z, Huang D, Huang K, et al. Finbert: A pre-trained financial language representation model for financial text mining [C]//Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence. 2021: 4513-4519.

Lortet M C, Borde S, Ochsenbein F. Second reference dictionary of the nomenclature of celestial objects. [J]. Astronomy and Astrophysics Suppl., Vol. 107, p. 193-218 (1994), 1994, 107: 193-218.

Lu H p, Michel R, Zhang L y, et al. Magnetic activity and period variation studies of the short-period eclipsing binaries. iii. v1175 her, nsvs 2669503, and 1swasp j133417. 80+ 394314.4 [J]. The Astronomical Journal, 2018, 156(3): 88.

Lu H p, Zhang L y, Michel R, et al. Magnetic activity and period variation studies of the four w uma-type eclipsing binaries: Uv lyn, v781 tau, nsvs 4484038, and 2mass j15471055+ 5302107 [J]. The Astrophysical Journal, 2020, 901(2): 169.

Luo L, Yang Z, Yang P, et al. An attention-based bilstm-crf approach to document-level chemical named entity recognition [J]. Bioinformatics, 2018, 34(8): 1381-1388.

Lyu Q, Havaldar S, Stein A, et al. Faithful chain-of-thought reasoning [J]. arXiv preprint arXiv:2301.13379, 2023.

Marculewicz M, Nikolajuk M, Różańska A. Deep absorption in sdss j110511. 15+ 530806.5 [J]. Astronomy & Astrophysics, 2022, 668: A128.

Mazzarella J M, Team N, et al. Ned for a new era [C]//Astronomical Data Analysis Software and Systems XVI: volume 376. 2007: 153.

McKenna N, Li T, Cheng L, et al. Sources of hallucination by large language models on inference tasks [J]. arXiv preprint arXiv:2305.14552, 2023.

Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in neural information processing systems, 2013, 26.

Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention [J]. Advances in neural information processing systems, 2014, 27.

Munnangi M, Feldman S, Wallace B C, et al. On-the-fly definition augmentation of llms for biomedical ner [J]. arXiv preprint arXiv:2404.00152, 2024.

Murphy T, McIntosh T, Curran J R. Named entity recognition for astronomy literature [C]// Proceedings of the Australasian Language Technology Workshop 2006. 2006: 59-66.

Nguyen T D, Ting Y S, Ciucă I, et al. Astrollama: Towards specialized foundation models in astronomy [J]. arXiv preprint arXiv:2309.06126, 2023.

Niu J R, Zhu W W, Zhang B, et al. Fast observations of an extremely active episode of frb 20201124a. iv. spin period search [J]. Research in Astronomy and Astrophysics, 2022, 22(12): 124004.

Office U S D A R P A I T. Sixth message understanding conference,(muc-6): Proceedings of a conference held in columbia, maryland, november 6-8, 1995: volume 353 [M]. Morgan Kaufmann Publishers, 1995.

Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

Perera N, Dehmer M, Emmert-Streib F. Named entity recognition and relation detection for biomedical information extraction [J]. Frontiers in cell and developmental biology, 2020, 8: 673.

Perkowski E, Pan R, Nguyen T D, et al. Astrollama-chat: Scaling astrollama with conversational and diverse datasets [J]. Research Notes of the AAS, 2024, 8(1): 7.

Piskorski J, Yangarber R. Information extraction: Past, present and future [J]. Multi-source, multilingual information extraction and summarization, 2013: 23-49.

Puccetti G, Giordano V, Spada I, et al. Technology identification from patent texts: A novel named entity recognition method [J]. Technological Forecasting and Social Change, 2023, 186: 122160.

Purandardas M, Goswami A, Shejeelammal J, et al. Lamost j045019. 27+ 394758.7, with peculiar abundances of n, na, v, zn, is possibly a sculptor dwarf galaxy escapee [J]. Monthly Notices of the Royal Astronomical Society, 2022, 513(4): 4696-4710.

QasemiZadeh B, Schumann A K. The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods [C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016: 1862-1868.

Qin Y, Zeng Y. Research of clinical named entity recognition based on bi-lstm-crf [J]. Journal of Shanghai Jiaotong University (Science), 2018, 23: 392-397.

Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey [J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.

Qu C X, Luo A L, Wang R, et al. Stellar atmospheric parameters for cool dwarfs in gaia data release 3 [J]. The Astrophysical Journal Supplement Series, 2024, 270(2): 32.

Sager N. Natural language information processing: a computer grammmar of english and its applications [M]. Addison-Wesley Longman Publishing Co., Inc., 1981.

Sager N. Computer analysis of sublanguage information structures [J]. Annals of the New York Academy of Sciences, 1990, 583(1): 161-179.

Sanderson H, Bonsor A, Mustill A. Can gaia find planets around white dwarfs? [J]. Monthly Notices of the Royal Astronomical Society, 2022, 517(4): 5835-5852.

Shang L H, Bai J T, Dang S J, et al. The "bi-drifting" subpulses of psr j0815+ 0939 observed with the five-hundred-meter aperture spherical radio telescope [J]. Research in Astronomy and Astrophysics, 2022, 22(2): 025018.

Shannon C E. A mathematical theory of communication [J]. The Bell system technical journal, 1948, 27(3): 379-423.

Shao W, Hu Y, Ji P, et al. Prompt-ner: Zero-shot named entity recognition in astronomy literature via large language models [J]. arXiv preprint arXiv:2310.17892, 2023.

Shi F, Chen X, Misra K, et al. Large language models can be easily distracted by irrelevant context [C]//International Conference on Machine Learning. PMLR, 2023: 31210-31227.

Shi Z, Comte G, Luo A L, et al. Emission lines properties of the radio-loud quasar lamost j1131+ 3114 [J]. Astronomy & Astrophysics, 2014, 564: A89.

Sobhana N, Mitra P, Ghosh S. Conditional random field based named entity recognition in geological text [J]. International Journal of Computer Applications, 2010, 1(3): 143-147.

Sotnikov V, Chaikova A. Language models for multimessenger astronomy [J]. Galaxies, 2023, 11 (3): 63.

Su H, Kasai J, Wu C H, et al. Selective annotation makes language models better few-shot learners [J]. arXiv preprint arXiv:2209.01975, 2022.

Sundheim B. Appendix c: Named entity task definition (v2. 1) [C]//Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8. 1995.

Sundheim B M, Chinchor N. Survey of the message understanding conferences [C]//Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993. 1993.

Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine [J]. Nature medicine, 2023, 29(8): 1930-1940.

Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models [J]. arXiv preprint arXiv:2307.09288, 2023.

Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

Wang S, Sun X, Li X, et al. Gpt-ner: Named entity recognition via large language models [J]. arXiv preprint arXiv:2304.10428, 2023.

Wang Y, Zhang C. Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing [J]. Journal of informetrics, 2020, 14(4): 101091.

Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

Wenger M, Ochsenbein F, Egret D, et al. The simbad astronomical database-the cds reference database for astronomical objects [J]. Astronomy and Astrophysics Supplement Series, 2000, 143(1): 9-22.

Weston L, Tshitoyan V, Dagdelen J, et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature [J]. Journal of chemical information and modeling, 2019, 59(9): 3692-3702.

Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models [J]. arXiv preprint arXiv:2401.11817, 2024.

Ye X, Iyer S, Celikyilmaz A, et al. Complementary explanations for effective in-context learning [J]. arXiv preprint arXiv:2211.13892, 2022.

Zhang B, Qian S B, Wang J J, et al. 1swasp j034439. 97+ 030425.5: a short-period eclipsing binary system with a close-in stellar companion [J]. Research in Astronomy and Astrophysics, 2020, 20 (4): 047.

Zhang M, Chen B Q, Huo Z Y, et al. A catalogue of hα emission-line point sources in the vicinity fields of m 31 and m 33 from the lamost survey [J]. Research in Astronomy and Astrophysics, 2020, 20(6): 097.

Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts [J]. Journal of biomedical informatics, 2013, 46(6): 1088-1098.

Zhang Y, Feng S, Tan C. Active example selection for in-context learning [J]. arXiv preprint arXiv:2211.04486, 2022.

Zhang Z, Zhang A, Li M, et al. Automatic chain of thought prompting in large language models [J]. arXiv preprint arXiv:2210.03493, 2022.

Zhao Q, Xu M, Gupta K, et al. The first to know: How token distributions reveal hidden knowledge in large vision-language models? [J]. arXiv preprint arXiv:2403.09037, 2024.

Zhao W X, Zhou K, Li J, et al. A survey of large language models [J]. arXiv preprint arXiv:2303.18223, 2023.

Zhao Z, Wallace E, Feng S, et al. Calibrate before use: Improving few-shot performance of language models [C]//International Conference on Machine Learning. PMLR, 2021: 12697-12706.

Zong C, Xia R, Zhang J. Information extraction [M]//Text Data Mining. Springer, 2021: 227-283.

致 谢

至此，求学路近二十载。还能依稀记得三年前拿到硕士录取通知书时对硕士研究生生涯的满怀期待，而转眼间，我又即将要与它挥手告别。在本硕士学位论文即将完成之际，我满怀感激之情，向以往关心、支持和帮助过我的人致以诚挚的谢意。

首先，我特别要感谢我的导师樊东卫副研究员。樊老师不仅以其深厚的专业知识和严谨的学术态度为我指引方向，而且在研究方法和论文撰写上给予了我宝贵的指导和建议。在遇到难题时，樊老师总是耐心地与我讨论，提供解决方案，帮助我克服难关。此外，樊老师对于科研的热情和对细节的把握也深深影响了我，让我在学术探索的道路上更加专注。在此，我诚挚地感谢樊老师在学术和生活上对我的支持与鼓励，以及给予的无私帮助。

其次，我还要特别感谢我们团组的首席崔辰州研究员。崔老师作为团组的领头人，不仅为我们提供了一个纯粹和充满机遇的学术环境，而且以其丰富的经验和远见卓识，为我们的研究工作指明了方向。崔老师的领导力和对科研的热情激励着我们每一位团队成员。此外，崔老师对于团队合作精神的强调，也为我树立了学术和人生的标杆。在此，我衷心地感谢崔老师在硕士期间给予的悉心指导，以及无私的帮助和支持。

感谢国家天文台天文信息技术团组的许允飞老师、米琳莹老师、陶一寒老师、韩军老师、李珊珊老师、何勃亮老师、李长华老师、杨丝丝老师、杨涵溪老师和王有芬老师。感谢你们的悉心教导和帮助，我将铭记在心。

感谢之江实验室大数据智能研究中心的严笑然老师、张睿老师、姬朋立老师、胡耀华同学、田时齐同学，你们的悉心指导和热情交流为我的研究之路增添了宝贵的知识和温馨的回忆。

感谢同门师兄弟姐妹们，包括陈朗、马鹏辉、汤超、吴莹、杨嘉宁、左肖雄、朱珈莹、张震、张琦乾。在攻读学位的这几年里，你们是我的学习伙伴和生活中的朋友。你们的智慧、耐心和幽默为我的学术探索带来了启发和动力。特别是在我遇到困难和挫折时，是你们给予了我鼓励和帮助，让我能够不断前进。这份深厚的友谊，我将永远珍惜。

感谢国台教育处的梁艳春老师、马怀宇老师和李响老师。在学习和研究的道路上，你们不仅是管理者，更是引导者和支持者。

感谢一路走来对我有深刻影响的其他恩师们，他们有蔡正权老师、张向云老师、廖胜权老师、许少双老师和宫大卫老师。你们的教导和帮助我时刻铭记在心。

感谢我硕士期间的两位舍友吕鑫和付秋阳，你们不仅给予我一个温馨而充满科研氛围的宿舍环境，更是我学习和生活中的支持者和倾听者。特别是在我撰

写论文的过程中，你们给予了我莫大的鼓励和帮助，同时也无私地倾注着你们的心血和分享着你们的经验。你们的支持和陪伴让我感到无比幸运和温暖，真诚地感谢你们。

感谢本科期间与我同住东十 514 宿舍的程超、丁一维、方志旺、兰亚鹏和王海浪五位舍友，我永远怀念我们一起度过的快乐且美好的时光。

感谢一路走来其他重要的同学、朋友和领导，他们有陈曦、韦必浩、张涛、程雨、张超、黄威、刘淑飞、杜朋亮、于燚、徐田华、柴晶晶、宫雪、刘越、桑荣庆、王艺萌。

特别要感谢我的家人，父亲、母亲和姐姐无私的照顾与支持让我得以完成硕士研究生学业，你们的辛苦与遗憾我将以十足的努力来回报！

最后我要特别感谢我们的党和国家。和平与繁荣发展为我们提供了宝贵的学习机会，我将继续充实自己的知识和技能，以期将来能为党和国家的发展和建设贡献自己的力量。

再次致以诚挚的谢意！

2024 年 6 月

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历:

2016 年 09 月——2020 年 07 月，在安徽科技学院信息与网络工程学院获得学士学位。

2021 年 09 月——2024 年 06 月，在中国科学院国家天文台攻读硕士学位。

已发表（或正式接受）的学术论文:

(1) **Shao Wujun**, Zhang Rui, Ji Pengli, Fan Dongwei, Hu Yaohua, Yan Xiaoran, Cui Chenzhou, Tao Yihan, Mi Linying, Chen Lang. Astronomical Knowledge Entity Extraction in Astrophysics Journal Articles via Large Language Models. Research in Astronomy and Astrophysics, 2024. （已正式接受）

申请或已获得的专利:

(1) 发明专利：**邵务俊**，樊东卫，崔辰州；LAMOST 文献天体标识信息提取方法、装置、设备及介质；CN202311341443.2（已进入实审）

(2) 发明专利：**邵务俊**，樊东卫，崔辰州；天文星表离群数据的挖掘方法、装置、设备及介质；CN202311507431.2（已进入实审）

(3) 发明专利：张睿，胡耀华，姬朋立，严笑然，**邵务俊**；基于论文 PDF 的天文多模态知识图谱构建方法和系统；CN202311333807.2（已授权）

参加的研究项目及获奖情况:

(1) 参与国家自然科学基金项目《LAMOST 数据与异构科研数据融合方法研究》

(2) 参与国家重点研发计划青年科学家项目《多模态天文科学数据知识关联推荐系统》

(3) 获评 2023 学年度中国科学院大学“优秀学生干部”称号