中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：22721482

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	面向科学文献的命名实体识别研究与应用

作　　者：王慧
学科专业：计算机应用技术
导　　师：韩越兴
完成日期：2025 年 5 月

姓　　名：王慧　　学号：22721482

论文题目：面向科学文献的命名实体识别研究与应用

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕/博士学位论文质量要求。

答辩委员会签名：

主　　席：

委　　员：

导　　师：

答辩日期：2025年6月10日

姓　　名：王慧　　学号：22721482

论文题目：面向科学文献的命名实体识别研究与应用

上海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指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：王慧

日 期：2025 年 6 月 10 日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密论文在解密后应遵守此规定）

学位论文作者签名：王慧　　导师签名：韩赅芳

日 期：2025年 6月 10日　　日 期：2025年 6月 10日

上海大学工学硕士学位论文

面向科学文献的命名实体识别研究与应用

作　　者： 王慧

学科专业： 计算机应用技术

导　　师： 韩越兴

上海大学计算机工程与科学学院

2024 年 5 月

A Dissertation Submitted to Shanghai University for the

Degree of Master in Engineering

Research and Applications of Named Entity Recognition for Scientific Literature

Candidate: 王慧

Major: Computer Application Technology

Supervisor: 韩越兴

School of Computer Engineering and Science,

Shanghai University

May, 2025

摘　要

从海量且非结构化的科学文献中精准高效地进行命名实体识别，是实现信息提取、支持知识发现和加速科研决策的关键环节。然而，科研文献数量的快速增长和学科交叉所带来的术语密集与知识体系复杂化问题，在材料科学、生物医学等高度专业化的领域尤为突出，这使得传统人工检索难以保证效率与准确性。本文针对上述问题提出了基于大语言模型的上下文一致的实体显式标注方法与结合直接偏好优化的双阶段训练方法，然后进一步针对高度专业化的材料科学和生物医学领域，提出了基于领域语言模型的语义融合方法。本文的主要研究内容如下：

（1）为了发挥大语言模型在科学文献场景下进行实体提取的潜力，本文提出了一种上下文一致的实体显式标注方法与双阶段训练方法，来解决大语言模型在生成式输出与命名实体识别任务序列标注特性之间的差异。然后，在训练阶段分为监督微调和直接偏好优化两个阶段，监督微调阶段先在已有标注数据上学习基本的实体识别能力；直接偏好优化阶段为了更有效地引导模型纠正错误，在负样本构造时尝试扩张、收缩实体边界，并对监督微调后的推理结果筛选制造类别混淆样本，通过利用正负样本对的偏好差进行约束，增强模型对错误判定的修正能力。

（2）为解决通用模型在处理材料科学、生物医学等高度专业化领域时，因大量低频专业术语导致命名实体识别精度不足的问题，本文提出了一种基于领域语言模型的语义融合方法，通过将不同领域语言模型和领域词级向量进行语义融合，来增强对科学文献的深层语义理解，通过实验验证了方法对材料科学和生物医学领域中复杂专业文本的有效性。最后将方法应用在具体领域，展现其在科研文本挖掘及辅助研发决策方面的实用价值。

关键词：大语言模型；领域语言模型；命名实体识别

ABSTRACT

Accurate and efficient named-entity recognition (NER) from the vast body of unstructured scientific literature is a key step toward information extraction, knowledge discovery, and accelerated research decision-making. However, the rapid growth of publications and the proliferation of interdisciplinary terminology have intensified the complexity of domain knowledge, particularly in highly specialized fields such as materials science and biomedicine, making traditional manual curation unable to guarantee efficiency or accuracy. In response, this paper proposes a context-consistent explicit tagging method for entities based on large language models, combined with a two-stage training strategy that incorporates direct preference optimization. Furthermore, for highly specialized domains such as materials science and biomedicine, a domain-aware semantic-fusion method is introduced. The main contributions are as follows:

(1) To leverage the potential of large language models for entity extraction in scientific literature, this paper designs a context-consistent explicit tagging scheme with a two-stage training framework that bridges the gap between generative outputs and the sequence-labeling nature of NER. The training process is divided into supervised fine-tuning and direct preference optimization (DPO). In the supervised phase, the model acquires basic entity-recognition ability from annotated data. In the DPO phase, the model is guided to correct errors more effectively through the construction of hard negatives (by expanding or shrinking entity boundaries) and the generation of label-confusion samples from the inference results post fine-tuning. The preference differences between positive and negative pairs guide the model, strengthening its capability to rectify boundary and type errors.

(2) This paper proposes a domain-aware semantic-fusion method to address the accuracy degradation caused by numerous low-frequency technical terms when generic models are applied to highly specialized domains such as materials science

and biomedicine. The method enriches deep semantic understanding of scientific texts by fusing representations from multiple domain-specific language models with domain-specific word-level embeddings. Experiments have verified its effectiveness on complex specialist corpora in materials science and biomedicine. Finally, the method is applied in practical scenarios, demonstrating its utility in scientific text mining and supporting research and development decision-making.

Keywords: Large language model; Domain language model; Named entity recognition

目 录

第一章　绪论

1.1 课题来源

本课题得到国家自然科学基金（项目编号：52273228、52373227）、国家自然科学基金青年科学基金项目（项目编号：62401352）、云南省科技计划重点项目（项目编号：202302AB080022）、上海市科技青年人才扬帆计划（项目编号：23YF1412900）、教育部上海大学硅酸盐文物保护重点实验室项目（项目编号：SCRC2023ZZ07TS、SCRC2024ZZ07TS）以及国家重点研发计划（项目编号：2017YFB0701502、2017YFB0702901）的资助。

1.2 课题研究的目的和意义

近年来，科研文献数量在各个领域呈指数级增长，学科交叉的趋势也日益显著，多领域术语和繁复的实验描述使学术信息的规模与复杂度持续攀升。对研究者而言，文献虽是获取关键信息的主要途径，但在庞大且不断更新的文献资源面前，依靠人工检索已难以兼顾效率与准确性。信息抽取作为一种基础的自然语言理解任务，能够将非结构化文本转化为结构化并为文献的自动化处理与大规模文本挖掘提供有力的技术支撑[1]。命名实体识别任务作为信息抽取的前置核心任务，能够准确识别和提取文献中的关键信息，为后续的关系抽取、信息检索及知识图谱构建提供坚实的数据支持；实时而精准的实体识别可显著提升科研决策的准确性与效率，帮助研究者更好地把握前沿动态、验证科学假设，并加速研发与设计进程。

随着深度学习的快速发展，不同学科开始探索如何利用语言模型进行信息提取来处理海量且复杂的数据。通过在海量文本数据上进行预训练，语言模型能够掌握语言的深层语义结构与模式，从而有效支撑各类自然语言处理任务。尤其是近年来，以庞大的参数规模和前所未有的训练数据量为特征的大语言模型经历了蓬勃发展，显著提升了语言理解与生成的性能基准，在文本生成、摘

要、翻译、问答乃至一定程度的推理等多样化任务中表现出色。这些模型通常拥有超过千亿级的参数，自然语言处理普遍将这些大规模的预训练语言模型称为大语言模型。与传统语言模型不同，大语言模型通常展现出令人惊叹的涌现能力[2][3]，能够同时适应大量自然语言处理任务并取得显著的表现。尽管如此，大语言模型在命名实体识别任务上的整体表现仍明显低于监督式基线模型[4]。主要原因在于二者输出形式的差异，命名实体识别任务本质上是序列标注任务，更强调精确标注，而大语言模型更偏向文本生成；语义标注任务与文本生成模型之间的差距，再加上幻觉等[5]问题，使得这一挑战更为突出。

除此之外，在训练通用语言模型时，通常会使用来自网络文本、书籍以及新闻报道等大规模通用语料，用以学习语言的一般特征与结构。然而，与通用领域相比，科学文献的信息抽取面临更加严峻的挑战。首先，科学文献常包含大量专业术语，对文本进行准确的标注需要丰富的领域知识，使得人工标注成本高昂，且高密度的专业领域化信息也给命名实体识别增加了难度。以材料科学为例，文献的快速增长使得依赖人工对大规模文献进行分析来总结规律与趋势的难度陡增，同时在海量文献中搜索具备特定理想属性的材料体系也变得更加复杂，这在很大程度上阻碍了后续基于数据驱动的研究与应用[6]。在生物医学领域，类似的情况同样存在。生物医学文献数量持续快速增长，大量文献中蕴含着与新发现相关的关键信息。然而，由于高质量标注所需的专业背景要求和人力成本，生物医学文本挖掘也极其依赖自动化技术来从海量文献中提取实体、关系及相关属性。加之生物医学语料在词分布、术语使用等方面与通用领域文本存在显著差异，若仅依赖通用自然语言处理方法或数据，往往难以取得理想效果[7]。

综上所述，在信息抽取领域，如何充分发挥大语言模型的上下文理解与生成优势，仍是需要深入探讨的研究方向。与此同时，面对诸多专业领域中不断涌现的大规模文献，尤其是对于材料科学、生物医学等拥有大量专业术语与特定知识体系的领域，如何高效使用兼具领域适配性与性能优势的领域语言模型也尤为关键。利用语言模型强大的语义表示和理解能力，通过自动化识别文献中的关键概念与术语信息，可以显著降低科研人员在数据收集与筛选过程中的

重复性劳动，将海量科研信息转化为可分析结构数据的起点，为后续的关系抽取、事件检测以及知识图谱构建等深层次研究奠定坚实基础[8]。

1.3 国内外研究概况

近年来，自然语言处理领域的研究取得了显著进展，预训练语言模型的发展尤其是现在的大语言模型，大幅提升了各种自然语言处理任务的表现。然而，科学文献作为一种特殊的文本类型，其领域特性和专业术语的丰富性依然对现有方法构成挑战。为此，本节首先回顾语言模型整体研究的进展，包括大语言模型的发展状况，然后聚焦于科学文献中命名实体识别的研究现状。

1.3.1 语言模型研究现状

自统计语言模型转向深度学习以来，语言模型在语义表示与复杂推理等方面取得了显著进展。早期研究主要依赖对词语共现概率的统计建模，并在此基础上不断改进平滑与回退算法，以缓解数据稀疏与参数规模之间的平衡难题。深度学习的兴起促使循环神经网络被广泛应用于语言模型领域，循环神经网络能够在序列方向上积累和更新状态，从而在一定程度上捕捉句子内部的上下文关联[9]。长短期记忆网络（LSTM）又通过引入输入门与遗忘门等门控机制缓解了梯度消失和梯度爆炸问题，使得对较长序列的建模成为可能[10]。这类模型在文本生成与序列标注等任务上表现出显著优势，但在大规模并行化及更高层次的语义推理上仍面临一定局限。

自注意力机制与 Transformer 框架的出现则进一步推动了语言模型向更深层次的语义理解迈进[11]。预训练语言模型逐渐成为主流范式。该方法通常先在海量文本上通过自监督任务，学习通用语义与句法表征，然后再通过微调来适配具体任务，从而在文本分类、机器翻译、自然语言推理等多个领域获得显著性能提升[12]。典型代表包括采用双向编码器 BERT[13]，通过自回归方式进行文本生成的 GPT 系列[14][15]以及编码-解码架构下统一多任务格式的 T5[16]。其中，BERT 首次引入掩码语言建模与下一句预测方法来获取深层双向语境信息，

GPT 系列则借助从左到右的生成式预训练，在文本生成、续写等应用中显示出强大灵活性。T5 将绝大多数自然语言处理任务都转化为“文本到文本”形式，为下游多任务训练和迁移学习提供了新的思路。随着研究者持续增大模型参数规模与训练数据规模[17][18]，预训练模型开始呈现出跨任务的“能力涌现”现象，为从预训练语言模型向超大规模的语言模型，即大语言模型的跃迁提供了重要契机[19]。

在预训练方法的基础上，通过进一步放大参数规模与训练数据，研究者逐渐发现模型的理解与生成性能会在一定阈值处出现质变，形成了大语言模型的研究热点。GPT-3[20]以千亿级参数规模并采用自回归 Transformer 架构在大规模无监督文本上完成预训练，在零样本或少样本学习场景下展现了远超普通模型的语言能力[21]，仅依赖提示[22]就能完成问答[23]、翻译[24]、摘要[25]等多项任务[26]。PaLM[27]是将参数规模扩展至 5000 亿以上，通过 Pathways 技术实现对大规模参数的有效管理和并行训练，在多项下游基准上进一步刷新记录。GPT-4[28]则从多模态输入、链式推理与安全性等角度进行了新的探索，为学术和工业应用中更复杂的人机交互提供了更高层次的可用性。与此同时，InstructGPT[23]和 ChatGPT[29]通过结合人工反馈（RLHF）或监督微调，使模型在对话系统与交互应用中对人类需求展现出更强的匹配度[29]。

1.3.2 面向科学文献的命名实体识别研究现状

在科学文献处理领域，命名实体识别被公认为信息抽取与语义理解的重要环节，不仅在文本信息提取、知识图谱构建等任务中占据关键地位[30]，也为多学科交叉研究提供了坚实的技术基础[31]。

早期的命名实体识别主要依赖于人工编写的规则和词典进行匹配，如典型的基于正则表达式或模式匹配的系统。这些方法对于专有名词相对固定、上下文场景比较有限时，效果可令人满意。然而它们也存在明显的缺点，规则的设计和维护成本高且难以迁移，当领域或语言环境发生变化，系统需要大规模的手动调整；因此这类方法对于极具多样性或开放域的文本表现较差。随着自然语言处理研究的不断推进，统计机器学习方法开始应用于命名实体识别，其中

条件随机场[32]等序列标注模型成为了这一时期的标志性工作。通过为文本构建特征模板并利用序列依赖关系，条件随机场在一定程度上实现了对字符、词及上下文信息的联合建模[33]。与规则系统相比，此类方法在实际应用中更具灵活性和可移植性，因为只需通过调整特征模板或训练数据就能适应不同领域。然而，这些模型对特征工程的依赖依然较高，往往需要结合人工知识和特征选择来提高效果。对数据规模和质量也有所要求，一旦缺乏足够的带标注数据或有效的特征模板，模型性能会受到明显制约[34]。

近年来，随着深度学习技术的成熟与大规模预训练模型的出现，命名实体识别进入了以神经网络为主要驱动力的新阶段。早期的神经网络方法多基于循环神经网络及其变体，如 LSTM 或双向 LSTM，通过结合字符级和词级嵌入，显著提升了对上下文依赖信息的刻画能力。随后，卷积神经网络[35]在提取局部特征方面的优势也被利用，用以辅助命名实体识别中的字符层表示。在这类模型中，特征工程被自动化，模型可以从数据中直接学习到对识别有用的模式。这带来了模型效果上的大幅提升，但与此同时，训练这些神经网络往往需要更大规模、更高质量的标注数据[36]。在数据不足的场景中，如果不进行适当的预训练或数据增广，模型易出现过拟合或者泛化能力不佳的情况，因此依然难以应用于科学文献领域。

随着大规模预训练语言模型的引入，尤其是 BERT 及其后续一系列变体进一步提升了命名实体识别的效果。由于预训练模型在海量无标注数据上学习了丰富的语言表示，这些模型在少量标注数据场景下也能具有较好的表现[38]。然而，与新闻、社会媒体等通用文本相比，科学文献通常具有更强的专业性和领域特征[39]，不同的文献类型在实体定义和关注重点上可能大相径庭[40]。例如，医学期刊可能关注疾病名称、药物和基因[41]，而面向材料科学的论文则更倾向于识别新材料成分及其实验条件。因此，各个学科训练了了各自领域的预训练语言模型。在生物医学领域，Lee 等[7]开发 BioBERT，以通用模型基础持续预训练，在生物任务中效果显著，但受通用领域词汇表限制。Yasunaga 等[42]提出 BioLinkBERT，利用文献引用图结构跨文档联合预训练，大幅增强了多跳推理能力，但复杂图结构也增加了模型训练负担。Gu 等[43]提出 PubMedBERT，从

头领域预训练，在生物医学任务中表现较好，但预训练资源需求较高。在材料科学领域，Gupta 等[6]开发了 MatSciBERT，通过特定文献预训练有效识别材料术语，但未充分考虑跨文档关系知识。Shetty 等[44]提出 MaterialsBERT，通过大规模摘要预训练与自动化抽取挖掘材料属性关联，但对少样本任务表现有限。

尽管这些基于 BERT 架构的领域预训练模型在各自领域内取得了一定的进展，但仍存在一些普遍的不足。具体而言，这类模型通常依赖较大的标注数据量进行微调，且在少样本和跨领域任务中的表现受限。针对这些问题，大语言模型的兴起为自然语言处理的研究提供了新的思路，其通过超大规模参数和更广泛的语料训练，在语言理解和生成能力上取得了突破性的进展，能够有效缓解标注数据稀缺问题并提高少样本任务的性能，对自然语言处理及其子任务包括命名实体识别任务在内的各个方向都带来了新的研究范式。这些模型凭借超大规模参数量和在广泛语料上的训练，表现出了极强的语言理解与生成能力，尤其是上下文内学习（in-context learning）方式，仅需少量示例便能完成新任务[45][46]。

然而，相较于端到端训练或基于微调的监督学习模型，大语言模型在命名实体识别任务上依然存在明显差距。一方面，命名实体识别本质上是一个序列标注任务，需要对句子中每个词元进行精确的实体类型分类；另一方面，GPT 等生成式模型侧重于自动生成文本，生成过程中还可能伴随“幻觉”问题，导致生成与上下文不一致或虚构的内容[47][48]，在严谨性要求较高的科学文献信息抽取场景下，这种不匹配会使模型表现不及预期[4][49]。除此之外，科学文献与通用文本在语言表达上存在显著区别，首先，科学文献有大量的专业词汇和符号系统，如材料领域的组织结构名称和性能单位以及特定的缩写和符号等[8]，这些专有名词和符号需要依赖专业知识才能准确解析，在通用文本中出现频率很少。其次，科学文献信息密集，并且常常跨越化学、物理、生物等多个学科领域，不同领域间的知识背景差异使得信息提取和理解变得更加困难。因此，大语言模型在预训练过程中由于训练语料覆盖不足或专业知识缺失，往往对新兴概念和专业术语的识别不够准确。Jiang 等[23]研究了提示工程和微调在材料信息提取中的应用，提升了材料预测与设计能力，但对领域知识整合仍不充分。

为了解决以上问题，一系列工作对如何更好地利用大模型的通用语言表示能力，并结合提示工程、自动数据构造或检索等技术，提出了相应的解决方法。针对开放领域命名实体识别，UniversalNER[50]通过任务聚焦蒸馏与指令调优，将 ChatGPT 的能力压缩至轻量级模型。GoLLIE[51]将复杂的标注指南嵌入微调流程，使模型能够在零样本信息抽取任务中遵循未见过的指南。在数据增强策略方面，GDA[52]提出了一种大语言模型驱动的指导式数据增强方法，通过抽象化的上下文与句子结构生成多样化训练样本。在科学文献领域信息抽取中，PGA[53]通过同义改写和隐含标签生成两种伪样本扩充训练集。此外，Li Qi 等人[54]结合实体链接与最优自训练策略，从大语言模型获取并筛选高质量外部知识，有效过滤非实体干扰，从而增强了模型对实体特征的捕获能力。但对于科学文献领域，如何在保证模型强大上下文理解能力的同时，实现对专业性、跨领域信息和精确标注的有效捕捉，仍是当前需要解决的重要挑战。

综上所述，面向科学文献的命名实体识别技术虽已从早期的规则、统计机器学习方法，发展到基于深度学习的序列标注模型，并随着领域语言模型如 BioBERT、MatSciBERT 等及大型语言模型的出现取得了显著进展。然而，现有方法在处理高度专业化、信息密集的科学文献时仍面临诸多挑战。新兴的大语言模型虽然展现了强大的上下文理解和少样本学习潜力，但其生成特性与命名实体识别这类序列标注任务存在固有矛盾，且易产生“幻觉”，且在处理科学领域，尤其是材料科学和生物医学这类包含大量独特术语、符号体系及复杂缩写的领域时，常因缺乏深度领域知识而表现不佳。近期虽有工作尝试通过提示工程、数据增强等方式优化大语言模型在命名实体识别上的应用，但在保证信息抽取精确性、可靠性，同时有效融合领域知识方面仍存在明显不足。因此，尽管先进的语言模型显著推动了自然语言处理的整体发展，但面向科学文献的命名实体识别任务中，有效应对专业术语的复杂性与低频性、克服标注数据相对稀疏的限制，以及实现领域知识的深度融合等问题，仍然构成关键的研究挑战。

1.4 论文的主要研究内容

本文聚焦于科学文献中的命名实体识别，目标是在高效处理庞大专业文献的同时兼顾识别精度与领域适配性。首先，提出了一种上下文一致的显式标注方法，通过在文本中插入成对的特殊符号，显式突出实体边界与类别信息，从而减少词级对齐与类别混淆的难度。接着，为了进一步提升模型对实体类别和边界的判定准确度,提出了由监督微调和直接偏好优化构成的双阶段训练方法，监督微调阶段用于建立模型的基础识别能力，直接偏好优化阶段[55]则结合正负样本对引导模型纠错。为使这一过程更具针对性，本文在负样本构造中从实体类别错误与实体边界错误两个方向出发，分别提出了扩张类、收缩类以及“多标”与类别错误类负样本。并且，为了确保正负样本在上下文中仅在实体标注层面不同的情况下有明显差异，本文利用有效标签数及其比例对样本进行筛选采样，来提高负样本的代表性与对比信息。将不同种类的负样本输入到直接偏好优化阶段进行训练，在保留模型原有知识分布的同时，强化其对实际推理场景下复杂错误类型的修正能力。对于材料科学和生物医学文献这种高度专业化的领域中出现的专业术语和缩写通常在通用语料中出现频率较低，单纯依赖通用预训练语言模型容易出现误判与漏标的问题，通过对领域语言模型与词向量进行语义融合，来增强对材料科学和生物医学文献的理解。考虑到材料科学的数据集规模较小且涉及领域知识较为单一，因此构建了一个统一标注规范的材料领域专用数据集作为补充，通过借助本文提出的数据集在具体领域中进行了实际应用验证。

1.5 论文组织架构

为便于读者更好地理解本文提出的研究方法及其实验验证过程，全文结构安排如下：

第一章：阐述了研究背景与研究动因，重点分析了在材料科学、生物医学等高度专业化领域中，科研文献数据规模快速增长所带来的信息抽取挑战。

第二章：系统回顾了本研究相关的技术基础，包括从早期词向量 Word2Vec 到 BERT、GPT 等在内的语言模型演化历程，以及命名实体识别中常见的长短记忆网络 (LSTM) 、条件随机场 (CRF) 等方法。随后介绍直接偏好优化 (DPO) 的核心思想，为后续章节奠定理论基础。

第三章：详细论述了在大型语言模型上进行科学文献命名实体识别的总体思路。首先提出上下文一致的显式标注方法，以在不破坏文本语义的前提下突出实体范围与类别信息。然后，针对实体识别常见的类别错误与边界错误两大难题，引入了正负样本对的训练思路，并依照有效标签数及其比例对样本进行筛选，以保证正负样本的区分度。在此基础上，通过扩张类、收缩类以及多标和类别错误类的负样本构造手段，配合直接偏好优化训练过程，对模型进行更细致的纠错和优化。

第四章：面向材料科学与生物医学的高度专业化领域场景，提出融合领域语言模型的方法，提升模型在不常见专业术语和领域缩写方面的识别能力。通过在具体领域的分析与应用，证明了该方法在复杂领域文献挖掘中的有效性与实践价值。

第五章：对全文进行了归纳与展望。

第二章　相关理论和方法概述

2.1　语言模型

语言模型（Language Model，　LM）是自然语言处理领域的核心技术之一，其目标在于学习并生成合乎语法与语义规律的文本。从早期的 N 元语法模型到基于 RNN、LSTM 的深度神经网络，再到通过 Word2Vec 学习词向量、BERT 利用 Transformer 架构进行双向上下文预训练的模型，逐步克服了远距离依赖和上下文捕捉不足的局限。近年来，大规模语言模型如 GPT-3/4、LLaMA、DeepSeek 等借助自注意力机制和海量参数，实现了更强的文本生成与理解能力，推动了自然语言处理在通用和特定领域的广泛应用。

2.1.1 Word2Vec 及其衍生模型

在深度学习尚未全面兴起之前，传统的统计语言模型（例如 N-gram）往往难以有效捕捉词语间潜在的语义关联，仅能基于相邻词出现频次来衡量语义相似度。Word2Vec 的出现为词向量表示带来里程碑式的发展。Word2Vec 通过在大规模的无监督文本语料上进行训练，成功学到了词和词之间的分布式向量表示，使语义相似度在高维向量空间中得以量化与度量[56]。

Word2Vec 的训练框架主要包含两种核心方法，Skip-Gram 和 CBOW（Continuous Bag-of-Words）。Skip-Gram 的目标是，在给定中心词的情况下去预测它的上下文；这一做法的直观思路是，通过最大化中心词向量与上下文词向量间的内积，让含义相近的词在向量空间中的距离更小，进而实现对语义相似度的捕捉。与之相对，CBOW 模型则采用了逆向思路，给定上下文词来预测中心词。训练时，会先将窗口内的上下文词向量进行平均或加和，然后输入至网络进行预测，从而学习到最能预测中心词的词向量。无论采用 Skip-Gram 还是 CBOW，模型的本质都在于通过大量无监督数据去捕捉词语语义与上下文的潜在关系。

尽管 Word2Vec 在文本分类、机器翻译、信息检索等多种下游任务中取得了明显成效，但它也存在一些局限性。首先，它只能学习词的静态表示，对于

多义词在不同上下文中的含义难以加以区分；其次，单纯用“词”作为最小建模单位，往往无法应对复杂词形变化或未知词（OOV）问题。为克服这些不足，出现了对于 Word2Vec 多种衍生和改进方法，其中最具代表性的便是 fastText。fastText[57]在 Word2Vec 的框架上进一步引入了“子词”概念，即将每个词拆分为字符级别的 n-gram 片段并为之学习向量表示；当遇到新词或罕见词时，模型能够通过已有的子词 n-gram 表示来生成近似向量，从而有效缓解未登录词的问题，也更好地处理了形态丰富语言中常见的词形变化。fastText 与 Word2Vec 在训练流程上大体一致，同样能够结合 Skip-Gram 或 CBOW 并使用负采样来降低计算开销，但额外的子词向量存储会在大型词典场景下带来一定的内存开销。总体而言，Word2Vec 及其衍生模型为自然语言处理中的词向量学习奠定了基础，让模型可以更好地捕捉词与词之间的语义或语法关联，并在随后的深层预训练网络如 BERT、GPT 等出现之前，极大推动了自然语言处理任务在表达能力和下游性能方面的提升。

2.1.2 BERT 模型

随着人们对深层语言特征与丰富上下文表示能力的需求不断提高，如何让模型捕捉到更准确且更具泛化能力的语义信息，成为了自然语言处理领域的核心课题之一。2018 年，由 Google 提出的 BERT[13]在此背景下应运而生。BERT 采用了预训练与微调相结合的框架：先在大规模无监督文本语料上进行预训练，再在具体下游任务上进行微调，从而在自然语言推断、情感分析、阅读理解等大多数主流测试中都取得了显著性能提升，标志着深层双向预训练模型时代的正式到来。

如图 2.1 所示，在 BERT 的训练阶段，掩码语言模型（Masked Language Modeling，MLM）与下一句预测（Next Sentence Prediction，NSP）是两项至关重要的训练任务。前者通过在输入序列中随机遮蔽部分单词，让模型同时“向左看”和“向右看”，进而根据双向上下文去预测被掩盖的词。由于模型可以借助全局上下文信息完成还原，这一过程使得 BERT 能够学习到更深入的双向语义表示；后者则通过随机选取一对句子来判断它们在原文中是否相邻，从而帮助模型理解句子级别的连续性和连贯性，对于多句子输入场景下的上下文

衔接和跨句推理至关重要。这两类训练目标的设计，使 BERT 在理论与实践上获得了比传统单向语言模型更完整的语义理解能力。

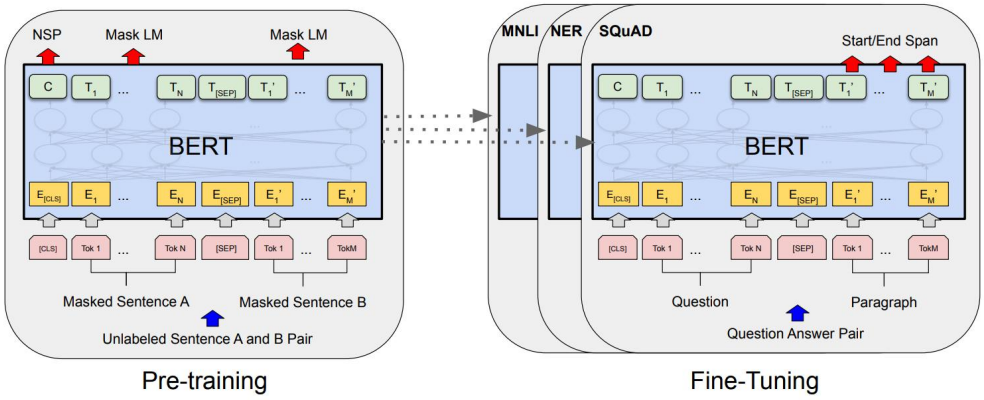


图 2.1 BERT 的整体预训练和微调过程[13]

在模型结构上，BERT 采用了基于 Transformer Encoder 的堆叠式架构，每一层都由多头自注意力和前馈网络组成，并通过残差连接与 LayerNorm 保证梯度传递的稳定性与训练的快速收敛。多头自注意力机制会将输入序列投影到多个不同子空间中，以捕捉句子中各词与其他词之间各种不同类型的关联，从而实现更加多样的语义对齐。此外，每一层的残差连接与归一化方法不仅帮助模型在深度堆叠时保持稳定，也在一定程度上缓解了训练初期梯度消失或梯度爆炸的问题。

BERT 的出现极大拓展了深度表征学习在自然语言处理中的应用空间，面向不同领域或语种的 BERT 模型也被提出。例如，在材料领域，研究者们基于材料文献进行再训练，提出了 MatSciBert[6]，用于材料文本分析任务；在金融领域，FinBERT[58]通过大量财经报道和股票市场新闻等数据进行预训练，以更好地捕捉财经文本中的行业术语与专家知识；法律领域则诞生了 LEGAL-BERT[59]，通过对法律条文、判例和各类法律文件进行大规模训练，提升了在法律文书分类、案件检索和判决预测等专业任务上的表现；其他如 ClinicalBERT[60]、SciBERT[39]等针对临床医学、科研论文等领域也在不断涌现。它们都继承了 BERT 双向预训练的核心思想，并在相应语料上进一步学习垂直

领域的语言特性，明显提升了专门场景的模型效果。正是通过这些扩展与改进，BERT 及其各种变体在几乎所有下游任务中都发挥了巨大的价值，也奠定了多任务、多领域预训练模型的基础思路。

2.1.3 大语言模型

在 Transformer 架构取得成功之后，研究者们不断尝试通过扩大模型规模与预训练数据量来强化模型的表达与泛化能力，提出了 GPT[14]、GPT-2[15]、GPT-3[61]、GPT-4[28]、PaLM[27]、LLaMA[62]等一系列大语言模型。这些模型通常拥有从百亿到千亿甚至上万亿级别的参数，通过海量无监督文本的预训练，使得在最少标注甚至零样本情况下，模型就能展现出惊人的文本生成与推理能力。与主要面向语言理解的双向编码模型 BERT 不同，大语言模型大多采用自回归训练方式，即在给定前文或上下文条件下不断预测下一词或下一步输出，这使得它们天然适用于文本续写和生成任务。通过这种连续序列建模方式，模型能够学习到丰富的语言规则，并在多个下游任务中表现出较高的一致性与泛化能力。

同时，在大语言模型的研究与应用中，提示学习逐渐兴起并发挥了关键作用[63]。研究者们发现，通过巧妙设计上下文提示，可以让同一模型在无需为每个任务单独微调的前提下，灵活适应数十乃至上百种下游任务，充分体现了其在零样本与小样本场景下的潜能[64]。此外，为了进一步提高模型在对话场景中的连贯性与安全性，近年出现了基于人类反馈强化学习 (Reinforcement Learning from Human Feedback， RLHF）的优化方法，该方法通过采集人类对模型输出的偏好信息来更新模型参数，从而使生成的回答更符合人类期望，减少有害或不恰当的内容[65]。

然而，由于大语言模型具有庞大的参数量，直接对其进行完整训练和微调所需的计算资源和数据成本极高，因此在某些特殊场景或资源受限的情况下，研究者提出了多种参数高效微调方法[66]。其中，LoRA[67]作为一种代表性方法如图 2.2，通过在已有预训练模型的基础上添加少量可训练的低秩矩阵来进行微调，从而使得大部分原始参数保持冻结状态，仅对额外的低秩参数进行更新。

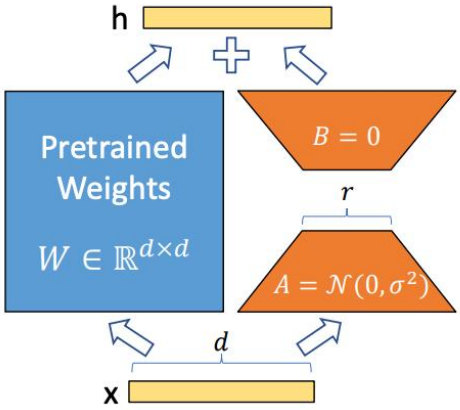


图 2.2 LoRA 的重新参数化[67]

具体而言，设预训练模型中某个需要更新的权重矩阵为$W_0 \in \mathbb{R}^{d\times k}$，LoRA 用低秩分解来约束它的更新：

$$W_0 + \Delta W = W_0 + BA \ , \tag{2.1}$$

这里$B \in \mathbb{R}^{d\times r}$和$A \in \mathbb{R}^{r\times k}$是两个低秩矩阵，且秩 $\mathrm{r} \ll \min(\mathrm{d}, \mathrm{k})$。训练时冻结$W_0$，仅对$A$和$B$进行梯度更新。初始时令$B = 0$，$A \sim N(0, \sigma^2)$，保证初始模型行为与原模型一致且更新幅度可控。这种低秩分解不仅显著降低了更新参数的规模和显存占用，同时也在一定程度上保持了原模型的预训练表示，使得模型可以在较低成本下灵活适应新的任务需求。

值得一提的是，当前大语言模型的技术路径正呈现多元化发展。以 DeepSeek[68]为代表，在保持自回归生成框架的基础上，在维持自回归生成能力的同时，通过动态语义索引机制实现生成与检索的协同优化。总体而言，大语言模型凭借超大规模参数和海量预训练数据，实现了从通用语言理解到文本生成的跨任务统一。然而，其在虚假信息生成、偏见、伦理风险以及高昂的计算资源需求等方面也提出了严峻的挑战。

2.2 命名实体识别框架

2.2.1 面向科学文献的命名实体识别

命名实体识别（Named Entity Recognition， NER）是自然语言处理的重要任务，主要目标是从非结构化文本中准确提取出具有特定意义的实体，并为关系抽取、事件检测和知识图谱构建等后续任务提供基础数据支持。为实现这一目标，命名实体识别不仅要求在语法和语义层面具备精确的识别能力，还需要对上下文信息进行详细分析，确保每个实体与预定义类别正确对应。近年来，深度学习技术为命名实体识别带来了新的突破。以 BiLSTM-CRF[69]模型为代表的神经网络方法，利用双向长短期记忆网络捕捉上下文信息，再借助 CRF 层对标签序列进行全局解码，有效解决了传统方法难以捕捉长距离依赖的问题。同时，预训练语言模型的引入，使模型能够从海量无标签数据中自动学习丰富的语义表示，进一步减少了对人工特征的依赖，并在多个领域和多种语言环境中展现出优异性能。

在有监督学习框架下，命名实体识别模型通常需要大量高质量标注数据进行训练，以确保高准确率和稳定性。然而，标注数据获取成本较高，尤其在低资源语言或专业领域中。为解决这一问题，无监督和自监督学习方法开始受到关注，它们利用未标注数据挖掘语义结构，为数据不足的场景提供支持。此外，少样本和零样本学习技术也显示出应用前景：少样本学习利用少量标注样本帮助模型迅速适应新领域，而零样本学习则尝试在完全没有标注样本的情况下，通过跨任务迁移或依靠大规模预训练模型识别全新类别[70]。

尽管上述方法在缓解数据稀缺和提升领域适应性方面取得了一定进展，但当前命名实体识别研究仍面临不少挑战。数据标注成本高、不同领域间语言表达和语义差异较大、多模态信息融合复杂，以及少样本和零样本学习在模型稳定性和精细识别上的局限性，这些问题仍需要进一步解决。

2.2.2 长短记忆网络

长短期记忆网络（LSTM）[10]是一种特殊的循环神经网络，用于解决传统循环神经网络在捕捉长期依赖关系时容易出现的梯度消失和梯度爆炸问题。通

过引入门控机制，LSTM 能够在时间序列中动态控制信息的保留与遗忘，从而更好地捕捉长时段内的信息依赖。如图 2.3 所示，LSTM 单元主要由三个门组成，即遗忘门、输入门和输出门，以及一个单元状态。

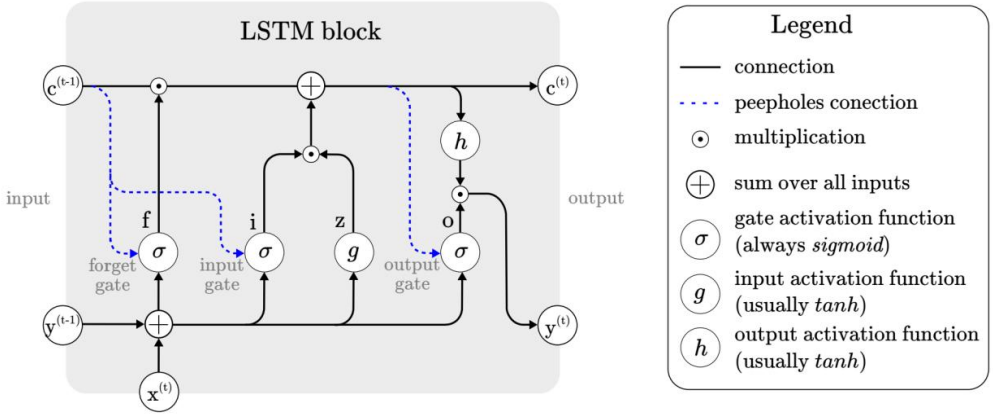


图 2.3　LSTM 架构[71]

遗忘门用于决定在当前时刻保留多少上一时刻的单元状态，设输入向量为$\mathrm{x_t}$，上一时刻的隐藏状态为h_{t-1}，则遗忘门的计算公式为：

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f), \tag{2.2}$$

其中，$\sigma(\bullet)$表示 Sigmoid 激活函数，其输出范围为[0，1]，表示保留信息的比例；W_f和b_f分别为权重矩阵和偏置项。

输入门确定当前输入中哪些信息需要写入单元状态，而候选状态则生成待添加的新信息。计算过程包括两个步骤：

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i), \tag{2.3}$$

$$\widetilde{C}_t = \tanh(W_C \bullet [h_{t-1}, x_t] + b_C), \tag{2.4}$$

其中，i_t表示输入门的控制因子，$\widetilde{C}_t$表示当前输入生成的候选单元状态，双曲正切函数 tanh 将候选状态的值映射到[-1，1]。单元状态的更新结合了遗忘门和输入门的作用，具体公式为：

$$\widetilde{C}_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t, \tag{2.5}$$

其中，$\odot$ 表示逐元素乘法，C_{t-1}为为前一时刻的单元状态。该过程保证了模型可以在较长序列中保持关键信息，同时舍弃不相关的信息。输出门控制当前时刻最终输出的隐藏状态，计算公式为：

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o), \tag{2.6}$$

$$h_t = o_t \odot \tanh(C_t), \tag{2.7}$$

隐藏状态h_t不仅作为当前时刻的输出，同时也传递到下一时刻，用于捕捉序列中长距离的依赖关系。

2.2.3 条件随机场

条件随机场（Conditional Random Field， CRF）是一种判别式概率模型，专用于序列标注任务。与生成模型（例如隐马尔可夫模型）不同，CRF 直接建模条件概率$P(Y|X)$，从而充分利用输入特征信息，同时考虑输出标签之间的依赖关系。这种全局建模方法能够确保整个序列的预测结果具有更高的一致性和准确性。

对于给定的输入序列$X = (x_1, x_2, \ldots, x_n)$以及对应的标签序列$Y = (y_1, y_2, \ldots, y_n)$，线性链 CRF 模型定义条件概率如下：

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{t=1}^{n} \sum_k \lambda_k f_k(y_{t-1}, y_t, X, t)\right), \tag{2.8}$$

$f_k(y_{t-1}, y_t, X, t)$为特征函数，用于描述标签转移和输入特征之间的关系，λ_k为相应特征函数的权重参数，通过训练数据学习得到；$Z(X)$是归一化因子，定义为：

$$Z(X) = \sum_{Y'} \exp\left(\sum_{t=1}^{n} \sum_k \lambda_k f_k(y'_{t-1}, y'_t, X, t)\right), \tag{2.9}$$

归一化因子确保所有可能标签序列的概率和为 1。在 CRF 模型中，通常通过最大化条件对数似然来训练参数，这可以转化为最小化负对数似然损失。简化后的损失函数公式为：

$$\mathcal{L}(\theta) = -\ell(\theta) = -\left(\sum_{t=1}^{n}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, X, t) - \log Z(X)\right), \quad (2.10)$$

其中，θ表示模型参数（即各特征权重λ）。在训练过程中，利用梯度下降等优化算法对上述损失函数进行最小化，从而不断调整参数，使模型输出的标签分布尽可能接近真实分布。梯度计算时会对比真实标签序列在各时刻所对应的特征与模型预测下的特征期望，从而为参数更新提供方向。

在推理阶段，训练好的 CRF 模型需要在给定输入序列 X 的情况下，找到使得$P(Y|X)$最大的标签序列。由于模型的结构允许将整体得分分解为各时刻的累加形式，因此使用动态规划算法通过递推的方式计算每个位置上各个标签的最优累积得分，最终利用回溯方法得到全局最优的标签序列。这样既保证了局部得分的合理性，也确保了整个序列预测的全局一致性。

2.3 直接偏好优化

在以往需要从人类偏好中训练方法的实践中，通常会将所有偏好数据先转换成某种奖励模型，然后再在一个含有 KL 约束的优化目标下使用强化学习算法进行微调。这样的流程往往涉及额外的建模与两阶段训练，既耗费时间也容易在方法更新中出现不稳定。直接偏好优化，即 Direct Preference Optimization（DPO）[55]方法试图绕过这个繁琐过程，如图 2.4 所示；它利用一个重参数化思想，将最优方法与隐式奖励建立起指数对应关系，从而直接用偏好数据来训练最终的方法模型，不再需要显式的奖励网络或强化学习过程。

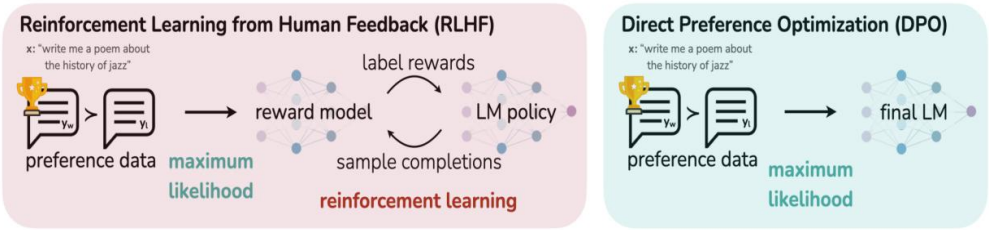


图 2.4　DPO 通过优化人类偏好来避免强化学习[55]

为了说明这一思路，可以先将最优方法假设为：

$$\pi^*(y|x) \propto \pi_{ref}(y|x)\exp\left(\frac{1}{\beta}r^*(x,y)\right), \tag{2.11}$$

其中$\pi_{ref}(y|x)$是一个参考分布，β为控制平衡项，而r^*则代表在人类偏好之下的“真实”或“最优”奖励函数。这样一来，如果公式取对数，就可以把每个回答(x,y)的的奖励值直接视作“对数概率比”形式:

$$r^*(x,y) \approx -\beta\log\frac{\pi^*(y|x)}{\pi_{ref}(y|x)}, \tag{2.12}$$

，

由于人类偏好在本质上常用一对对回答进行比较，直接偏好优化会让两个输出(y_w, y_l)的奖励差进入对比损失中，从而把通常需要训练奖励函数和做强化学习的过程简化成一次带对比的交叉熵优化。更具体地说，人类偏好可以被视为一组三元组$\{(x, y_w, y_l)\}$，其中y_w是偏好更高的回答，y_l是偏好更低的回答。直接偏好优化在建模时直接令方法π_θ同参考分布做比率, 带入一个对数差再经过逻辑函数，就能将判断y_w优于y_l的概率写为:

$$Pr(y_w \succ y_l|x) = \sigma\left(-\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} + \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right), \tag{2.13}$$

其中$y_w \succ y_l$表示在人类偏好标注中, 回答y_w被认为优于回答y_l。接着用一项简单的交叉熵或者极大似然目标去拟合方法π_θ, 让它倾向于给y_w更高的概率, 给y_l较低的概率。这个构造等价于在一个指数形式的奖励下做 KL 约束的最大化。

因此直接偏好优化实际上直接跳过了显式的奖励模型训练与强化学习循环, 却仍能达到相当的效果。它不仅让训练流程大幅简化，而且在实验中也展现出较好的训练稳定性和泛化能力。通过直接偏好优化, 可以将对人类偏好的采样、标注和模型微调一次性打通，避免了先学奖励网络再做强化学习的一系列繁琐步骤，从而让偏好对齐过程更为直观和高效。

2.4 评价指标

本章采用的评价指标为评估文本生成任务的 BLEU-4 和 ROUGE 系列（ROUGE-1、ROUGE-2 和 ROUGE-L），以及命名实体识别任务常用的精确率，召回率和 F1 指标。BLEU-4 侧重于生成文本与参考文本在不同层次 n-gram 的匹配，强调精度；而 ROUGE 系列指标主要度量召回率，反映生成文本对参考文本的覆盖情况，其中 ROUGE-1 关注单词级别，ROUGE-2 关注二元组级别，ROUGE-L 则通过最长公共子序列综合反映文本整体结构的一致性，为评价文本生成系统提供了多角度、定量化的评估标准如表 2.1。精确率，召回率和 F1 指标则是同来评估模型在命名实体识别任务上的表现。

表 2.1 文本生成任务的指标

指标	定义
BLEU-4	基于 n-gram 匹配的评估指标，用于衡量生成文本与参考文本间的相似性。
ROUGE-1	评估生成文本与参考文本在单词级（unigram）上的匹配情况。
ROUGE-2	衡量生成文本与参考文本在二元组（bigram）层面的匹配情况。
ROUGE-L	基于最长公共子序列（LCS）的指标，衡量文本间整体结构连续性的匹配。

衡量命名实体识别任务的指标精确率 P，召回率 R 和 $F1$ 值公式如下：

$$P = \frac{TP}{TP + FP}, \tag{2.14}$$

$$R = \frac{TP}{TP + FN}, \tag{2.15}$$

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{2.16}$$

TP（真正例）代表模型正确预测为属于相应实体类别的样本数量。FP（假正例）表示模型错误地将不属于该实体类别的样本预测为该类别的数量。FN（假负例）表示本应属于特定实体类别的样本，但被模型错误地预测为其他类别或未被识别的数量。精确率（Precision）衡量的是模型对正例预测的准确性，即在模型所预测的正例中真正例所占的比例；而召回率（Recall）则衡量模型捕获实际正例的能力，即模型正确识别出的真正例占所有真实正例的比例。$F1$ 值作为精

确率与召回率的调和平均值，通过同时考虑这两个指标，提供了一个较为平衡的模型性能度量方式。

2.5 本章小节

本章回顾了自然语言处理的核心理论和关键方法。首先，梳理了语言模型的发展历程，介绍了词向量模型 Word2Vec 及其衍生的 fastText，然后是预训练模型 BERT 再到最近的大语言模型，展示了模型在捕捉语法和语义规律方面的不断进步。接着，在命名实体识别部分，本章回顾了基于规则和统计方法的传统命名实体识别方法，以及基于深度学习改进方法，并介绍了 LSTM 与 CRF 在序列标注中的基本原理与应用。最后，讨论了直接偏好优化方法。利用人类偏好数据，通过对比学习简化强化学习流程，为大语言模型的偏好对齐提供了一种简洁高效的解决方法。

第三章 基于大语言模型的科学文献命名实体识别

在命名实体识别任务中，尽管大语言模型（LLM）在理解和生成自然语言文本方面取得了显著的成果，但在命名实体识别任务时，仍面临着精确识别实体的挑战。主要原因在于，大语言模型一般为生成式模型，对于需要对实体进行精确标注的命名实体识别任务，生成式模型的输出方式和序列标注任务之间的差异使得模型在面对词级精细标注时，难以保持对实体类别和边界的严格约束。此外，生成模型在处理实体识别时，还常常面临“幻觉”问题，即模型可能生成一些并不存在的实体，这进一步加剧了命名实体识别任务的难度。为了解决这些问题，本章提出了一种上下文一致的实体显式标注方法。通过插入成对的特殊标记，有效地突出实体类别和边界，在不破坏原文本上下文结构的条件下，使得模型能在理解文本语义的同时，准确地识别实体。在训练阶段，先进行监督微调，然后通过直接偏好优化结合正负样本对进行精细化约束。借助两阶段的训练方法，模型既能保留在监督微调阶段学得的知识分布，又能在直接偏好优化阶段更准确地倾向于正确标注，实现更稳健的命名实体识别效果。

3.1 方法概述

本章以 Llama3.1-8B-chat 作为基础大语言模型，提出了一种上下文一致的实体显式标注方法，通过在文本中显式插入成对的特殊符号突出实体边界与类别信息，使模型在识别实体时能够兼顾文本的语义完整性与对实体范围的明确标注。随后在训练环节，先以大规模标注数据进行监督微调，帮助模型掌握基础的实体识别能力；紧接着，通过直接偏好优化在同一输入条件下比较正负样本的对数概率差，使模型在保持原有知识分布的同时，进一步减少对实体类别和边界的错误识别。为辅助直接偏好优化训练，构造三种负样本场景，包括通过扩张或收缩实体边界，并通过对监督微调阶段的真实错误样本进行采样筛选，模拟模型易犯的“多标”与“漏标”错误，生成类别混淆的负样本，以便在正负样本的对比中强化模型对不确定实体的识别能力。整体方法如图 3.1。

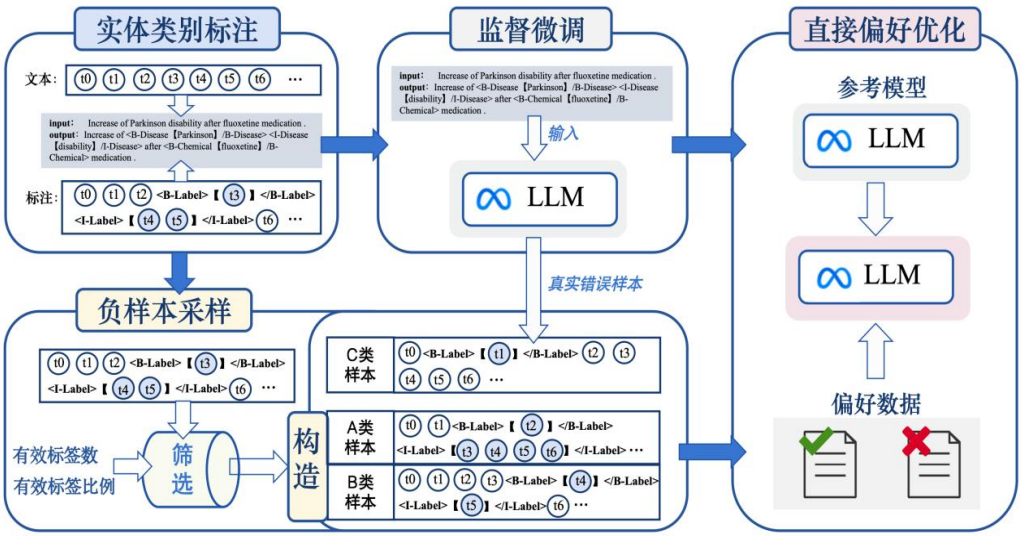


图 3.1 上下文一致的实体显式标注和结合直接偏好优化负样本构造的双阶段训练方法

3.2 上下文一致的实体显式标注方法

传统的实体标注方法主要分为 BIO 标注方式[72]和跨度表示方式两类[73]。在 BIO 标注方式中[74]，文本中的每个词都被标注为三种标签之一：B-X（实体的开始），I-X（实体的内部），以及 O（非实体）。这种标注方式通过标明每个词的实体类别，帮助模型进行词级别的处理。跨度表示方法通过标记实体的起始位置和结束位置来表示一个完整的实体，避免了对每个词的重复标注。在跨度表示中，一个实体可以用其开始和结束位置的索引对来表示，其中一个是实体的开始位置，一个是实体的结束位置，一个实体可以用跨度表示为其对应索引和实体类别的三元组形式。

为了更好地让大语言模型与命名实体识别任务适配，本章结合以上两种标注方法的优点，提出了一种上下文一致的实体显式标注方法，让标注与原句的上下文共存于同一序列中，在不破坏原有上下文的前提下，采用一种显式标注方法，将实体类别与实体部分显式绑定来突出实体范围和类别信息，使得模型在生成过程中能够自然地聚焦于这些标记区域，同时避免模型过分关注标记的形式而忽略实体内容的语义信息。具体来说，给定一个输入文本序列：

$$t = (t_0, t_1, \ldots, t_n), \tag{3.1}$$

其中t_i表示序列中的第i个词，假设此序列中的一个实体是由 m 个词组成，记该实体为：

$$E = (t_i, t_{i+1}, \ldots, t_{i+m-1}), \tag{3.2}$$

其中，t_i是实体的开始词，t_{i+m-1}是实体的结束词。

采用 BIO 的对实体的标注思想，把实体分为开始 B-和内部 I-两部分，来突出类别信息；为了标记实体的边界和范围，分别对实体两部分的开始和结束位置进行显式标注，用一种注入实体类别的成对括号方式来包裹相应的实体信息，实体标注的形式可以表示为如图 3.2 所示：

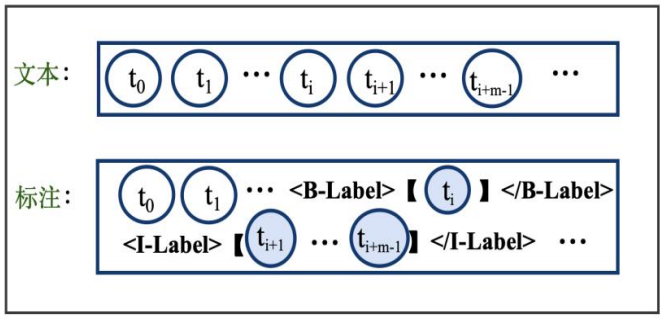


图 3.2 上下文一致的实体显式标注

3.3 结合直接偏好优化负样本构造的双阶段训练方法

本章提出一种双阶段训练方法，在基于监督微调获得稳定基础模型后，利用直接偏好优化进一步整合人类偏好信号，从而提高模型在实体识别上的精细表现。该方法的核心目标是在保留监督学习阶段已学得知识分布的同时，通过直接偏好优化对模型输出进行约束，减少错误判断，提升生成结果的质量。

3.3.1 监督微调阶段

在第一阶段中，对基础模型进行监督微调，此时模型通过标准的交叉熵损失函数在大量标注数据上进行优化，学习到对给定输入生成正确回答的能力。具体而言，对于一个训练样本，没有任何标注的的原文本作为输入，输出是标

注了命名实体标签的目标回答.在模型训练过程中，采用条件生成概率来指导模型学习，目标是最大化模型在给定输入下生成正确的实体标注。生成的文本可以表示为：

$$y = \{y_1, y_2, \ldots, y_i, \ldots, y_n\}\ , \tag{3.3}$$

其中y_i表示模型生成回答的第i个词。模型在生成过程中根据前一个词生成下一个词，并计算每个词的条件概率：

$$P(y|t) = \prod_{i=1}^{T} P(y_i|t, y_1, \ldots, y_{i-1})\ , \tag{3.4}$$

这里，$P(y_i|t, y_1, \ldots, y_{i-1})$是模型在位置$i$的生成概率，$t$表示输入序列。在进行监督微调时，输入序列会通过模型进行前向传播，并计算每个位置的 logit 向量z_i。该向量的维度等于词汇表的大小|V|，其中每个元素$\mathrm{z}_{i,j}$表示位置i上对应词汇表 V 中第j个词的未经归一化分数。随后用 softmax 函数将此 logit 向量z_i转换为概率分布。这个概率分布中目标词y_i对应的概率值等于模型在位置i的生成概率$P(y_i|t, y_1, \ldots, y_{i-1})$。具体公式为：

$$P(y_i|t, y_1, \ldots, y_{i-1}) = \frac{\exp(z_{t,y_i})}{\sum_{j=1}^{|V|} \exp(z_{i,j})}, \tag{3.5}$$

其中，z_{i,y_i}是 logit 向量z_i中与目标词y_i相对应的 logit 值，$\sum_{j=1}^{|V|} \exp(z_{i,j})$是对词汇表$V$中所有词对应的 logit 值取指数后求和，作为归一化项。对整个输出序列，模型将各位置的对数概率累加，得到完整序列y的对数概率：

$$\log P(y|t) = \sum_{i=1}^{n} \log P(y_i|t, y_1, \ldots, y_{i-1})\ 。 \tag{3.6}$$

在训练阶段，以上述对数似然的负值作为交叉熵损失，通过反向传播优化模型参数，使其在输出时能够更精准地生成符合标注的答案或实体标签。

3.3.2 直接偏好优化阶段

在第一阶段监督微调训练的基础上，为进一步降低错误实体识别的概率，采用直接偏好优化进行精细化微调。并将第一阶段经过监督微调的模型参数固定，作为参考模型P_{ref}，来防止在偏好优化过程中大幅偏离监督训练阶段学得的知识分布。

直接偏好优化[55]的基本思想是利用成对的偏好数据，即人类标注偏好较高的正例（chosen）和偏好较低的负例（rejected），直接优化模型在相同输入下对正负回答的概率差异，从而使模型更偏向于生成高质量回答，因此每个样本包含同一输入下的两个候选回答。具体到本研究的命名实体识别任务，正例是包含了准确无误的实体标注的输出序列，而负例则是依据特定方法构造的、模拟了常见标注错误，如将实体边界错误地扩张或收缩、将实体错误地归类到其他类别、将非实体错标为实体即“多标”、或未能识别出本应存在的实体即“漏标”等的输出序列。对于每个回答y无论正例回答还是负例回答，在给定输入x下，根据公式 3.4 计算生成的对数概率。然后，定义回答的奖励为：

$$r(y,t)=\log P_{\theta}(y|t)-\log P_{ref}(y|t), \tag{3.7}$$

其中P_{θ}为当前待更新的模型，初始状态同P_{ref}，在直接偏好优化阶段中参数被持续优化，以学习人类偏好。对于同一样本中的正例回答和负例回答，计算奖励差Δ：

$$\Delta=r(y^{+},t)-r(y^{-},t), \tag{3.8}$$

这里的y^{+}是正例回答，y^{-}是负例回答。如果$\Delta>0$，表示对正例回答的对数概率高于对负例回答的；$\Delta<0$ 则反之亦然。为了将该差值映射到概率空间，并通过交叉熵损失进行优化，于是引入温度参数β并通过 Sigmoid 函数来计算偏好概率p_{f}：

$$p_f = \sigma(\beta \bullet \Delta) = \frac{1}{1 + e^{-\beta\Delta}}, \tag{3.9}$$

其中$\sigma(\bullet)$表示 Sigmoid 函数。温度系数β用来控制奖励差Δ的放大/收缩力度，从而决定 Sigmoid 函数的陡峭程度。对直接偏好优化阶段的损失函数则定义为：

$$\mathcal{L}_{DPO} = -\log(p_f), \tag{3.10}$$

为了保证在直接偏好优化训练阶段时，模型依然保持的原有监督微调知识，加入平衡因子γ用于在直接偏好优化损失中施加对监督微调模型的约束或正则化，具体公式如下：

$$\mathcal{L} = -\log(p_f) + \gamma \cdot \mathrm{KL}(P_\theta \| P_{ref}), \tag{3.11}$$

其中P_{ref}是经过监督微调阶段微调得到的参考分布，$\mathrm{KL}(P_\theta \| P_{ref})$为 KL 散度正则项，用来约束模型不要过度偏离初始的监督微调分布。在联合优化过程中，当模型为了捕获并放大这种奖励差Δ，进而试图显著降低偏好损失$-\log(p_f)$而导致输出分布偏离P_{ref}过大，那么 KL 散度项的值就会显著上升。此时，KL 项的梯度也会同步增大，形成一个有效的“拉回”约束。这个约束机制会抑制模型参数向着可能导致与P_{ref}过大偏离的方向更新，即便那个方向可能在当前批次数据上带来更大的奖励差Δ。反之，当偏离较小时，KL 项的惩罚减弱，模型可在更大搜索空间内继续优化偏好损失。这样既保证了偏好一致性的提升，又保持了与原有知识的连贯性与稳定性。γ为用来控制约束强度的超参数，γ较大意味着对 KL 散度的惩罚更强，模型更倾向保留监督微调阶段分布，γ较小则允许模型在直接偏好优化阶段有更大幅度的调整。

整体算法流程如算法 3.1。

算法 3.1：DPO 训练流程	
	输入：包含 (输入 t, y+, y-) 的样本对 D，初始模型 M
	输出：更新后的模型 M_θ
1.	for each (t, y+, y-) in D:
2.	计算 $P_\theta \leftarrow \log P(y \mid t; M\_ref)$;
3.	计算 $P_{ref} \leftarrow \log P(y \mid t; M\_ref)$;
4.	计算奖励 $r(y, t) \leftarrow \log P_\theta(y\|t) - P_{ref}(y\|t)$;
5.	计算奖励奖励差 $\Delta \leftarrow r(y+, t) - r(y-, t)$;
6.	P_pref $\leftarrow$ sigmoid(Δ)
7.	根据损失函数$\mathcal{L}$计算梯度并进行反向传播
8.	更新模型参数 M_θ
9.	return M_θ

3.3.3 采样方法

为了让监督微调阶段和直接偏好优化阶段输入输出一致，让正负样本的上下文保持不变，只有实体的标注类别或者位置发生变化。考虑到正负样本之间的差异需要足够明确才能让模型在直接偏好优化训练阶段继续学习，所以本章提出了一种采样方法，依据有效标签数及其比例对样本进行筛选，留下标签占比较大的样本，达到加大正负样本之间的差异的效果。具体来说，对训练数据中每个样本进行标注统计，并依据有效标签数及其比例对样本进行筛选。设定每个样本包含一个 BIO 标签序列$Y = \{y_1, y_2, \ldots, y_N\)$，其中$N = |Y|$表示该样本中的总标签数，而$y_i$表示第$i$个词对应的标签。对于每个样本定义有效标签数$E$和有效标签比例$R$ ：

$$E = \left|\{y_i \in Y \mid y_i \neq \text{"}O\text{"}\}\right|, \tag{3.12}$$

$$R = \frac{E}{N}, \tag{3.13}$$

对每条样本按照以上公式计算总标签数N、有效标签数E以及有效标签比例R，并根据设定的阈值对样本进行筛选采样。

3.3.4 直接偏好优化阶段的负样本构造

本章把目前的命名实体识别任务面临的难题分为两类。

第一类是实体类别错误，此类错误可进一步细分为实体与非实体之间的识别错误以及实体内部类别混淆。具体而言，实体与非实体之间的识别错误指模型在判断文本片段是否应视为预定义实体时发生误判。这可能导致漏标，即本应标注的实体被错误地识别为非实体，例如，在生物医学文本中将疾病名称“diabetes mellitus”错误地标注为代表非实体的“O”标签；也可能导致错标或多标，即将本身不属于任何预定义实体类别的词语错误地判定为某个实体类别，例如，将普通名词“treatment”误标为“Chemical”或“Disease”。实体内部类别混淆则发生在模型已正确识别出实体边界，但却将其归属到错误的预定义类别中的情况，例如，将化学（Chemical）实体“aspirin”错误地分类为疾病（Disease）类别。

第二类是实体边界识别错误，指模型在确定实体于文本中的准确起始和结束位置时产生偏差。此类错误可能表现为边界收缩，即未能完整识别实体跨度，例如，对于疾病实体“diabetes mellitus”，模型可能仅识别并标注了“mellitus”部分；或表现为边界扩张，即将实体邻近的非实体词语错误地纳入实体范围，例如，将短语“aspirin was”整体识别为一个化学（Chemical）实体。这些边界错误会导致实体被不当切分或与邻近文本合并，从而降低整体识别性能[75][76]。针对以上的问题，提出了以下来三种负样本的构造方法。对每条样本按照计算总标签数 N、有效标签数 E 以及有效标签比例 R，并根据设定的阈值对样本进行筛选。

（一）扩张类负样本构造方法

对于命名实体识别中出现的边界过度预测，原始实体区域被有意地扩展以包含额外的信息，从而模拟可能由于模型预测偏差而产生的实体过度预测现象，并构造出扩张类负样本。

算法 3.2 给出了以上构造方法（一）的算法流程。

算法 3.2：扩张类负样本构造

输入：实体标签序列 E, $e_i \in E$
输出：扩张后的负样本实体标签序列 E_{new}

```
for e_i in E:
    实体左边界 s ← e_i.start_idx;
    实体右边界 t ← e_i.end_idx;
    向左尝试扩展一词 s' ← (s > 0 ? s−1 : s);
    向右尝试扩展一词 t' ← (t < |words| ? t+1 : t) ;
    if s' ≠ s or t' ≠ t then:                    // 区域重新标注
        ner[s'] ← "B - " + e_i.label             //新区间首词标为 B - 标签
        for i = s'+1 to t'−1:                    //其余词标为 I - 标签
            ner[i] ← "I - " + e.label;
        end for
    end if
end for
return E_new
```

对于方法（一）构造的扩张类负样本，简称为 A 类样本，具体输入示例如图 3.3：

input: We introduce a new interactive corpus exploration tool called InfoMagnets .

chosen: We introduce a new <B-Method【interactive】/B-Method> <I-Method【corpus exploration tool】/I-Method> called <B-Method【InfoMagnets】/B-Method> .

rejected: We introduce a <B-Method【new】/B-Method> <I-Method【interactive corpus exploration tool】/I-Method> <B-Method【called】/B-Method> <I-Method【InfoMagnets .】/I-Method>

图 3.3 扩张类负样本（A 类样本）

（二）收缩类负样本以及“漏标”类负样本构造方法

对于实体边界捕捉不全而导致预测边界偏小和实体“漏标”问题，对实体边界进行边界收缩，构造收缩类负样本。

算法 3.3 给出了以上构造方法（二）的算法流程。

算法 3.3：收缩类及“漏标”负样本构造

输入：实体标签序列 E, $e_i \in E$
输出：扩张后的负样本实体标签序列 E_{new}

```
for e_i in E:
    实体左边界 s ← e_i.start_idx;
    实体右边界 t ← e_i.end_idx;
    实体长度 length← t – s ;   //若存在两个或以上单词实体，则过滤该样本
    if length > 1 then              // 区域重新标注
        排除此样本          //新区间首词标为 B - 标签
    end if                  //新区间首词标为 B - 标签
    if length = 1 then      //作为漏标负样本：将单词实体置为非实体
        ner[s] ← "O";
    elseif length = 2 then          //收缩为单词：删除右侧 I-标签
        ner[s + 1] ← "O" ;
    else
        ner[s] ← "O" ;                  // a. 原 B-置 O
        ner[s + 1] ← "B-" + e_i.label;  // b. 左边界收缩：下一个词 标为新的 B-
        for i = s + 2 to t - 2:         //c. 中间部分保持 I-
            ner[i] ← "I-" + e_i.label;
        end for
       ner[t - 1] ← "O"                 //d. 右边界收缩：删除最后一个 I-标签
    end if
end for
return E_new
```

对于方法（二）构造的收缩类负样本以及“漏标”类负样本，简称为 B 类样本，具体输入示例如图 3.4：

input: We introduce a new interactive corpus exploration tool called InfoMagnets .

chosen: We introduce a new <B-Method 【interactive】 /B-Method> <I-Method 【corpus exploration tool】 /I-Method> called <B-Method 【InfoMagnets】 /B-Method> .

rejected: We introduce a new interactive <B-Method 【corpus】 /B-Method> <I-Method 【exploration】 /I-Method> tool called InfoMagnets .

图 3.4 收缩类负样本以及“漏标”类负样本（B 类样本）

（三）“多标”类负样本以及类别错误负样本构造方法

对于非实体“O”错误标为实体的“多标”问题，采用监督训练后的模型对验证集的数据进行推理，然后对推理出错的数据随机采样，作为“多标”类负样本以及类别错误负样本。对于方法（三）构造的“多标”类负样本以及类别错误负样本，简称为 C 类样本，具体输入示例如图 3.5：

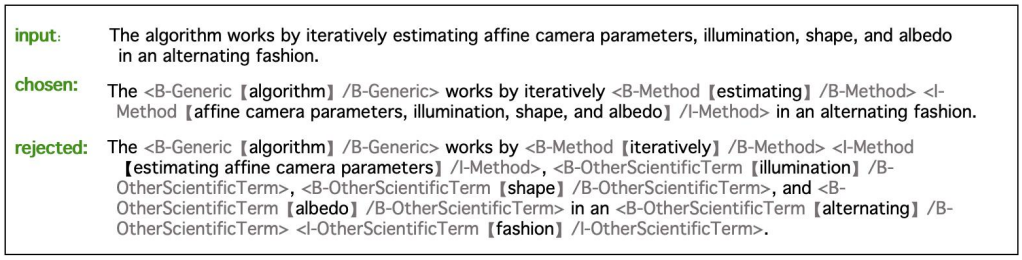


图 3.5 “多标”类负样本以及类别错误负样本(C 类样本)

3.4 实验分析

3.4.1 数据集构建和实验设置

为了验证方法在不同科学文献领域的适用性，使用了以下两个不同领域的公共数据集进行实验：

SCIERC 数据集[77]：该数据集是由 Yi Luan 等人创建的，该数据集由 500 篇计算机领域科学文献的摘要构成，涉及来自 12 个 AI 会议和研讨会的论文，具体包括人工智能、自然语言处理、语音识别、机器学习等多个领域的会议。这些文献摘自 Semantic Scholar 数据库。SCIERC 数据集的设计延续了 SemEval 2017 任务 10[78]和 SemEval 2018 任务 7[79]的数据集，并在此基础上进行了扩展，数据集已被预先划分为训练集（1857 个样本），验证集（275 个样本）和测试集（551 个样本）。该数据集涵盖了如表 3.1 所示的六种实体类型以及每种实体对应的含义：

表 3.1 SCIERC 的实体类别及其含义

实体类别	实体含义
Task（任务）	描述研究的任务或问题。
Method（方法）	描述论文中提出的技术或方法。
Metric（度量）	与研究中所采用的评估指标相关的实体。
Material（材料）	涉及实验中使用的材料或对象。
Other-ScientificTerm（其他科学术语）	不属于上述类别，但仍对领域相关研究具有重要意义的术语。
Generic（通用术语）	不特定于任何领域的通用术语。

BC5CDR 数据集[80]：该数据集是由 Jiao Li 等人创建的，该数据集由 1500 篇 PubMed 文章的摘要构成，涉及化学物质与疾病之间的关系。数据集的设计旨在支持 BioCreative V 挑战任务，提供关于疾病和化学物质的命名实体识别。BC5CDR 数据集已被预先划分为训练集（4611 个样本）、验证集（4606 个样本）和测试集（4818 个样本）。该数据集涵盖了如表 3.2 所示的两种科学实体类型：

表 3.2 BC5CDR 的实体类别及其含义

实体类别	实体含义
Chemical（化学物质）	描述文章中提到的化学物质
Disease（疾病）	描述文章中提到的疾病

本实验在监督微调阶段采用 LoRA 对大语言模型 Llama3.1-8B-chat 进行优化。为兼顾训练效率与模型性能，实验开启了 bf16 数值精度，同时将输入文本统一截断至 1024 个 token，批处理大小设置为 2，梯度累积步数为 1，以确保参数更新的稳定性。优化器采用 PyTorch 实现的 AdamW，初始学习率为 5.0×10^{-5} ，并利用余弦退火调度器[85]对学习率进行动态调节，训练过程中设定最大梯度范数为 1.0 以防止梯度爆炸。针对 LoRA 模块，本实验配置了 lora_alpha 为 32、lora_rank 为 8、lora_dropout 为 0.1，并指定对所有目标层进行更新，同时将 LoRA 模块的学习率放大 16 倍，监督阶段训练 20 轮。

在直接偏好优化阶段，训练过程中采用 bf16 数值精度，并将输入文本统一截断至 1024 个 token，从而在保证上下文完整性的前提下提高训练效率。同

样使用 LoRA 微调方法并自动启用 Flash Attention，以加速注意力计算；每个设备的批量大小设为 4，梯度累积步数为 1。采用 AdamW 优化器，初始学习率为 2.0×10^{-6} ，并使用余弦退火调度器对学习率进行动态调节，同时最大梯度范数设置为 1.0 以防止梯度爆炸。LoRA 相关参数配置为：lora_alpha 为 32、lora_rank 为 8、lora_dropout 设为 0（不使用 dropout）、lora_target 指定为所有层，且 loraplus_lr_ratio 设为 16。训练周期为 2 轮。DPO 阶段偏好优化的参数设置中，pref_beta 设置为 0.05、pref_ftx 为 0.5。

本实验采用 Intel Xeon Gold 6248R 处理器与 NVIDIA A100-PCIE-40GB GPU 搭建硬件平台，软件环境配置为 Python 3.10.8 和 PyTorch 2.1.2。

3.4.3 标注方法在监督微调阶段的影响

对于词级精确理解的命名实体识别任务，大语言模型需要充分利用上下文信息的同时并兼顾对实体文本的精确关注，因此可以将问题归纳为两个方面：第一，如何在标注和模型输入过程中尽量保持丰富的上下文；第二，如何平衡模型对实体内容与实体标签的注意力，使其在生成式预测时兼顾整体语义与局部准确度。为此，本章提出了 7 种不同的标注方法来研究在监督阶段不同标注方法对模型表现的影响。

原文：We propose a Coupled Marginalized Denoising Auto - encoders framework to address the cross - domain problem .

标注方法：

I	We propose a <B-Method【Coupled】/B-Method> <I-Method【Marginalized Denoising Auto - encoders framework】/I-Method> to address the <B-Task【cross】/B-Task> <I-Task【- domain problem】/I-Task> .
II	We propose a <B-Method【Coupled】/B-Method> <I-Method【Marginalized 】/I-Method> <I-Method【Denoising】/I-Method> <I-Method【Auto】/I-Method> <I-Method【-】/I-Method> <I-Method【encoders】/I-Method><I-Method【framework】/I-Method> to address the <B-Task【cross】/B-Task> <I-Task【- domain problem】/I-Task><I-Task【- domain problem】/I-Task><I-Task【domain】/I-Task><I-Task【problem】/I-Task> .
III	We propose a <Method:B-Method【Coupled】/B-Method> <Method:I-Method【Marginalized Denoising Auto - encoders framework】/I-Method> to address the <Task:B-Task【cross】/B-Task> <Task:I-Task【- domain problem】/I-Task> .
IV	We propose a <span:B-Method>Coupled<span:/B-Method> <span:I-Method>Marginalized Denoising Auto - encoders framework<span:/I-Method> to address the <span:B-Task>cross<span:/B-Task> <span:I-Task>- domain problem<span:/I-Task> .
V	We propose a <Method>Coupled Marginalized Denoising Auto - encoders framework</Method> to address the <Task>cross- domain problem</Task> .
VI	We propose a <B-Method>Coupled</B-Method> <I-Method>Marginalized Denoising Auto - encoders framework</I-Method> to address the <B-Task>cross</B-Task> <I-Task>- domain problem</I-Task> .
VII	We propose a [:B-Method]Coupled[:/B-Method] [:I-Method]Marginalized Denoising Auto - encoders framework[:/I-Method] to address the [:B-Task]cross[:/B-Task] [:I-Task]- domain problem[:/I-Task].

图 3.6 7 种标注方法的具体示例

图 3.6 展示了 7 种标注方法的具体示例，分别是：

- 方法 I：使用最基础的 B- / I-标注，将实体开头和内部区分为两个标签，但对实体内部不做进一步精细切分，利用【】引起模型对实体本身的注意。
- 方法 II：在方法 I 的基础上，对实体进行更细粒度的拆分，每个词乃至符号都使用 B- / I-标注。
- 方法 III：在方法 I 的基础上，对 B- / I-前添加实体类型作为前缀，提供双层类别信息。
- 方法 IV：以<span:B-Method>这种类似 HTML/XML 的方式标注，保留 B- / I-结构。
- 方法 V：仅用类别标签<type>包裹整个实体，不再区分开头或内部。
- 方法 VI：采用<label></label>的方式包裹实体，结合 B- / I-形式，但对实体内部不做进一步精细切分。
- 方法 VII：用[:label][:/label]的方式包裹实体，结合 B- / I-形式，但对实体内部不做进一步精细切分。

为了评估这 7 种方法在监督训练阶段的表现，分别用这 7 种方法在 SCIERC 数据集进行实验。

表 3.3 在 SCIERC 7 种标注方法的实验表现

方法	Blue-4	ROUGE-1	ROUGE-2	ROUGE-L
I	**84.14**	95.25	89.98	**90.06**
II	80.86	98.08	**91.09**	88.46
III	81.87	98.11	89.22	89.37
IV	82.40	98.04	89.90	88.65
V	82.37	**98.16**	89.03	88.57
VI	82.24	98.01	89.85	88.76
VII	84.12	98.15	90.50	90.05

如表 3.3 所示，方法 I 与方法 VII 在 BLEU-4[81]指标上均表现较为突出，说明这两种方法在捕捉局部精细信息方面表现较好。对于 ROUGE 系列指标[82]，ROUGE-1 主要关注单词级别的重叠情况，表明生成文本能够覆盖参考文本的大部分单词，结果显示方法 V 和方法 VII 在这一指标上表现较好；而 ROUGE-2 则关注连续两词的匹配，方法 II 获得最高分，说明其在捕捉词组搭配上的表现

较好；ROUGE-L 通过最长公共子序列来反映文本整体结构的一致性，方法 I 和 VII 分别取得 90.06 和 90.05 的高分，表明这两种方法在保持文本整体连贯性上同样具有优势。总体上，各种标注方法在评价指标上呈现出不同的侧重，方法 II 更善于捕捉 n-gram 级别的局部细节，而方法 I 与 VII 则兼顾了文本局部精度与整体结构的平衡。为了进一步探究不同标注方法在命名实体识别任务上的具体表现，采用精确率，召回率和 F1 作为评价指标[83]，在 SCIERC 数据集上的表现如表 3.4：

表 3.4 在 SCIERC 的实体识别实验表现

方法	精确率	召回率	F1
I	**70.6**	**71.3**	**70.9**
II	68.6	64.5	66.5
III	69.4	66.8	68.1
IV	68.3	64.7	66.4
V	68.9	65.4	67.1
VI	70.0	65.6	67.7
VII	69.0	70.2	69.6

从结果可以看到，方法 I 拥有最好的表现。该方法仅对实体开头与内部进行 B- / I-区分，并利用方括号对实体进行简单突出，既避免了过细粒度拆分导致的标注噪音，又能帮助模型明确识别实体边界。相较之下，方法 II 虽然将实体拆分到词乃至符号级别，使模型可在局部位置精确对齐，但在召回率和 F1 方面表现相对欠佳，表明过细粒度的标注在一定程度上影响了整体识别的连贯性。方法 III 在 B- / I-之前添加实体类型前缀，进一步强化了标签信息，其 F1 值比方法 II 来说有提升，但依然没有方法 I 表现优秀，说明过度强化实体标签反而可能会提高模型的生成的难度，并可能抢夺模型对本身实体语义的关注。方法 IV 则采用了类似 HTML/XML 的标注形式，保留 B- / I-结构以维持通用性，但其 F1 与方法 II 相近，说明并未显著提高识别效果，且引入了除了实体类别以外的文本‘span’可能会对模型造成干扰。方法 V 仅用大类别标签整体包裹实体，无开头与内部之分，F1 居中，说明 BIO 的标注思想不仅能在一定程度上强化实体对标签的注意力，也对实体的边界有一定的强调作用。

3.4.4 预测错误结果分析

监督微调阶段使用表现最好的方法 I 在 SCIERC 数据集和 BC5CDR 数据集进行实验，然后对测试集中预测的错误结果进行分析，结果如图 3.7 和图 3.8 所示：

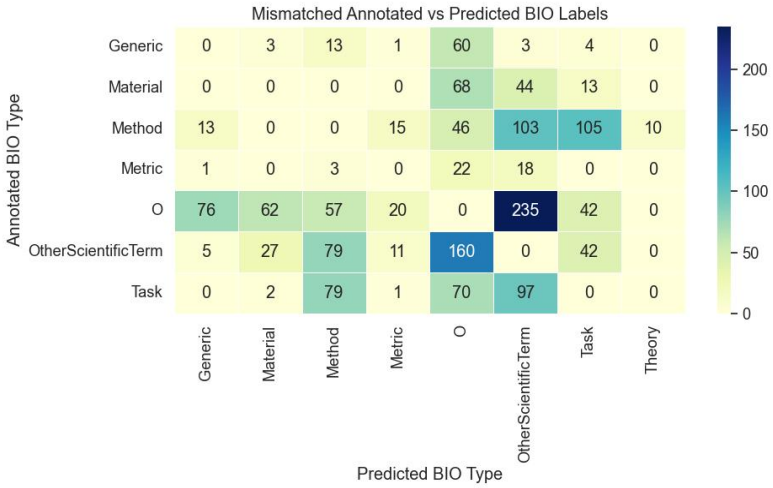


图 3.7 SCIERC 数据集的预测错误结果

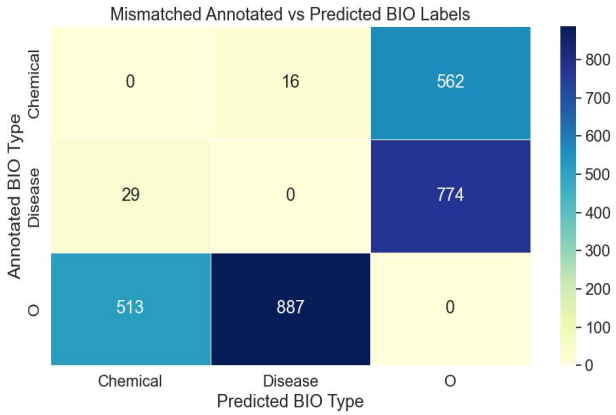


图 3.8 BC5CDR 数据集的预测错误结果

从图 3.7 和 3.8 可以看出，最常出现的错误现象是将标记为“O”的非实体与实际的实体标签混淆。这一现象的主要有两个原因：一是实体的边界不清；二是训练数据中非实体部分占比较高、导致实体与非实体之间的识别错误，产生“漏标”或“错标”。而当在实体类别数量扩充时，不同实体之间的混淆现象会变

得更加明显。在只有少数实体类别时，比如仅包含 Chemical 和 Disease 时，错误主要体现在非实体和实体之间的区分上；而当类别扩展到包括 Method、Metric、Material、Task、Theory 和 OtherScientificTerm 等更为细化的类型时，由于各实体之间在语义上存在相似性以及定义边界存在部分重叠，不同实体之间出现误分类的情况将显著增加。为了进一步分析 SCIERC 数据集中各实体类别之间的联系，本实验使用 SciDeBERTa_v2[86]模型对验证集中的所有非”O”的实体进行编码，取出[CLS]的向量作为整个实体的表示，然后利用 TSNE 将高维向量降至二维后的分布情况可视化结果如图。

从初步结果来看，不同实体类别在向量空间中呈现出一定程度的聚集效应，但部分类别之间仍存在明显重叠。为进一步探究这一现象，采用 KMeans [87]算法对所有实体嵌入进行聚类，聚类数设定为真实实体类别数量。对每个聚类，通过多数投票方法确定代表性标签，并将聚类结果映射到相应的实体标签上进行比较。

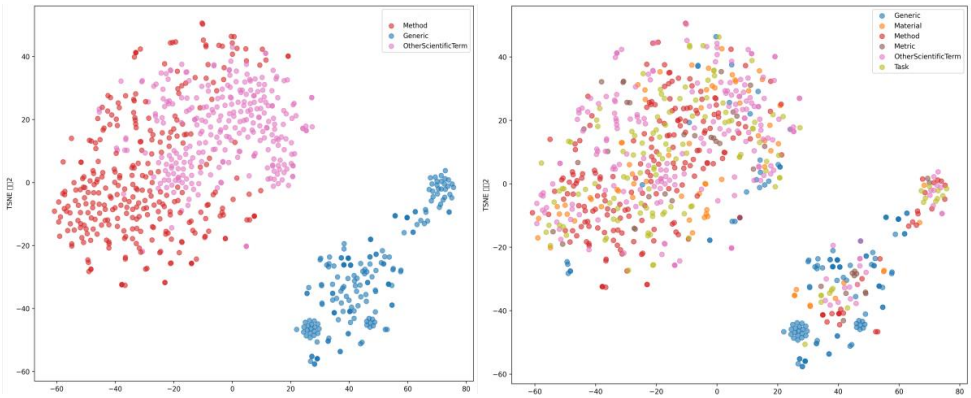


图 3.9 SCIERC 的验证集实体向量可视化

结果如图 3.9 显示，一些原本定义明确的细粒度标签在聚类后发生了合并现象，例如 Material、Metric 和 Task 等类别的实例被归入 Method、Generic 或 OtherScientificTerm 的簇中，而 OtherScientificTerm 类别则显示出较为松散的分布。在 KMeans 投票中大量合并了其他类别的实体，说明其定义在语义上本就宽泛且分布松散，与其他标签区分度不足；这也图与模型对 SCIERC 数据集的

错误结果对应，数据集中语义较为宽泛与其他标签区分度不足的类别也会明显影响模型的表现。

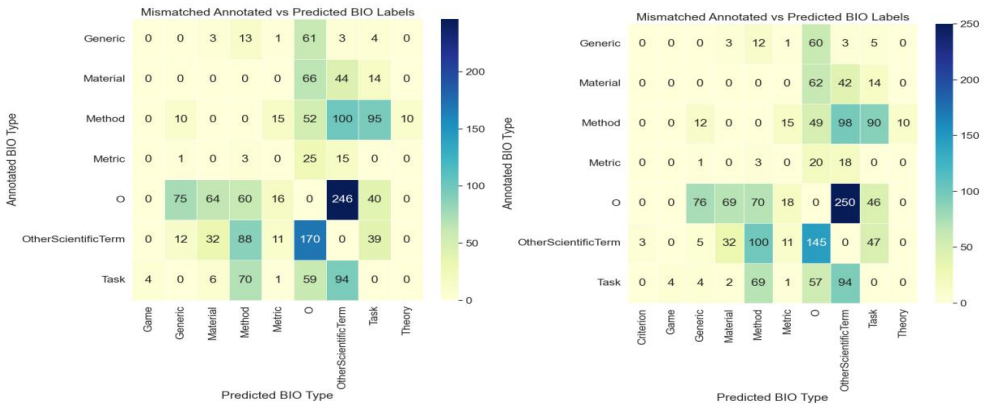
3.4.5 负样本的不同构造方法在直接偏优化阶段的影响

本实验采样时，对于扩张类负样本、收缩类负样本以及“漏标”类负样本，对于有效标签数 E 设定阈值 $4 \le E \le 8$，对于有效标签比例 R 设定阈值 $4 \le R \le 8$，确保正负样本之间拥有充分差异进行对比学习的同时，能有拥有足够的空间进行边界扩张和收缩。

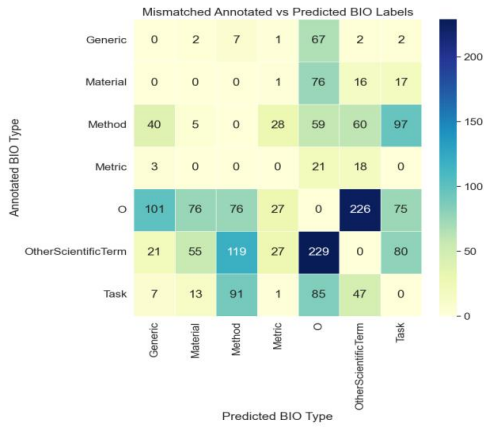
对于在验证集上经过监督训练的模型推理产生的真实错误样本，可能存在多种错误并存的情况，因此在消融实验单独测试 C 类样本时，使用全部的真实错误实例对模型进行训练，在与前两类样本合并时，按照前两种方法构造的负样本总量的 1/4，对真实错误样本进行随机抽样加入。

表 3.5 在 SCIERC 的消融实验结果

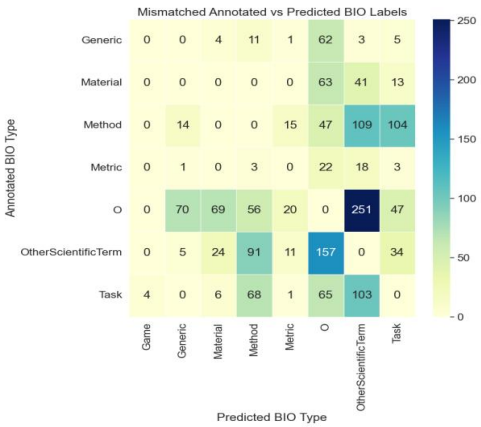
训练阶段	负样本类别			精确率	召回率	F1
	A 类样本	B 类样本	C 类样本			
监督训练				70.4	71.5	70.9
DPO 训练	√			70.6	71.2	70.9
		√		70.2	71.6	70.9
			√	66.8	64.6	65.7
	√	√		70.6	71.6	71.1
	√	√	√	**71.2**	**71.6**	**71.4**



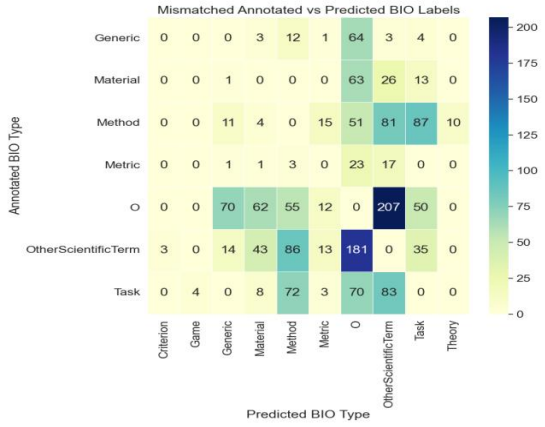
(a) A 类样本　　(b) B 类样本



(c) C 类全部样本



(d) A 类+B 类样本



(e) A 类+B 类+C 类样本

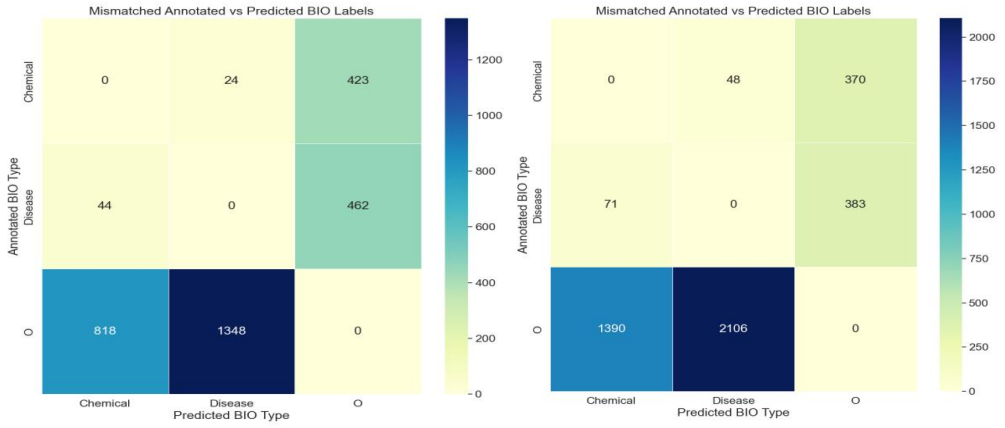
图 3.10 SCIERC 数据集的每类实体预测错误结果

从表 3.5 结果可见，在完成监督训练后，进一步对直接偏好优化阶段结合不同类型负样本，对模型性能会产生明显影响：单独使用 A 类（扩张边界）或 B 类（收缩边界）对整体性能并没有明显提升；仅使用 C 类（多标及类别错误）负样本则导致 F1 大幅下降，主要原因在于真实推理中的错误往往包含多重边界与类别混淆，直接纳入这些复杂负样本会干扰模型原有识别能力。而将 A 与 B 两类模拟性错误相结合后，F1 提升，说明边界扩张与收缩具有一定互补性，能够更有效地校正模型在实体边界上的判断。当进一步融合 C 类（真实推理错误），即三类类样本联合训练时，模型在精确率与召回率上均稳步提升，F1 最

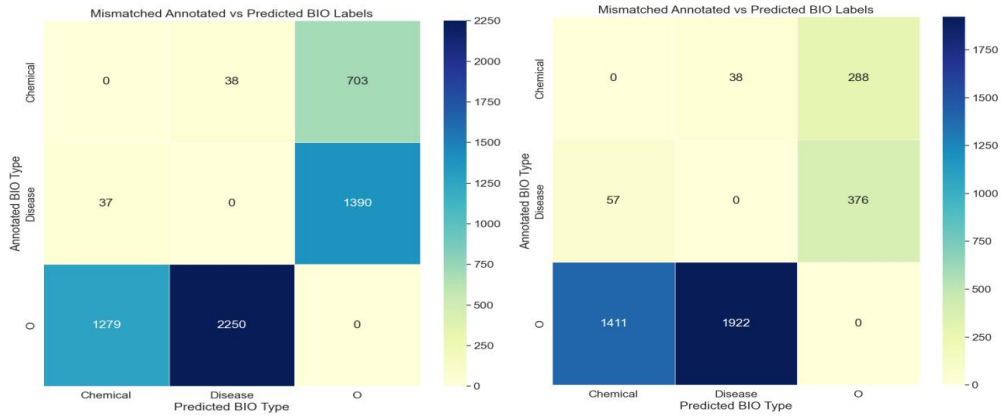
终达到 71.4。这表明“模拟错误”（A、B 类）与“真实错误”（C 类）结合能覆盖更多失误场景的同时有效帮助模型在 DPO 阶段更全面地修正错判，最大化其泛化和纠错能力。对各类别的预测错误情况进行分析从图 3.10 表明，不同负样本类型会影响模型的错误分布，尤其在“OtherScientificTerm”与“O”之间常出现互相误判，并且会出现少量“幻觉”现象，即模型会生成不存在的实体类别。C 类负样本源自真实推理过程，错误场景更复杂，若单独使用容易导致模型短期内在多类别上均出现较大偏差。相比之下，A + B 的联合负样本可在一定程度上弥补彼此不足，使模型对过度或不足标记的边界有更好的识别，而当再加入 C 类真实推理错误后，能够覆盖从实体边界到具体类别的多种误判类型，使得模型在纠错过程中获得更全面的反馈。因此，A + B + C 多样化负样本的结合让预测错误情况中关键类别之间的错误大幅减少，在实体边界与类别区分上均得到一定提升，进而在精确率、召回率和 F1 值上得到了最佳效果。

表 3.6 在 BC5CDR 的消融实验结果

训练阶段	负样本类别			精确率	召回率	F1
	A 类样本	B 类样本	C 类样本			
监督训练				88.7	**89.5**	**89.1**
DPO 训练	√			85.4	91.4	88.3
		√		81.2	89.9	85.4
			√	76.0	81.6	78.7
	√	√		81.1	91.6	86.1
	√	√	√	**88.9**	89.0	88.9

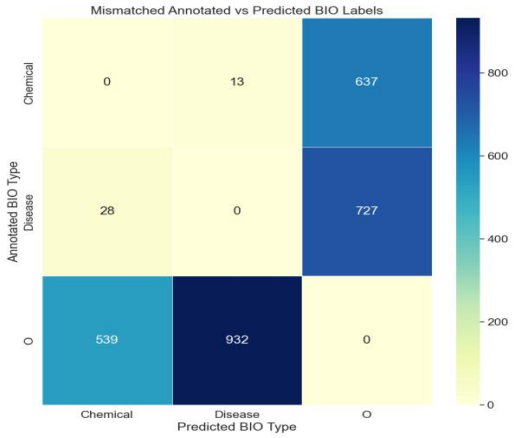


(a) A 类　　(b) B 类



(c) C 类全部样本

(d) A 类+B 类样本



(e) A 类+B 类+C 类样本

图 3.11 BC5CDR 数据集的每类实体预测错误结果

从表 3.6 结果来看，BC5CDR 数据集在经过监督训练后，取得了较高的基线表现。然而，当引入直接偏好优化训练并分别或组合地使用三类负样本时，除了 A + B + C 的组合能带来精确率的小幅提升, 其余情形下的召回率和 F1 均出现了不同程度的下降，说明针对生物医学场景中大量专有名词、符号与缩写而构造的负样本，未能在整体上显著提高命名实体识别的性能。特别是单独使用 B 类或 C 类负样本时，模型的精确率与召回率均大幅下滑，导致 F1 值显著下降；即便 A 类与 B 类的联合在一定程度上缓解了这种退化，整体 F1 仍难以超过基线。结合类别的预测错误情况如图 3.11 进一步发现其中的原因。该数据

集仅包含 “Chemical”、“Disease”和“O”三个 BIO 标签，因此当模型识别出现偏差时，往往会在“Disease”与“O”之间，或“Chemical”与“O”之间发生误判。当联合使用 A、B、C 三类负样本时（图 e），虽然整体“Disease”与“O”之间的混淆明显减少，模型对实体边界和类别的区分也得到一定提升，因而精确率随之上升，但在召回率方面仍不及监督基线，*F*1 也略有下降。这说明针对生物信息领域复杂的术语变体和专业缩写，仅靠扩张、收缩等模拟的边界错误与有限的真实错误样本，难以全面修正模型对高专业度命名实体的识别偏差。

3.4.6 对比实验

为了进一步证明方法的有效性，本章提供了与其他方法的对比实验结果，对比模型包括 PL-Marker[88]、ATG[89]、P-MHE[90]、SpERT[91]、UniRE[92]、TriMF[93]、PGA-SciRE[53]、Qi Li 等[54]、WordNet[94]、EDA[95]、Naïve DA[96]、GDA[52]、SciDeBERTa[86]和 SciBERT[39]。这些模型分为联合抽取、数据增强、知识注入和领域语言模型 4 类。

其中，联合抽取模型一般是通过单一框架内同步识别实体与判定关系，以提升任务间信息交互。PL-Marker 利用双标记打包建模实体边界与依赖；ATG 将任务重塑为生成式图序列输出；P-MHE 针对科学文献长句和多实体场景构建异构图并借助消息传递优化实体表征；SpERT 通过枚举并分类候选跨度进行实体判定；UniRE 将实体识别映射为表格对角线单元格分类，在统一标签空间内完成实体与关系预测；而 TriMF 通过三元组矩阵分解与记忆流迭代，把实体映射到共享记忆矩阵中进行特征更新。

对于数据增强的方法来说，传统方法如 WordNet 同义词替换与 EDA 通过同义词替换、随机插入/交换/删除等浅层扰动生成伪样本。近年来，大语言模型驱动的语义级增强受到关注：Naïve DA 直接向大语言模型输入“待替换实体+类型”提示，生成仅替换实体的平行句；PGA - SciRE 进一步让大语言模型对带实体边界的句子进行释义，得到语义等价而表面形式多样的样本；GDA 则先抽象出“上下文骨架+实体角色占位符”，再将该中间表示拼入提示，引导同一大语言模型产生结构多样、语义一致的句子。上述方法从规则级到语义级形

成了渐进式增强体系。Qi Li 等人提出的框架则是通过实体链接将外部知识对齐到文本中，并结合最优伪标签筛选策略，利用大语言模型生成的上下文语义进行自训练，增强 跨度抽取模型的实体表征能力。这一思路强调在数据增广之外，引入知识与噪声控制并行提升模型质量。

领域语言模型 SciBERT 和 SciDeBERTa 的提出，是为了减小科学文本与通用语料在词汇与句法上的鸿沟，使其在科学领域的词汇、实体边界及长依赖表示方面更加贴合下游任务。

表 3.7 在 SCIERC 的对比实验结果

模型	精确率	召回率	F1
PL-Marker	-	-	68.8
ATG	-	-	69.7
P-MHE	68.4	68.2	68.3
SpERT	68.5	70.1	69.3
UniRE	65.8	71.1	68.4
TriMF	70.2	70.2	70.2
PGA-SciRE	70.1	70.9	70.5
Qi Li 等	-	-	70.3
WordNet	-	-	50.2
EDA	-	-	54.3
Naïve DA	-	-	53.4
GDA	-	-	54.6
SciDeBERTa	-	-	71.1
SciBERT	-	-	67.6
Ours	**71.2**	**71.6**	**71.4**

表 3.7 展示了在 SCIERC 数据集上的系统对比结果，可以看到不同类型方法对应的表现差异较明显。对于数据增强类方法，传统规则增强 WordNet 同义词替换与 EDA 随机插入/交换/删除等操作仅在词面层对文本做浅层改写。它们几乎不触及实体的边界或类型信息，修改也缺乏全句一致性检查，因此对科学术语抽取帮助有限；用大语言模型驱动进行数据增强的 Naïve DA 与 GDA 都调用大语言模型生成伪样本，Naïve DA 只要求大语言模型把指定实体换成同类型新词，句子其余部分保持原样，GDA 先将原句抽象成“上下文骨架 + 实体角色占位符”，再把该抽象连同指导提示输入大语言模型，生成整体语法连贯且保留

实体语义的新句子。相较于传统规则，大语言模型增强在语法多样性和上下文流畅度上明显优越，但可能缺乏足够的格式约束，如 Naïve DA，仍易引入边界错位或类型漂移。GDA 通过骨架约束大幅降低了这类噪声，效果优于 Naïve DA，但可能仍受限于对科学长复句的覆盖深度。当对大型语言模型的生成过程加入更严格的语义或结构约束时，性能有了明显跃升。PGA - SciRE 让大语言模型在保持原句语义的前提下重写句子，同时保留显式的实体边界标记，表现明显提升；Qi Li 等方法又在此基础上引入实体链接和置信度过滤，用大语言模型生成上下文与伪标签进行自训练，表现明显优于单纯数据增强的方法。

对于联合抽取框架，PL - Marker、ATG 等方法借助显式标记或生成式建模把实体边界与关系信息放入同一流程，整体优于单纯数据增强方法；SpERT、UniRE、TriMF、P - MHE 等在边界建模和信息交互上进一步细化，表现比较稳定。对于领域语言模型来说，SciBERT 依靠科学文献语料训练中得到的领域知识，表现比单纯数据增强方法表现更好；SciDeBERTa 强化了词表和语料规模，使表征更贴合科学文献领域，表现更加突出。

本章提出的上下文一致的显式标注方法，在此基础上，将大语言模型作为核心识别模型并通过双阶段训练让模型进行充分学习。一方面，成对特殊标记同时注入了实体类别与边界信息，避免了如 UniRE 那样的对角线标签弱化边界问题。另一方面，两阶段训练策略使模型先在监督微调阶段学习任务的基本形式，再通过直接偏好优化阶段结合扩张 / 收缩边界与类别混淆等难度较大的负样本进行细粒度偏好对比，引导模型学习偏好。可以看到，本章的方法有助于将大语言模型的生成与推理能力适用科学文献命名实体识别的任务形式，且优于其他大语言模型的相关方法。

3.4.7 BC5CDR 数据集的预测错误案例分析

在 BC5CDR 数据集的预测结果中，选取四个具有代表性的错误案例对模型在生物医学场景中的局限性进行了剖析。

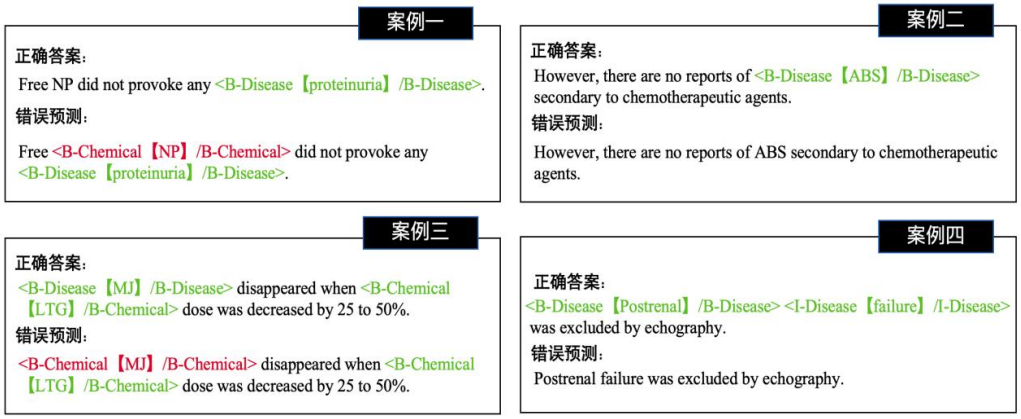


图 3.12 BC5CDR 数据集的预测错误案例

如图 3.12 所示, 在第一个案例中, NP 是内源性多肽激素 (natriuretic peptide) 的缩写，并不属于 BC5CDR “Chemical” 类别所定义的小分子化合物，模型却将其误标为化学实体；案例二中被漏标的疾病实体“ABS”为血液病 Aplastic Bone-Marrow Syndrome 的通用缩写。案例三中的 MJ 为肌阵挛 (myoclonic jerks) 的缩写，应标注为疾病实体，但仍被错判为化学物质，说明模型无法正确理解专业缩写的含义而产生“错标”，“多标”和“漏标”的错误预测现象；案例四中，模型未能识别疾病实体“Postrenal failure（后肾性肾衰竭）”，表明其对专业术语及其领域知识的欠缺。这四例集中体现了生物医学文本中专有名词、符号与缩写的高度专业化特点；在该情形下，仅依靠类别混淆或边界扰动生成的负样本，难以显著提升大语言模型对复杂术语变体和专业缩写的命名实体识别性能，模型对高专业度实体的识别偏差仍难以彻底纠正。

3.5 本章小结

本章提出了一种上下文一致的实体显式标注方法，通过在句子中为实体嵌入成对、带类型的特殊标记，为大语言模型提供清晰的边界提示同时保留原义。该方法结合了“监督微调+直接偏好优化”的双阶段训练策略：监督微调阶段学习基础实体知识，直接偏好优化阶段则引入针对边界扩张、收缩及类别混淆等典型错误的负样本，与正样本配对进行偏好学习，以强化模型对边界与类别

的精细判别力。在计算机科学领域的数据集 SCIERC 上的结果表明，该方法能有效提升大语言模型在科学文献场景下命名实体识别任务的表现，并取得了当前最佳性能。然而，在术语高度集中且缩写频繁的生物医学数据集 BC5CDR 上，由于负样本设计未能充分契合领域特性，性能提升有限，这表明了在高度专业化领域，若缺乏相应的领域知识或针对性的负样本策略，大语言模型的识别精度可能受限，凸显了领域适应的重要性。

第四章　基于领域语言模型的科学文献命名实体识别与应用

上一章节的实验表明，基于大语言模型（LLM）的命名实体识别方法在计算机科学领域 SCIERC 数据集上取得了较好的表现。 SCIERC 数据集的实体如任务、方法、评价指标等相对更好理解，通用预训练模型能够充分学习这些概念的表示。相较之下，材料科学和生物医学领域含有大量低频、缩写复杂的专业术语，通用语料覆盖度有限，大语言模型难以直接捕获其细粒度语义；当领域专业化程度升高，仅依赖通用预训练已难以保证识别精度。为此，本章提出了基于领域语言模型的语义融合方法，通过融合特定领域拥有不同的领域语言模型及其对应的领域词级向量，实现对领域文本中深层语义信息的精准捕捉和表达。目前生物医学已拥有多个公共的命名实体识别任务基准数据集；相比之下，材料科学的数据集规模较小且涉及领域知识较为单一，因此本章构建了一个统一标注规范的材料领域专用数据集作为补充, 并在具体领域进行实际应用，展现了该方法的应用价值。

4.1　基于领域语言模型的科学文献命名实体识别

4.1.1 方法概述

整体结构如图 4.1。首先，输入的原始文本序列会经过分词器处理，将其分解为模型可处理的词元序列。接下来，这个词元序列会被分别输入到两个不同的预训练领域语言模型 LM1 和 LM2 中，以获取目标领域中复杂的上下文依赖和细致语义特征。与此同时，为了增强模型在词级别的理解能力，利用领域词向量模型为输入文本生成对应词级表示，来补充通用语言模型在处理专业术语和低频词汇时存在的不足。最后，将来自两个预训练领域语言模型的上下文表示与来自词向量模型的词级表示进行语义融合, 并传递给 BiLSTM 层。BiLSTM 网络通过同时从前向后和从后向前两个方向处理整个序列，能够有效捕捉词元

之间长距离的依赖关系，并生成充分考虑了前后文信息的隐藏状态。这些隐藏状态随后通过一个全连接层进行转换，为每个词元计算出其对应于各个预定义命名实体标签的发射得分。这个分数反映了在当前位置，模型认为该词元属于某个特定标签的初步判断。然而，直接基于最高分选择标签可能会产生不符合语法或标注规则的标签序列，例如，实体内部标签“I-X”出现在实体开始标签“B-X”之前。因此，使用 CRF 进行解码。CRF 不仅考虑每个位置的发射得分，还会学习并利用标签之间合理的转移模式。通过综合评估所有可能的标签路径，并结合发射得分与学到的转移偏好，CRF 能够确定并输出整体上最可能、最连贯的命名实体标签序列。

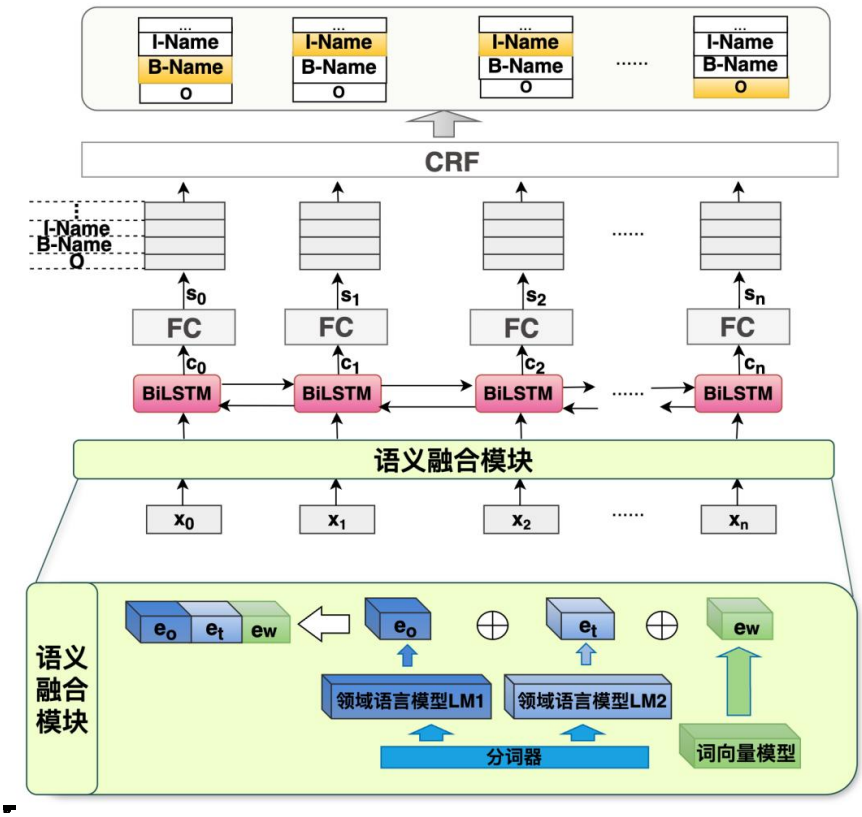


图 4.1 语义融合模块结合 BiLSTM-CRF 框架

4.1.2 语义融合模块

语义融合模块分别获取输入文本的上下文和词级别表示。对于输入的文本序列x，分词器将每个单词拆分成固定长度的词元，形成词元序列，随后由领域语言模型LM_1和领域语言模型LM_2进行嵌入表示：

$$e_o = E_o(spl(x)), \tag{4.1}$$

$$e_t = E_t(spl(x)), \tag{4.2}$$

$spl(x)$表示由分词器分词后的词元序列，E_o和E_t分别表示LM_1和LM_2的嵌入功能，分别用于获取词元序列$spl(x)$的向量表示e_o和e_t。为了增强模型在词级别的理解，通过领域词向量模型W获取词向量表示，记作e_w，具体公式如下：

$$e_w = E_w(x), \tag{4.3}$$

其中，E_w表示领域词向量模型W的嵌入功能。最后，将获得的向量表示e_o，e_t和e_w拼接在一起，形成语义融合向量e_{otw}，即：

$$e_{otw} = e_o \oplus e_t \oplus e_w, \tag{4.4}$$

其中，$\oplus$表示向量表示的拼接。通过语义融合的向量表示e_{otw}可以为输入文本x提供深层语义信息的精准捕捉和表达。

4.1.3 BiLSTM-CRF 框架

BiLSTM 模块接收由语义融合后的向量表示，并对每个向量进行双向编码，以捕捉更丰富的上下文信息。通过双向扫描输入序列，并将其隐藏状态$(h_1, h_2, \ldots, h_n)$进行拼接，公式如下：

$$(h_1, h_2, \ldots, h_n) = \mathrm{LSTM}(e^1_{otw}, e^2_{otw}, \ldots, e^n_{otw}), \tag{4.5}$$

其中，e_{otw}^{i}为第i个输入向量，通过语义融合模块获取。BiLSTM 由一个前向 LSTM 和一个后向 LSTM 组成，分别记作 $\mathrm{LSTM_s}$ 和 $\mathrm{LSTM_b}$，表示为：

$$h_i = \mathrm{LSTM_s}(h_i, h_{i-1}), \tag{4.6}$$

$$\tilde{h}_i = \mathrm{LSTM_b}(h_i, h_{i+1}), \tag{4.7}$$

h_i表示通过前向 LSTM 获取的隐藏状态，它基于当前输入h_i和前一个时间步h_{i-1}的隐藏状态；$\tilde{h}_i$表示通过后向 LSTM 获取的隐藏状态，它基于当前输入h_i和下一个时间步h_{i+1}的隐藏状态。最终，结合h_i和$\tilde{h}_i$生成最终上下文向量c_i，公式如下：

$$c_i = h_i \oplus \tilde{h}_i, \tag{4.8}$$

其中，符号$\oplus$表示拼接操作。接下来，使用线性变换（FC）将 BiLSTM 输出c_i换为每个位置的发射得分s_i，即：

$$s_i = W \cdot c_i + b, \tag{4.9}$$

其中，W 是权重矩阵，b 是偏置向量。通过 BIO 标签标注的方法，$\mathrm{s_i}$表示对标签的预测得分。BiLSTM 通过双向上下文建模，增强了对序列依赖关系的捕捉，为后续 CRF 层提供了更准确的顺序特征。标签序列记为：

$$y = \{y_1, y_2, \ldots, y_i, \ldots, y_n\}, \tag{4.10}$$

y_i表示是位置i对应的标签；$y_i \in \mathcal{C}$，$\mathcal{C}$为预先定义的实体类别标签。考虑到 BIO 标签的序列性质，CRF 接收从s_i构建的发射得分矩阵，并计算转换得分矩阵A，表示从一个标签转换到下一个标签的得分。公式如下：

$$\mathrm{Score}\,(x, y) = \sum_{i=1}^{n} \left(A_{y_{i-1}, y_i} + s_{i, y_i} \right), \tag{4.11}$$

其中 A_{y_{i-1}, y_i}表示从标签y_{i-1}到y_i的转移得分，用于捕捉标签之间的相邻依赖；s_{i, y_i}是位置 i 将对应的标签y_i对应的发射得分。CRF 层通过最大化条件概率来

学习转换得分矩阵，公式如下：

$$P(y|x) = \frac{\exp(\text{Score}(x, y))}{\sum_{\widehat{y} \in Y(x)} \exp(\text{Score}(x, \widehat{y}))}, \quad (4.12)$$

$Y(x)$表示所有可能的标签序列，$\exp(\text{Score}(x, y))$通过指数函数将得分函数转换为非负值，表示特定标签序列的重要性。$\sum_{\widehat{y} \in Y(x)} \exp(\text{Score}(x, \widehat{y}))$表示所有可能标签序列的指数得分的总和，用于对每个标签序列的得分进行归一化，从而确保条件概率$P(y|x)$落在 0 和 1 之间。因此，对于单个样本，负对数似然损失 NLLLoss[97]为：

$$NLLLoss = -\log P(y|x), \quad (4.13)$$

$\mathrm{P}(y|x)$表示在给定输入序列x时，生成标签序列y的条件概率。通过最大化正确标签序列的条件概率，模型能够学习到输入与标签之间的最佳映射关系，从而提高预测准确性。

4.2 实验分析

4.2.1 数据集构建和实验设置

为了验证方法在不同领域的适用性，加入了以下两个不同材料子领域的数据集进行补充：

PolymerAbstracts 数据集[44]：该数据集是由 Pranav Shetty 等人创建的，旨在通过自然语言处理技术自动提取聚合物文献中的材料属性数据。该数据集由 750 篇与聚合物相关的文献摘要构成。数据集已被划分为训练集（85%）、验证集（5%）和测试集（10%）。该数据集的标注方法涵盖了 8 种科学实体类型：

表 4.1 PolymerAbstracts 的实体类别及其含义

实体类别	实体含义
POLYMER（聚合物）	描述聚合物材料实体。
POLYMER_CLASS（聚合物类别）	指代某一类聚合物的实体。
PROPERTY_VALUE（属性值）	对应材料属性的数值及其单位。
PROPERTY_NAME（属性名称）	描述材料的特性或属性。
MONOMER（单体）	描述聚合物单体的实体。
ORGANIC_MATERIAL（有机材料）	指有机材料，通常用于作为塑化剂或交联剂。
INORGANIC_MATERIAL（无机材料）	指无机材料，通常作为添加剂使用。
MATERIAL_AMOUNT（材料量）	指特定材料在材料配方中的含量。

PolymerAbstracts 数据集是主要面向高分子材料的命名实体识别任务，本体设计与语料选择主要侧重于聚合物及其相关属性。这使其难以直接应用于高熵合金等金属材料系统，其术语体系、实体构成及关键性能指标存在显著差异。高熵合金作为一类重要的金属材料，因其在结构、耐磨及高温等应用中展现的独特性能优势而受到广泛关注，对其力学性能，如硬度、强度及韧性等，进行精准的自动化信息提取具有重要的科研与工程价值。然而，实现针对高熵合金的高效信息提取面临特定挑战。一方面，高熵合金的命名通常涉及四至六种近等原子比元素的复杂组合，缺乏标准分隔符；另一方面，其硬度等力学性能的表征涉及多种测试方法和不同的结果单位如 HV、HB、GPa，显著增加了属性值抽取与规范化的复杂性。

因此，本研究构建一个高熵合金-硬度性能数据集作为材料领域的补充，通过与现有 PolymerAbstracts 数据集的互补性，来支持对跨材料子领域模型泛化能力及语义融合模块有效性的验证。具体来说，该数据集是通过对 Elsevier 数据库中用"HEA+Hardness"（高熵合金与硬度）进行了 2020-2024 年间文献检索，共得到 2698 篇相关文献。为系统收集合金名称、实验方法及对应硬度数据，从 2698 篇文献中随机选取 400 篇摘要进行 BIO 标注。针对超过 200 词的摘要，采用分句处理形成两个独立数据条目。最终构建包含 565 条样本的数据集，按 8:2 比例随机划分为训练集和测试集，然后经由材料专家进行标注，如图 4.2。该数据集的标注方法涵盖了 3 种科学实体类型如表 4.2:

表 4.2 高熵合金-硬度性能数据集的实体类别及其含义

实体类别	实体含义
Name（名称）	描述合金材料实体。
Preparation（工艺制备）	指代制备合金材料的工艺实体。
Hardness（硬度）	对应合金材料的硬度值。

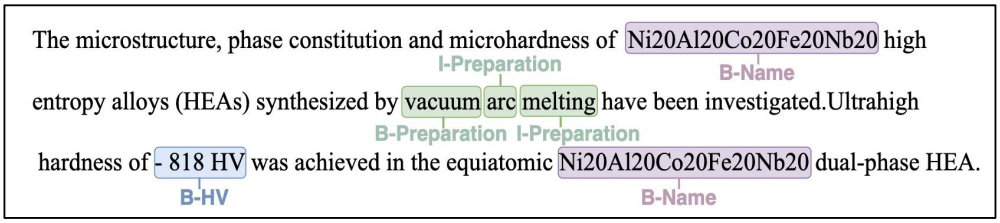


图 4.2 高熵合金-硬度性能数据集的标注示例

本实验采用 Intel Xeon Gold 6248R 处理器与 NVIDIA A100-PCIE-40GB GPU 搭建硬件平台，软件环境配置为 Python 3.10.8 和 PyTorch 2.1.2。编码器基于 BERT 架构，设置隐藏层维度 768，LSTM 层采用 0.4 的 dropout 率。优化器选用 AdamW，初始学习率 0.002，批次大小 32。训练过程持续 200 个 epoch。

4.2.2 消融实验

考虑到不同领域之间对应领域语言模型的差异性，对不同领域采用对应的领域预训练基模型。对生物科学文献领域的数据集 BC5CDR 选用基模型：BioBERT[7]，BioLinkBERT[42]，生物 WordVec[98]；对材料科学文献领域的数据集 PolymerAbstracts 和高熵合金-硬度性能数据集选用基模型：MatSciBERT[6]，MaterialsBERT[44]，材料 fastText[99]。

为了研究语义融合模块在不同领域中的适用性和有效性，本实验在三个数据集上进行了消融实验：生物文献领域的 BC5CDR 数据集，以及材料文献领域的 PolymerAbstracts 数据集和高熵合金-硬度性能数据集。其中，表 4.3 展示了生物文献领域的 BC5CDR 数据集上语义融合模块的消融实验定量结果；表 4.3 和表 4.4 分别对应了材料文献领域中 PolymerAbstracts 与高熵合金-硬度性能数据集的总体表现。

表 4.3 语义融合模块在 BC5CDR 数据集的消融结果

语言模型	精确率			召回率			F1		
	chem	disease	total	chem	disease	total	chem	disease	total
Bio	93.7	86.3	90.7	92.9	84.4	89.4	93.3	85.4	90.1
Pub	93.3	84.8	90.0	92.9	84.4	89.6	93.1	84.6	89.8
BioL	92.8	82.9	88.9	93.7	83.0	89.5	93.2	82.9	89.2
Bio+BioL	**94.4**	85.0	90.5	93.0	87.1	90.6	93.6	86.1	90.5
Bio+WV	94.2	**86.9**	**91.2**	94.7	86.6	91.4	94.4	86.7	91.3
Bio+Pub+BioL	69.7	47.0	58.3	25.2	26.1	25.6	37.0	33.6	35.6
Ours$_{(Bio+BioL+WV)}$	93.7	86.4	90.7	**95.2**	**88.0**	**92.3**	**94.4**	**87.2**	**91.5**

本表格中使用“Bio”表示“BioBERT”，使用“Pub”表示 PubMedBERT[43]，“BioL”表示 BioLinkBERT，“WV”代表生物 WordVec。

表 4.4 语义融合模块在 PolymerAbstracts 与高熵合金-硬度性能数据集的总体结果

语言模型	PolymerAbstracts			高熵合金-硬度性能数据集		
	精确率	召回率	F1	精确率	召回率	F1
MatS	62.71	71.40	66.80	79.25	85.46	82.24
Mate	61.04	68.24	64.42	65.05	69.94	67.41
MatT	61.31	71.84	66.11	81.29	82.24	81.76
MatS+Mate	62.50	70.99	66.48	78.42	81.33	79.85
MatS+FT	63.66	**72.41**	67.75	79.17	**86.62**	82.72
MatS+Mate+MatT	54.24	65.73	59.43	75.65	69.46	72.42
Ours$_{(Mate+MatS+FT)}$	**66.57**	71.21	**68.81**	**87.19**	79.85	**83.36**

本表格中使用“MatS”表示“MatSciBERT”，使用“Mate”表示 MateialsBERT，“BioL”表示 BioLinkBERT，“MatT”表示 MatTPUSciBERT，“WV”代表生物 fastText。

如表 4.3 所示，对于生物文献领域的 BC5CDR 数据集，从整体结果来看，本章提出的语义融合模块在总召回率和总 F1 值上均取得了最高的成绩，显示了其在平衡精确率与召回率方面的优势。具体来说，在化学物质（chem）类别中，双预训练模型的融合方式在化学物质（chem）类实体获得了较高的精确率，而单预训练模型与词向量的融合方式在疾病（disease）两类实体上获得了较高的精确率，但它们的召回率都相对较低，导致其 F1 值没有达到最佳表现。多预训练模型的融合方式尽管融合了多个模型，但由于过度增加了特征维度，可能导致了冗余信息的引入，因此表现明显变差。相比之下，本章提出的语义融合模块在在化学物质（chem）和疾病（disease）两类实体识别上都拥有最高的召回率。表 4.4 所示，针对材料文献领域的 PolymerAbstracts 与高熵合金-硬度

性能数据集，可以看出，语义融合模块在这两个数据集上均表现出较强的性能，尤其在平衡精确率与召回率方面展现了优势。对于这两个数据集而言，模型在多源表征融合后的表现普遍优于单一的预训练模型，在引入词级向量 fastText 后，召回率得到了显著提升，进而提高了 F1 值。从表 4.3 和表 4.4 的结果可以看出，不管是在生物文献领域还是材料文献领域，语义融合模块通过合理融合多源预训练模型和词级向量，都能较好地适应领域科学文献领域的低频词汇和复杂术语的变化，在 F1 值上获得了最好的表现。

表 4.5 语义融合模块在 PolymerAbstracts 上 7 类实体的消融结果

语言模型	指标	PL	PLC	PRV	PRN	MON	ORG	INO	MAT
MatS	精确率	70.7	52.8	80.0	60.1	58.4	34.5	50.0	62.8
	召回率	75.5	53.7	86.7	69.8	82.2	18.9	75.2	77.1
	F1	73.0	53.2	83.2	64.6	68.3	**24.5**	59.9	69.2
Mate	精确率	70.1	59.3	80.0	62.4	62.8	13.2	46.7	60.9
	召回率	71.7	52.0	88.6	71.3	72.8	14.8	75.5	87.7
	F1	70.9	55.4	84.1	66.5	67.4	14.0	57.7	71.9
MatS+Mate	精确率	68.0	67.7	84.8	61.7	55.1	28.3	48.1	67.1
	召回率	80.8	50.2	83.9	70.1	80.1	12.6	56.5	89.4
	F1	73.8	57.7	**84.4**	65.6	65.3	17.5	52.0	**76.7**
MatS+FT	精确率	69.3	54.3	77.4	64.6	62.6	31.9	37.7	56.2
	召回率	74.2	53.7	83.9	71.5	80.4	16.4	65.9	63.1
	F1	71.7	54.0	80.6	**67.9**	70.4	21.7	48.0	59.5
$Ours_{(Mate+MatS+FT)}$	精确率	70.5	60.7	78.8	75.0	61.0	50.0	57.2	53.2
	召回率	78.7	66.3	84.0	53.2	89.7	11.0	63.0	71.9
	F1	**74.3**	**63.4**	81.3	62.3	**72.6**	18.0	**60.0**	61.9

本表格中使用“MatS”表示“MatSciBERT”，使用“Mate”表示 MateialsBERT，“BioL”表示 BioLinkBERT，“WV”代表生物 fastText。黑体为该实体类别在 F1 指标上取得的最高值；波浪线表示该实体类别精确率取得的最高值；下划线表示该实体类别在召回率上取得的最高值。

为了进一步探究语义融合模块在数据集上的表现，表 4.5 进一步细化展示了 PolymerAbstracts 各个实体类别的实验结果，可以看出，通过将领域语言模型 MatSciBERT 与 MatetialsBERT 进行语义融合，可以在一定程度上互补各自对专业术语细微差异的理解，因此在实体类别属性名称 POLYMER 与属性值 PROPERTY_VALUE 的识别上拥有更好的表现，虽然能够对预训练语言模型的融合能实现语义补充，但依然无法避免预训练语言模型在处理低频词汇或专业

术语时，尤其是在分词和 OOV 问题上细粒度语义的丢失问题。因此，引入了词级向量 fastText，补充了语言模型在处理低频或形态变化复杂的专业术语时可能存在的不足，从而实现了在 PROPERTY_VALUE、PROPERTY_NAME 等类别上更精确的识别，并最终取得了最高的综合 F1 值。该结果说明，深层上下文表示的语义融合能够充分利用不同预训练模型的领域覆盖范围和上下文感知能力，而词级向量则能够提供在处理特殊词形、低频词汇以及数字单位等细粒度要素上的稳健性，共同提高了模型对聚合物文献中各类专业实体的识别精度和鲁棒性。为了验证结论的普适性，本章也对语义融合模块在高熵合金-硬度性能数据集上的各类实体的表现进行了具体分析。

表 4.6 语义融合模块在高熵合金-硬度性能数据集上 3 类实体的消融结果

语言模型	评价指标	Name	Preparation	Hardness
MatS	精确率	94.2	60.4	88.8
	召回率	94.4	63.0	71.9
	F1	94.3	61.6	79.4
Mate	精确率	93.3	53.0	95.3
	召回率	95.7	64.8	78.2
	F1	**94.5**	58.3	85.9
MatS+Mate	精确率	95.3	59.8	96.6
	召回率	90.6	45.0	74.3
	F1	92.6	51.4	84.0
MatS+FT	精确率	95.0	61.5	93.2
	召回率	90.9	54.3	70.5
	F1	92.9	57.7	80.2
Ours	精确率	91.2	56.1	90.4
	召回率	97.0	70.3	84.6
	F1	93.0	**62.4**	**87.4**

本表格中使用“MatS”表示“MatSciBERT”，使用“Mate”表示 MateialsBERT，“BioL”表示 BioLinkBERT，“WV”代表生物 fastText。黑体为该实体类别在 F1 指标上取得的最高值；波浪线表示该实体类别精确率取得的最高值；下划线表示该实体类别在召回率上取得的最高值。

如表 4.6 所示，单一的领域语言模型 MatSciBERT 和 MaterialsBERT 能够较为准确地识别拥有较为统一结构的文本；但在处理含有复杂工艺描述的工艺实体 Preparation 与数值信息频繁出现的硬度属性值实体 Hardness 时，往往难以兼顾低频词汇或工艺缩写等带来的分词和 OOV 问题。具体来说，两种领域语

言模型在合金名称实体（Name）的识别上均展现了较高的准确性，这主要得益于它们在大规模材料领域语料上对化学式、元素符号及命名规范的充分学习。MatSciBERT 在 Name 上取得了 94.3 的 F1 分数，而 MaterialsBERT 则略优，为 94.5。尽管二者在名称实体上表现接近，但在其他两类实体上表现差异较大：MatSciBERT 对包含“solution - treated at 1000 ° C followed by aging”等复杂制备工艺描述的 Preparation 实体具有更好的分词和上下文捕捉能力，其 F1 达到 61.6；而 MaterialsBERT 则在硬度属性值（Hardness）的识别上更为鲁棒，能够更加准确地处理“HRC 65”或“HV 650”格式的数值 + 单位表达，其 F1 达到 85.9，相较于 MatSciBERT 的 79.4 有显著提升。这样的差异反映出两种模型在各自预训练语料中对不同类型专业术语的敏感度和表达形式的偏好。

将领域语言模型 MatSciBERT 与 MatetialsBERT 进行语义融合后，对工艺制备与硬度数值的捕捉能力进一步提升，说明语义融合能让二者在学习各自预训练语料中细微差异的工艺和硬度表达形式方面进行一定的知识互补。然而，模型在识别 Preparation 与 Hardness 时仍易受分词不完整或专业名词频繁变化的影响，因而在精细化识别上并未达到最佳效果。为此，再引入词级向量 FT 以加强对罕见术语、数字单位及多形态工艺名称的鲁棒表征，有效弥补了预训练语言模型在处理低频或专业术语时的局限。

4.2.3 不同语料预训练模型语义融合的影响

为了探讨不同领域语言模型的语义融合效果，本章以生物科学文献领域的数据集 BC5CDR 为例，选用了经过不同语料训练的 BioBERT[7]，如下：

BioBERT-Gene：针对基因类实体，模型在 JNLPBA 数据集[100]和 BC2GM 数据集[101]上进行了微调。这些数据集包含了与基因相关的命名实体识别任务，主要用于识别基因类的实体。

BioBERT-Diseases：针对疾病类实体，BioBERT 模型在 BC5CDR-diseases 和 NCBI-diseases 数据集[102]上进行了微调。BC5CDR-diseases 数据集提供了疾病名称的标注，而 NCBI-diseases 数据集也涉及疾病的实体识别任务。

BioBERT-Chem：针对化学物质实体，BioBERT 模型在 BC5CDR-chemicals 和 BC4CHEMD 数据集[103]上进行了微调。这些数据集用于识别文章中提到的化学物质实体。

表 4.7 不同领域语言模型在数据集 BC5CDR 的语义融合效果

预训练数据集语料			实体类别	精确率	召回率	F1
chem	disease	gene				
✓			chem	94.10	88.65	92.42
			disease	88.19	90.80	88.42
			total	88.59	85.64	87.09
	✓		chem	87.38	92.50	89.87
			disease	88.24	88.57	88.40
			total	87.73	90.90	89.28
		✓	chem	89.35	90.69	90.02
			disease	77.56	77.12	77.34
			total	84.58	85.13	84.86
✓	✓		chem	92.66	94.71	93.67
			disease	87.73	89.39	88.55
			total	**91.97**	**92.38**	**92.18**
✓		✓	chem	95.06	93.89	94.47
			disease	82.77	80.28	81.51
			total	90.08	88.32	89.19
✓	✓	✓	chem	92.70	95.37	94.02
			disease	88.91	87.89	88.40
			total	91.19	92.31	91.75

黑体为 total 在各指标取得的最高值；波浪线表示实体类别 chem 在各指标取得的最高值；下划线表示实体类别 disease 在各指标取得的最高值。

可以从表 4.7 看出，在每个领域语料的微调过程中，模型通过学习该领域特定的实体标签及其上下文关系，从而有效提升了模型在各自领域内的表现。然而，单独使用单一领域的语料训练模型时，模型的泛化性能存在一定的局限性。比如，在化学物质（chem）实体识别任务中，仅使用 BC5CDR-chemicals 数据集时，模型的精确率为 94.10%，而疾病实体（disease）识别时的精确率为 88.19%。这表明，单一领域的语料虽然能有效提升模型在该领域内的表现，但其跨领域的泛化能力较弱。

为了验证不同领域语言模型的语义融合效应，将来自化学、疾病和基因领域的预训练模型进行语义融合拼接。具体而言，将这三类领域的模型进行组合，

利用领域间的互补信息，提升模型对多种实体类型的识别能力。在实验中发现，经过语义融合拼接后的模型在多个实体类别的识别精度上有了显著提升。比如，在 chem 和 disease 两个领域融合时，模型的总精确率，召回率以及 F1 达到了最高分数，然而，在进一步的实验中发现，当将基因领域的预训练模型加入融合体系后，总体的精确率、召回率以及 F1 反而出现了下降。这主要是因为基因语料所蕴含的信息与化学物质和疾病语料的特征存在不一致性，导致在进行语义融合时，不同领域知识之间出现冲突或信息噪声，从而削弱了整体模型的识别能力。因此可以看出，在合理的情况下，预训练模型的语义融合能够提升模型性能；而在不同领域语料差异较大的情况下，盲目的语义融合可能导致整体性能下降。

4.2.4 对比实验

为了进一步证明方法的有效性，本章提供了与其他方法的对比实验结果，对比模型包括 UniversalNER[50]、GoLLIE[51]、RDANE[104]、RUIE[105]、PubMedBERT[43]、BioDistilBERT[66]和 SciDeBERTa[86]。

其中 UniversalNER 和 GoLLIE 都依赖大语言模型的指令跟随能力：前者通过蒸馏将 ChatGPT 的开放式实体识别知识压缩到中型模型，后者则让大语言模型在训练阶段充分阅读标注规范，从而在推理时用“指南遵循”的方式完成零样本抽取。RUIE 在此基础上加入检索增强环节，利用双编码器先为测试样本检索少量相似示例，再把检索到的示例与待抽取文本拼接给大语言模型进行上下文推理，以缓解纯零样本场景下的知识缺口。RDANER 不依赖外部词典，而是通过噪声对比学习和域对抗训练，在低资源情况下最大限度保持跨域语义一致性，目的是在源域和目标域分布差异较大时仍能维持稳定抽取性能；这一策略强调“模型结构 + 学习目标”对噪声和转移的内在抵御，而非外部知识补充。PubMedBERT 自底向上地使用 PubMed 全量语料从零开始训练，提供纯生物医学视角的词汇和句法表征；SciDeBERTa 在 S2ORC 大规模科学论文上继续预训练 DeBERTa，利用解耦注意力与相对位置偏置对科研语料做更精细建模；BioDistilBERT 则将大型生物医学 BERT 压缩为 6 层网络，以降低推理成本，适合资源受限场景。

表 4.8 不同模型在数据集 BC5CDR 的效果

语言模型	精确率			召回率			F1		
	chem	disease	total	chem	disease	total	chem	disease	total
UniversalNER	-	-	-	-	-	-	-	-	89.3
GoLLIE	-	-	-	-	-	-	-	-	88.4
RDANER	-	-	-	-	-	-	-	-	87.4
RUIE	-	-	-	-	-	-	-	-	89.6
PubMedBERT	93.3	84.8	90.0	92.9	84.4	89.6	93.1	84.6	89.8
BioDistilBERT	83.4	69.0	77.9	78.5	59.4	70.9	80.9	63.9	74.2
SciDeBERTa	92.9	82.2	88.9	93.1	84.5	89.9	93.0	83.4	89.4
Ours (LLM)	92.2	85.4	88.7	91.7	87.3	89.5	91.9	86.3	89.1
Ours (领域语言模型)	**93.7**	**86.4**	**90.7**	**95.20**	**88.03**	**92.3**	**94.5**	**87.2**	**91.5**

从表 4.8 可见，大语言模型驱动框架在缺乏额外标注的前提下即可取得与传统预训练模型接近的成绩，尤其是 RUIE 通过检索示例显著缩小了与领域模型之间的差距；然而指令蒸馏或零样本指南仍难完全覆盖高专业度实体的形态变体，导致对疾病类实体的辨识明显受限。鲁棒域适应的 RDANER 在跨域一致性方面具备优势，但在高度领域化的 BC5CDR 上因缺少显式生物医学知识，其精度与召回均处中游。

对于面向科学文献领域的领域语言模型 PubMedBERT 和 SciDeBERTa，在 BC5CDR 数据集上的化学实体（chem）与疾病实体（disease）识别上均取得较为接近的性能。BioDistilBERT 的表现不太理想，可能是因为 BioDistilBERT 虽部署友好，但对长依赖和稀有缩写的处理能力有所折损。与通用模型 RDANER[104]等相比，基于专业预训练的 PubMedBERT 和 SciDeBERTa 在精准率和召回率方面普遍表现更优，因此可以看出领域语言模型在生物医学领域对于专有名词、缩写等方面等的优势。

相较于第三章中基于 LLM 的方法，基于领域语言模型的方法展现了在特定专业领域中的优势。再次证明了对于在术语高度集中且缩写频繁的专业领域，LLM 缺乏相应的领域知识，导致识别精度也可能受限，因此基于领域语言模型的方法通过针对性的预训练，能更好地融入了相应领域的专业词汇与知识。对于生物文献领域的 BC5CDR 数据集，从整体结果来看，本章提出的语义融合模

块在化学实体（chem）与疾病实体（disease）识别上均取得了最好的表现，具体来说，词级 Word2Vec 嵌入显式校正了化学名词的形态和缩写差异，使化学实体召回显著提高；而领域语言模型融合的 BioBERT 和 BioLinkBERT 补足了疾病实体的领域知识。

4.3 方法应用

本章用提出的语义融合模块将其与 BiLSTM-CRF 网络整合，以本工作创建的高熵合金领域为具体应用领域，用标注的高熵合金-硬度性能数据集进行微调后的模型从 2689 篇高熵合金（HEA）文章摘要中提取信息。实现了对过去五年内发表的 2698 篇论文的自动化分析，从中提取了 8067 个数据点。通过在数据分析中融入材料领域知识，识别出具有高潜力的元素和关键加工条件，从而指导机器学习数据集的设计与构建。经过对目标文献的人工总结和整理，构建了一个包含 13 种元素硬度数据的数据集，用于训练性能预测的模型。然后采用遗传算法（GA）[106]与粒子群优化（PSO）[107]相结合，第一阶段用于探索新型合金体系，第二阶段则对成分比例进行精细调整。最后结合了 SHAP 特征重要性[108]和皮尔逊相关系数[109]，并辅以材料领域知识，以验证研究成果并指导合金体系的选择。最后，成功设计出三种与现有数据集不同的高熵合金。预测结果显示，其硬度的平均相对误差低于 5%，且最优合金的硬度仅比历史记录低 38 HV，总框架如图 4.3。

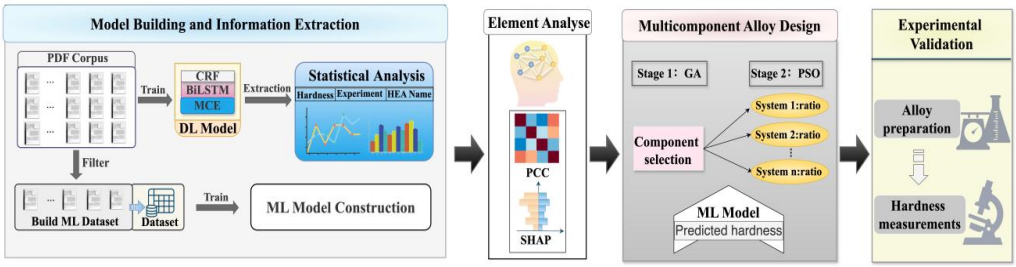


图 4.3 应用总框架

从检索到的 2698 篇文章摘要中提取了信息，并对提取出的名称进行了成分分析。图 4.4 展示了每年出现频率最高的前十个成分，根据材料专家的建议，相比其他元素，考虑到添加 Mo 能通过固溶强化显著提高合金硬度，将 Mo 作为主要元素进行关注，并在图中予以突出显示。

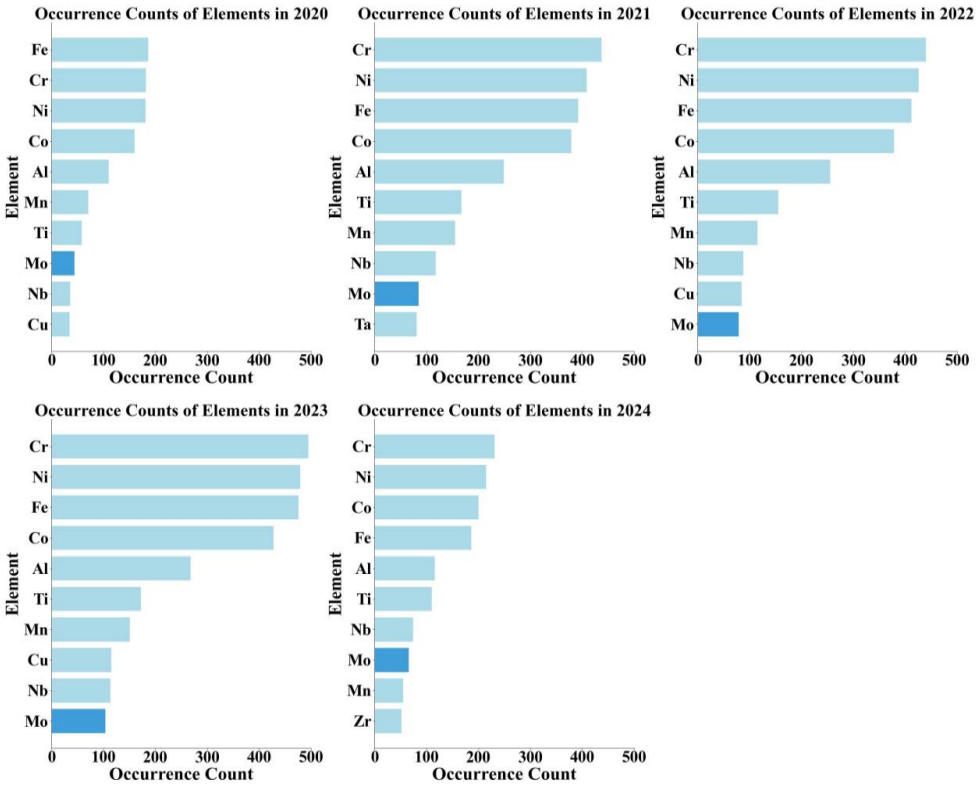


图 4.4 出现频率最高的前十种成分。

如图 4.4 所示，Cr、Ni、Co 和 Fe 在过去五年中一直是出现频率最高的元素，这表明基于 CoCrFeNi 的合金因其在高熵合金体系中卓越的机械性能而受到广泛认可和深入研究。Mn 和 Al 的高出现频率对应于最早的高熵合金之一——五元素 CrMnFeCoNi 合金（亦称 Cantor 合金[110]），以及另一种典型的基于 CoCrFeNi 的合金，即 CoCrFeNiAl 合金。这表明，高熵合金机械性能的研究主要集中在 CoCrFeNiMn 和 AlCoCrFeNi 高熵合金体系上[111]。近期研究表明，由于 Mo 拥有多种协同强化机制及低密度[112]，它作为高硬度 HEA 的候选元素

展现出巨大潜力[113]。并且 Mo 在图 4.4 所示的文献频率统计中，在 2020 至 2024 年间每年均能稳定出现在出现频率较高的元素行列，这表明围绕 Mo 的研究持续活跃，并积累了充足的文献资源可供数据挖掘和机器学习分析。因此，将 Mo 作为机器学习驱动的 HEA 系统设计中的关键组分予以优先考虑，其原因在于其在固溶强化、析出相调控及高温稳定性方面的协同效应。

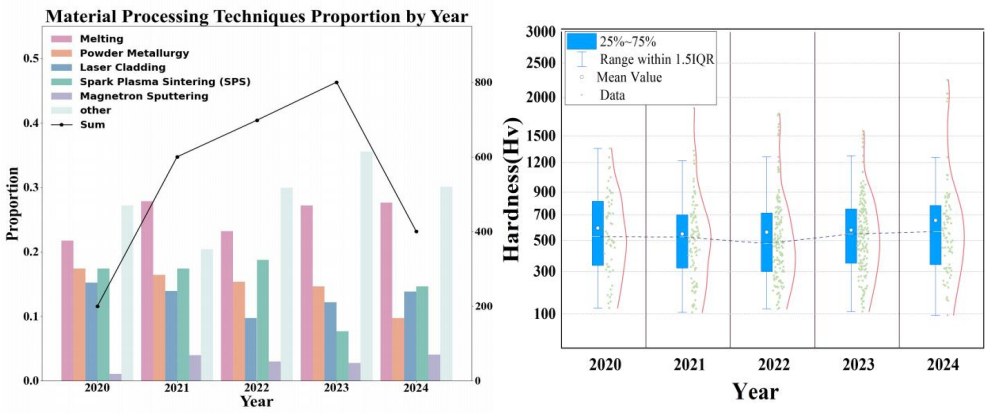


图 4.5 实体信息提取的统计结果。（a）实验方法统计图。（b）硬度统计图。

图 4.5(a)展示了 2020 年至 2024 年期间用于高熵合金硬度研究的材料加工技术的变化情况。熔炼技术仍然是最常用的制备方法，其使用量在 2020 年至 2024 年期间保持稳定。解释了为何通过机器学习收集的大多数数据主要基于电弧熔炼。图 4.4(b)显示，2020 年硬度主要集中在 500～900 HV；2021 年中位数和均值小幅上升，分布向高硬度扩展；2022 年整体硬度进一步提升且波动性增强；2023 年中位数和均值略降，但高硬度值仍较显著；2024 年整体硬度下降，但仍出现最大 2350.9 HV 的极高值。这些变化反映出在追求高硬度的同时，高熵合金在韧性和耐磨性等方面面临平衡优化的挑战。

借鉴了图 4.4 和图 4.5(a)中的分析结果。基于出现频率最高的前十种元素，选择 Mo 作为关键组分。采用了最广泛使用的电弧熔炼工艺作为过滤标准，从而指导机器学习（ML）数据集的特征设计，最后经过对目标文献的人工总结和整理，构建了一个包含 13 种元素硬度数据的数据集，并采用了支持向量机[115]、人工神经网络[116]、随机森林[117]和 XGBoost[118]来建立预测模型。鉴于数据集中包含大量设计的组分组合且各数据点之间的特征冗余极少，每个数据点都提供

了独特的信息，有助于模型的学习。因此，使用了留一法交叉验证（LOOCV）来进行交叉验证，以避免数据集的过拟合或欠拟合。最后，采用了均方根误差（$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}$）来评价模型的性能，以及决定系数（$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\widehat{y_i})^2}{\sum_{i=1}^{n}(y_i-\overline{y_i})^2}$）作为评估指标，用于评估模型的拟合精度并选择性能最佳的模型。图 4.6 展示了这四种机器学习算法在数据集上的拟合精度：

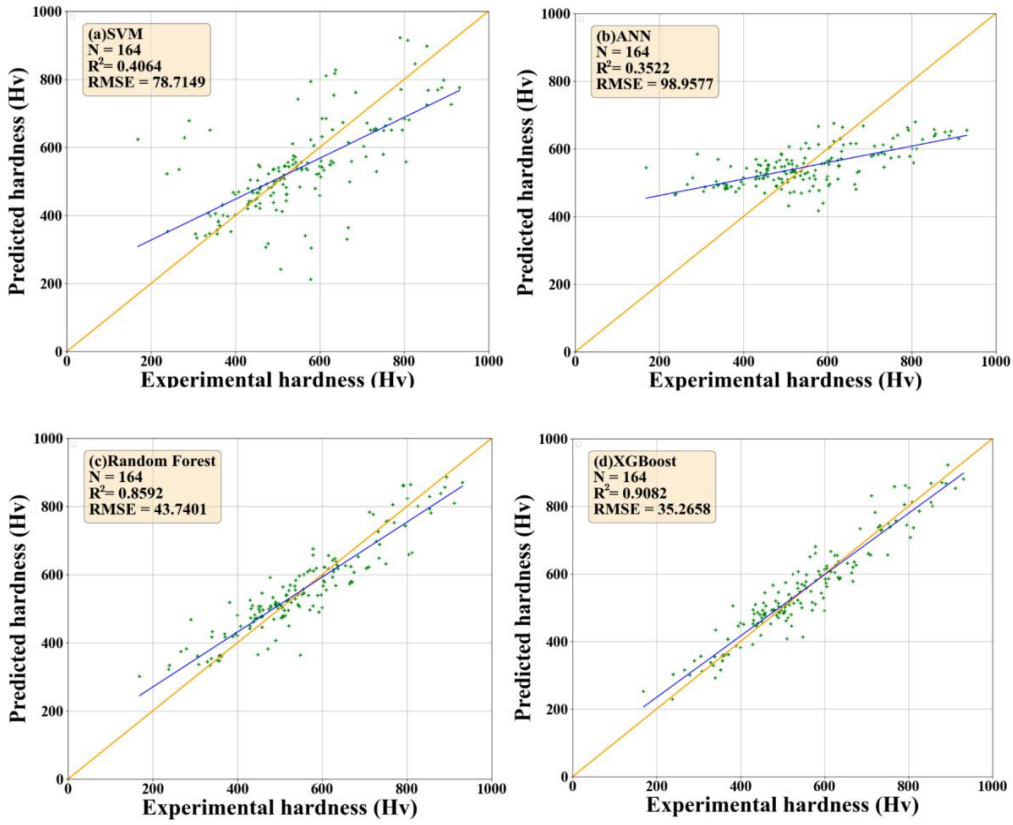


图 4.6 使用不同机器学习模型预测和实验硬度的 HEA 比较。拟合曲线展示了(a) SVM、(b) ANN、(c)随机森林和(d) XGBoost 模型在预测硬度方面的表现。黄色线表示理想拟合，即预测硬度与实验值相符（y = x），而蓝色线则显示了每个模型的实际回归拟合结果。

可以从图 4.6 看出，基于树的算法表现异常优异，其中 XGBoost 模型的表现最佳，决定系数达到最高值 0.9082，均方根误差降至最低，仅为 35.2658。结

合图 4.6 的机器学习拟合结果, 选用 XGBoost 模型来预测硬度, 并且结合 SHAP 可解释性分析和皮尔逊相关系数分析定量评估了各元素对硬度的贡献, 如图 4.7 所示：

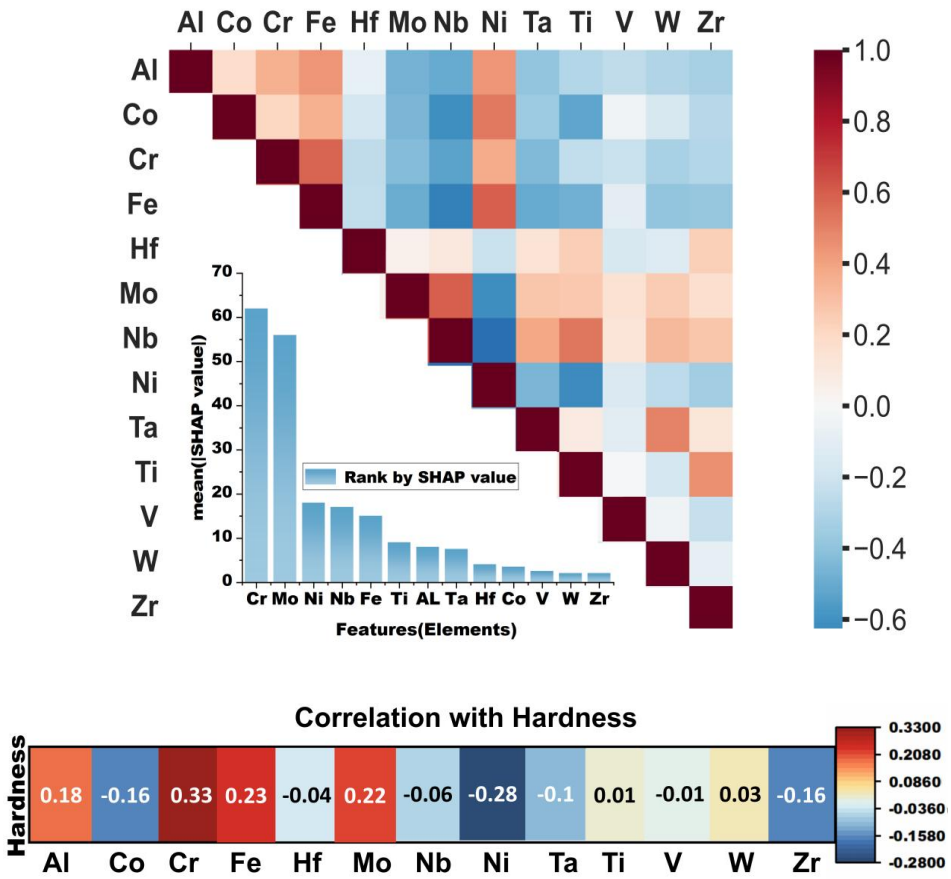


图 4.7 模型解释与成分分析。(a) 元素的皮尔逊相关系数 (PCC) 热力图；左下角的插图展示了机器学习优化的元素的 SHAP 重要性评估排名。(b) 材料成分与硬度之间的相关性分析；单个元素与 HEA 硬度之间的相关性通过 PCC 方法计算。

本研究先用 SHAP 与皮尔逊相关系数将合金元素按对硬度的正向、负向或中性贡献分为四类，然后基于仅含正面或中性元素的候选方案，通过采用遗传算法[106]与粒子群优化[107]相结合，第一阶段用于探索新型合金体系，第二阶段则对成分比例进行精细调整，最终设计并制备了三种高硬度高熵合金并通过维氏硬度测试验证了其性能。

表 4.9 设计的高熵合金的实验硬度和预测硬度

Composition	ML 预测硬度(HV)	实验平均硬度(HV)	MRE
Cr20.6Fe22.5Mo20.6Ti18.3V18	811.3	808.5	0.34%
Al9.32Cr20.62Fe21.71Mo27.09Ti21.26	805.8	790.4	1.95%
Al6Cr20.3Fe19.5Mo20.1Nb18.8Ti15.3	921.5	893	3.19%

表 4.9 显示，设计的三种高熵合金的实验硬度均超过 750 HV，其中最高硬度达到 893 HV，在收集的数据集中排名第三，仅比最高值低 38 HV。实验结果与预测结果之间的平均相对误差（MRE）低于 5%，表明该模型在对不同合金成分进行硬度预测时表现良好，误差极小。

4.4 本章小结

本章围绕领域语言模型的语义融合策略，探讨了在高度专业化语境下提升科学文献命名实体识别准确性的方法。首先选取与目标学科相匹配的领域 语言模型以捕捉深层上下文语义，再将其输出与同一领域的词级向量进行融合，结合 BiLSTM-CRF 框架承载融合后的词表示，使序列标注器保持对长程依赖的敏感度。为了验证方法在不同领域的适用性，除第三章的生物领域的 BC5CDR 数据集外，加入了材料文献领域的 PolymerAbstracts 数据集和构建的高熵合金-硬度性能数据集进行补充。实验结果表明无论是在生物医学文献还是材料科学文献，语义融合模块都能显著提高模型在精确率与召回率之间的平衡。最后，将模型迁移至具体领域，大规模提取材料-工艺-性能三类关键信息，设计出三种具备高硬度潜力的合金，验证了方法在真实科研场景中的应用价值与可扩展性。

第五章　总结与展望

5.1 总结

本论文围绕科学文献命名实体识别（NER）中的关键挑战展开研究，对于当前科学文献在数据规模、专业性与跨学科复杂性持续增长的背景下，传统人工检索方式难以兼顾效率与准确性的问题，提出了一种结合显式标注与直接偏好优化的双阶段训练方法，并进一步引入基于领域语言模型的语义融合方法，以提升语言模型在科研文献领域中命名实体识别的边界判定能力，类别辨别能力以及跨领域的适应性。论文所提出的各项方法均在多个实际数据集上进行了系统性实验验证，展现了良好的通用性和应用潜力。

（1）在方法设计方面，本文首先引入了上下文一致的实体显式标注方法，通过在实体前后插入类别标识符对，使得实体边界在不破坏句法结构的情况下被明确标注。随后，在训练阶段提出了“监督微调 + 直接偏好优化”的双阶段方法。在监督微调阶段，模型首先在大规模标注数据上学习基础的实体识别能力；在此基础上，进一步引入直接偏好优化算法，将构造的正负样本对作为输入，利用对数概率差优化策略，引导模型在实体边界扩张、收缩及类别混淆等典型错误场景中更倾向于正确标注，从而提升整体判别性能。在负样本构造策略上，本文通过模拟推理阶段可能出现的实体识别误差，构建了包括边界错误与类别错误在内的三种负样本类型，并对样本进行筛选，保证偏好学习过程中模型能够聚焦于实体标注的细粒度差异。

（2）为了进一步提高模型对专业术语、缩写、数值单位等低频知识的识别能力，本文提出了基于领域语言模型的语义融合方法，选取与目标任务相匹配的领域语言模型作为编码器，将其生成的上下文表示与领域内训练得到的词级向量进行融合，并使用 BiLSTM-CRF 框架对融合后的表示进行序列建模。

（3）在实验评估方面，本文在多个科学文献领域的数据集上对提出的方法进行了系统验证。在计算机科学领域的 SCIERC 数据集上，通过与已有多种 NER 模型的对比实验发现，本文方法整体性能在该数据集上达到目前已知的最佳水

平，证明了显式标注与 DPO 阶段对模型判别能力的有效增强。同时，在生物医学领域的 BC5CDR 数据集上也表明，领域存在大量缩写和专业名词的情况下，大语言模型的表现受限。针对这一问题，提出了基于领域语言模型的语义融合方法，让模型对关键术语的识别表现得到明显提升，并且为了验证方法在不同领域的普适性，进一步使用了 PolymerAbstracts 和高熵合金-硬度性能两个材料科学数据集实验。实验结果显示，无论是在生物医学文献还是材料科学文献中，语义融合方法均能够提升模型在精确率与召回率之间的平衡能力，特别是在低频实体、形态变体识别上的提升尤为显著，验证了深层语义表示与词级语义信息之间的互补效应。

（4）在高熵合金文献中的实际应用表明，模型能够自动识别并提取“材料－工艺－性能”三类关键信息，并结合合金设计的多阶段优化框架，辅助指导设计了三种具备高硬度潜力的新型高熵合金设计方案，进一步证明了所提方法在科研文本结构化与智能化设计流程中的实际落地能力。

5.2 展望

尽管本论文提出的方法在多领域文献的命名实体识别任务上取得了较为理想的效果，仍有若干尚待深化的研究方向。首先，在实体类别与边界标注的复杂交互中，本研究主要集中于非嵌套实体，但对于超长文本或多实体交错频繁的场景，如何提高识别准确度仍是一个值得进一步探究的问题。其次，在更加开放的科学研究场景中，实体识别本身往往只是信息抽取流程中的第一步。后续的关系抽取、事件检测以及知识图谱构建等更深层次的任务相互结合，将有望在科研文献的分析和管理中发挥更大功效。尤其在材料科学、生物医学等新兴交叉领域中，对大量实验过程和研究数据的快速扫描与深度解析，更有利于推动科研创新与决策支持。

参考文献

[1] NIKLAUS C, CETTO M, FREITAS A, et al. A Survey on Open Information Extraction[C]//27th International Conference on Computational Linguistics. 2018.

[2] 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: 24824-37.

[3] SCHAEFFER R, MIRANDA B, KOYEJO S. Are emergent abilities of large language models a mirage? [J]. Advances in Neural Information Processing Systems, 2023, 36: 55565-81.

[4] WANG S, SUN X, LI X, et al. Gpt-ner: Named entity recognition via large language models [J]. arXiv preprint arXiv:230410428, 2023.

[5] DING Y, LI J, WANG P, et al. Rethinking Negative Instances for Generative Named Entity Recognition[C]//Findings of the Association for Computational Linguistics ACL 2024. 2024: 3461-3475.

[6] GUPTA T, ZAKI M, KRISHNAN N A, et al. MatSciBERT: A materials domain language model for text mining and information extraction [J]. npj Computational Materials, 2022, 8(1): 102.

[7] 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): 1234-40.

[8] ZHANG Q, DING K, LV T, et al. Scientific large language models: A survey on biological & chemical domains [J]. ACM Computing Surveys, 2025, 57(6): 1-38.

[9] MIKOLOV T, KARAFIáT M, BURGET L, et al. Recurrent neural network based language model[C]//Interspeech. 2010, 2(3): 1045-1048.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-80.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[12] RUDER S, PETERS M E, SWAYAMDIPTA S, et al. ransfer learning in natural language processing[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials. 2019: 15-18.

[13] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[14] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [J]. 2018.

[15] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[16] KALE M, RASTOGI A. Text-to-Text Pre-Training for Data-to-Text Tasks[C]//Proceedings of the 13th International Conference on Natural Language Generation. 2020: 97-102.

[17] Snell C V, Wallace E, Klein D, et al. Predicting Emergent Capabilities by Finetuning[C]//First Conference on Language Modeling.

[18] Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models [J]. Transactions on Machine Learning Research.

[19] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways [J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[20] DALE R. GPT-3: What’ s it good for? [J]. Natural Language Engineering, 2021, 27(1): 113-8.

[21] ZHUANG H, QIN Z, HUI K, et al. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). 2024: 358-370.

[22] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-901.

[23] 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: 27730-44.

[24] LU Y, ZHU W, LI L, et al. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 10748-10772.

[25] TAN J C M, MOTANI M. Large Language Model (LLM) as a System of Multiple Expert Agents: An Approach to solve the Abstraction and Reasoning Corpus (ARC) Challenge [J]. arXiv e-prints, 2023: arXiv: 2310.05146.

[26] EKIN S. Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices [J]. Authorea Preprints, 2023.

[27] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways [J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[28] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report [J]. arXiv preprint arXiv:230308774, 2023.

[29] KALLA D, SMITH N, SAMAAH F, et al. Study and analysis of chat GPT and its impact on different fields of study [J]. International journal of innovative science and research technology, 2023, 8(3).

[30] FANOURAKIS N, EFTHYMIOU V, KOTZINOS D, et al. Knowledge graph embedding methods for entity alignment: experimental review [J]. Data Mining and Knowledge Discovery, 2023, 37(5): 2070-137.

[31] CHINCHOR N A, SUNDHEIM B. Message Understanding Conference (MUC) tests of discourse processing[C]//Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation. 1995: 21-26.

[32] SUTTON C, MCCALLUM A. An introduction to conditional random fields [J]. Foundations and Trends® in Machine Learning, 2012, 4(4): 267-373.

[33] LIU S, HE T, DAI J. A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese [J]. Mobile Networks and Applications, 2021, 26(5): 1891-903.

[34] WANG Y, TONG H, ZHU Z, et al. Nested named entity recognition: a survey [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2022, 16(6): 1-29.

[35] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012, 25.

[36] WANG H, LI J, WU H, et al. Pre-trained language models and their applications [J]. Engineering, 2023, 25: 51-65.

[37] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[38] SHARMA A, AMRITA, CHAKRABORTY S, et al. Named entity recognition in natural language processing: A systematic review[C]//Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021. Springer Singapore, 2022: 817-828.

[39] BELTAGY I, LO K, COHAN A. SciBERT: A pretrained language model for scientific text [J]. arXiv preprint arXiv:190310676, 2019.

[40] GHOSH M, MUKHERJEE S, SANTRA P, et al.BLINKtextsubscriptLSTM: BioLinkBERT and LSTM based approach for extraction of PICO frame from Clinical Trial Text[C]//Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD). 2024: 227-231.

[41] ALSENTZER E, MURPHY J, BOAG W, et al. Publicly Available Clinical BERT Embeddings[J]. NAACL HLT 2019, 2019: 72.

[42] YASUNAGA M, LESKOVEC J, LIANG P. LinkBERT: Pretraining Language Models with Document Links[C]//Association for Computational Linguistics (ACL). 2022..

[43] 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.

[44] SHETTY P, RAJAN A C, KUENNETH C, et al. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing [J]. npj Computational Materials, 2023, 9(1): 52.

[45] LI Z, FAN S, GU Y, et al. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(17): 18608-18616.

[46] MOSLEM Y, ROMANI G, MOLAEI M, et al. Adaptive Machine Translation with Large Language Models[C]//Proceedings of the 24th Annual

Conference of the European Association for Machine Translation. 2023: 227-237..

[47] SU W, TANG Y, AI Q, et al. Mitigating entity-level hallucination in large language models[C]//Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. 2024: 23-31.

[48] HUANG L, YU W, MA W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions [J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[49] YAN F, YU P, CHEN X. LTNER: Large language model tagging for named entity recognition with contextualized entity marking[C]//International Conference on Pattern Recognition. Springer, Cham, 2025: 399-411.

[50] ZHOU W, ZHANG S, GU Y, et al. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition[C]//The Twelfth International Conference on Learning Representations.

[51] SAINZ O, GARCíA-FERRERO I, AGERRI R, et al. GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction[C]//ICLR. 2024.

[52] KANG H, SEO H, JUNG J, et al. Guidance-Based Prompt Data Augmentation in Specialized Domains for Named Entity Recognition[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2024: 665-672.

[53] ZHOU Y, SHAN S, WEI H, et al. PGA-SciRE: Harnessing LLM on Data Augmentation for Enhancing Scientific Relation Extraction [J]. arXiv preprint arXiv:240520787, 2024.

[54] LI Q, XIE T, ZHANG J, et al. Enhancing named entity recognition with external knowledge from large language model [J]. Knowledge-Based Systems, 2025: 113471.

[55] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model [J]. Advances in Neural Information Processing Systems, 2023, 36: 53728-41.

[56] CHURCH K W. Word2Vec [J]. Natural Language Engineering, 2017, 23(1): 155-62.

[57] JOULIN A, GRAVE E, MIKOLOV P B T. Bag of Tricks for Efficient Text Classification [J]. EACL 2017, 2017: 427.

[58] HUANG A H, WANG H, YANG Y. FinBERT: A large language model for extracting information from financial text [J]. Contemporary Accounting Research, 2023, 40(2): 806-41.

[59] CHALKIDIS I, FERGADIOTIS M, MALAKASIOTIS P, et al. LEGAL-BERT: The Muppets straight out of Law School[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. 2020: 2898-2904.

[60] HUANG K, ALTOSAAR J, RANGANATH R. ClinicalBert: Modeling Clinical Notes and Predicting Hospital Readmission [J]. arXiv preprint arXiv:190405342, 2019.

[61] MANN B, RYDER N, SUBBIAH M, et al. Language models are few-shot learners [J]. arXiv preprint arXiv:200514165, 2020, 1: 3.

[62] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models [J]. arXiv preprint arXiv:230213971, 2023.

[63] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing [J]. ACM computing surveys, 2023, 55(9): 1-35.

[64] MESKó B. Prompt engineering as an important emerging skill for medical professionals: tutorial [J]. Journal of medical Internet research, 2023, 25: e50638.

[65] HAIDER Z, RAHMAN M H, DEVABHAKTUNI V, et al. A framework for mitigating malicious RLHF feedback in LLM training using consensus based reward [J]. Scientific Reports, 2025, 15(1): 9177.

[66] FU Z, YANG H, SO A M-C, et al. On the effectiveness of parameter-efficient fine-tuning[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(11): 12799-12807.

[67] HU E J, WALLIS P, ALLEN-ZHU Z, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.

[68] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report [J]. arXiv preprint arXiv:241219437, 2024.

[69] PANCHENDRARAJAN R, AMARESAN A. Bidirectional LSTM-CRF for named entity recognition[C]//Proceedings of the 32nd Pacific Asia conference on language, information and computation. 2018.

[70] MOSCATO V, POSTIGLIONE M, SPERLí G. Few-shot named entity recognition: definition, taxonomy and research directions [J]. ACM Transactions on Intelligent Systems and Technology, 2023, 14(5): 1-46.

[71] VAN HOUDT G, MOSQUERA C, NáPOLES G. A review on the long short-term memory model [J]. Artificial Intelligence Review, 2020, 53(8): 5929-55.

[72] RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning [M]. Natural language processing using very large corpora. Springer. 1999: 157-76.

[73] SOHRAB M G, MIWA M. Deep exhaustive model for nested named entity recognition[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 2843-2849.

[74] SANG E F T K, DE MEULDER F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition [J]. Development, 1837, 922: 1341.

[75] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [J]. 2018.

[76] DING Y, LI J, WANG P, et al. Rethinking Negative Instances for Generative Named Entity Recognition[C]//Findings of the Association for Computational Linguistics ACL 2024. 2024: 3461-3475.

[77] LUAN Y, HE L, OSTENDORF M, et al. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3219-3232.

[78] AUGENSTEIN I, DAS M, RIEDEL S, et al. SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017: 546-555.

[79] BUSCALDI D, SCHUMANN A-K, QASEMIZADEH B, et al. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers[C]//International Workshop on Semantic Evaluation (SemEval-2018). 2017: 679-688.

[80] LI J, SUN Y, JOHNSON R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction [J]. Database, 2016, 2016.

[81] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[82] LIN C-Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of the Workshop on Text Summarization Branches Out, 2004. 2004.

[83] SUNDHEIM B M. Overview of results of the MUC-6 evaluation[C]//Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995. 1995.

[84] HU E J, WALLIS P, ALLEN-ZHU Z, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.

[85] LOSHCHILOV I, HUTTER F. SGDR: Stochastic Gradient Descent with Warm Restarts[C]//International Conference on Learning Representations. 2022.

[86] JEONG Y, KIM E. Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks [J]. IEEE Access, 2022, 10: 60805-13.

[87] LLOYD S. Least squares quantization in PCM [J]. IEEE transactions on information theory, 1982, 28(2): 129-37.

[88] YE D, LIN Y, LI P, et al. Packed Levitated Marker for Entity and Relation Extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 4904-4917.

[89] ZARATIANA U, TOMEH N, HOLAT P, et al. An autoregressive text-to-graph framework for joint entity and relation extraction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(17): 19477-19487.

[90] LV H, HAN X, WANG P, et al. Joint Entity and Relation Extraction Based on Prompt Learning and Multi-channel Heterogeneous Graph Enhancement[C]//2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA). IEEE, 2024: 1584-1590.

[91] EBERTS M, ULGES A. Span-based joint entity and relation extraction with transformer pre-training [M]. ECAI 2020. IOS Press. 2020: 2006-13.

[92] WANG Y, SUN C, WU Y, et al. UniRE: A Unified Label Space for Entity Relation Extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 220-231.

[93] SHEN Y, MA X, TANG Y, et al. A trigger-sense memory flow framework for joint entity and relation extraction[C]//Proceedings of the web conference 2021. 2021: 1704-1715.

[94] FELLBAUM C. WordNet: An electronic lexical database [M]. MIT press, 1998.

[95] WEI J, ZOU K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 6382-6388.

[96] WHITEHOUSE C, CHOUDHURY M, AJI A. LLM-powered Data Augmentation for Enhanced Cross-lingual Performance[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 671-686.

[97] BISHOP C M, NASRABADI N M. Pattern recognition and machine learning [M]. Springer, 2006.

[98] KIM E, JENSEN Z, VAN GROOTEL A, et al. Inorganic materials synthesis planning with literature-trained neural networks [J]. Journal of chemical information and modeling, 2020, 60(3): 1194-201.

[99] ZHANG Y, CHEN Q, YANG Z, et al. BioWordVec, improving biomedical word embeddings with subword information and MeSH [J]. Scientific data, 2019, 6(1): 52.

[100] COLLIER N, OHTA T, TSURUOKA Y, et al. Introduction to the bio-entity recognition task at JNLPBA[C]//Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). 2004: 73-78.

[101] SMITH L, TANABE L K, ANDO R J N, et al. Overview of BioCreative II gene mention recognition [J]. Genome biology, 2008, 9: 1-19.

[102] DOğAN R I, LEAMAN R, LU Z. NCBI disease corpus: a resource for disease name recognition and concept normalization [J]. Journal of biomedical informatics, 2014, 47: 1-10.

[103] KRALLINGER M, LEITNER F, RABAL O, et al. CHEMDNER: The drugs and chemical names extraction challenge [J]. Journal of cheminformatics, 2015, 7: 1-11.

[104] YU H, MAO X-L, CHI Z, et al. A robust and domain-adaptive approach for low-resource named entity recognition[C]//11th IEEE International Conference on Knowledge Graph, ICKG 2020. Institute of Electrical and Electronics Engineers Inc., 2020: 297-304.

[105] LIAO X, DUAN J, HUANG Y, et al. RUIE: Retrieval-based Unified Information Extraction using Large Language Model[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 9640-9655.

[106] SAMPSON J R. Adaptation in natural and artificial systems (John H. Holland) [Z]. Society for Industrial and Applied Mathematics. 1976

[107] EBERHART R, KENNEDY J. Particle swarm optimization[C]//Proceedings of ICNN'95-international conference on neural networks. ieee, 1995, 4: 1942-1948.

[108] COHEN I, HUANG Y, CHEN J, et al. Pearson correlation coefficient [J]. Noise reduction in speech processing, 2009: 1-4.

[109] Yule G U. On the theory of correlation[J]. Journal of the Royal Statistical Society, 1897, 60(4): 812-854.

[110] YEH J W, CHEN S K, LIN S J, et al. Nanostructured high‐entropy alloys with multiple principal elements: novel alloy design concepts and outcomes [J]. Advanced engineering materials, 2004, 6(5): 299-303.

[111] TSAI M-H, YEH J-W. High-entropy alloys: a critical review [J]. Materials Research Letters, 2014, 2(3): 107-23.

[112] GUO N, WANG L, LUO L, et al. Microstructure and mechanical properties of refractory MoNbHfZrTi high-entropy alloy [J]. Materials & Design, 2015, 81: 87-94.

[113] SHUN T-T, CHANG L-Y, SHIU M-H. Microstructure and mechanical properties of multiprincipal component CoCrFeNiMox alloys [J]. Materials Characterization, 2012, 70: 63-7.

[114] LIU Y, YANG Z, ZOU X, et al. Data quantity governance for machine learning in materials science [J]. National Science Review, 2023, 10(7): nwad125.

[115] CORTES C, VAPNIK V. Support-vector networks [J]. Machine learning, 1995, 20: 273-97.

[116] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. nature, 1986, 323(6088): 533-6.

[117] BREIMAN L. Random forests [J]. Machine learning, 2001, 45: 5-32.

[118] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

攻读硕士学位期间取得的研究成果

一、论文

[1] HAN Y，WANG H，XU P，et al. Deep Learning-Based Framework for Efficient Design of Multicomponent High Hardness High Entropy Alloys [J]. ACS Applied Materials & Interfaces， 2025.（已发表，导师一作，本人二作，中科院二区）

二、软著

软件名称：基于深度学习的合金材料文献数据挖掘和材料性能分析平台 V1.0，开发人：韩越兴、王慧、王冰。登记号：2023SR1430101，登记日期：2024.10.21，申请人：上海大学。（导师一作， 本人二作）

致　谢

时光荏苒，两年多的研究生岁月如白驹过隙；回首求学之路，离不开各位师长与挚友的帮助，在此谨致以我最诚挚的谢意。

首要感谢我的导师韩越兴老师。从论文的选题立意、实验设计至最终成文，每一步都凝聚着韩老师的心血与指导。韩老师治学严谨，其对学术的热忱与坚持，不仅培养了我独立思考、慎思明辨的能力，更教会我在科研探索中要勇于坚持自我。生活上，韩老师言传身教，他勤勉不辍的工作态度和张弛有度的处世智慧，使我深受启发，让我懂得了为人处事的道理，更明白了拼搏向上的真谛。此外，我同样要由衷感谢陈侨川老师对我论文的精准指点，以及张老师在科研习惯与细节把握上的谆谆教导。

其次，感谢我的父母总是尊重和包容我的选择；感谢我的妈妈，总是无条件陪伴支持我，教会我做一个考虑他人感受并且坚韧独立的人，感谢我的爸爸，教导我做一个正直勇敢并且自尊自爱的人。感谢和我一起上普拉提课的姐姐和教练们，在我生病时给我带的巧克力，还有好久不见的拥抱以及闲来无事的关切寒暄。感谢给予过我帮助的人们，可爱的学弟学妹们，学习榜样池姐，以及美丽可爱的朋友们，一起躲雨吃麻布屋并总是给予温暖关怀的派派，快乐深夜海底捞后来异地约饭顺带游戏厅的骊宝，还有雨天包场脱口秀相互加油打气的潘甜。

然后是亲爱的同门们，我们陪伴彼此走了好长一段难熬的路。和大家在一起的时光好像总会温暖一点点，难过的事也会被烘干一些些，变得不再那么沉重，那些稀松平常的日子已经成为了我不想放手的快乐时刻。现在按字母顺序依次提姓感谢性格有趣各异的各位：首先是生活习惯良好规律、善用工具且性格活泼的包姓同学，向您学习到了善待自己身体的生活习惯和健身技巧，还有感谢您在我论文返修阶段给予的可靠经验和有效建议。然后是经过努力拼搏已经勇猛精进（不止在羽毛球上）的凌姓同学，总是进步很快尤其是游泳，以后就是能线下约饭交流工作经验的同地同行了。接下来是已经从隔夜柠檬茶换成枸杞红枣桂圆茶，但对咖啡的爱始终如一的阮姓同学，虽然初见是不善言辞不

闲聊人设，但其实是能一起快乐吃瓜逻辑缜密且相对客观的优秀瓜友，而且现在已熟练掌握夸赞和安慰技能了。最后是积极主动、感性抒情、舞出生命的赵姓同学，就算不向 j 人转型，其实也是在某些方面靠谱很有责任心且目的性和执行性很强、很擅长做自己的 E 人朋友。

研究生的生活像上海漫长的雨季，一度将我淹没。当我终于放弃自我画饼，学会平视过去和现在的自己，我得以被拉出水面呼吸。谢谢从前那个白天步履匆匆穿梭于各个教室第一排，晚上踩着斑驳白雪和惨白路灯从图书馆回寝的自己；也谢谢现在这个厌倦了困难总是在哭但每次都坚定选择冷静面对的自己。我渐渐发现生活是一出如此艰难的剧本，还可能追加一些荒诞残忍的随机场景；也渐渐明白听上去成词滥调的祝愿承载了人们多少欲言又止的情绪和浓郁深切的关怀。祝愿我爱的人和爱我的人，祝愿努力生活着的人们都能被世界温柔以待。希望我能一直勇敢自由且温柔坚定，而生命见证过的真实和付出的努力都会成为我作为“我”在不同幕的叙事底色。