Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(6), 1398-1409
Published Online November 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.146131

Hans汉斯

基于信息抽取的基因编辑技术领域大模型偏见评估研究

陈　梅[1,2*]，高　扬[1,2]，石林钢[1,2]，何明净[1,2]，刘小英[1,2]

[1]中央民族大学民族语言智能分析与安全治理教育部重点实验室，北京
[2]中央民族大学信息工程学院，北京

收稿日期：2025年10月23日；录用日期：2025年10月31日；发布日期：2025年11月10日

摘　要

基因编辑技术是一项兼具重大应用潜力与生物安全风险的前沿技术。当前，大语言模型在其相关科研、传播中的应用日益广泛。如果大模型对基因编辑技术存在认知偏见，将可能引发生物安全风险。为系统评估大模型对基因编辑技术的认知偏见，本研究构建了一种基于事件要素的信息抽取分析框架。该框架聚焦“人物、组织、技术、对象、效果、发表期刊”六类核心事件要素，采用“Basic-Rethink-Multi-Query”的级联评估流程，并在gpt-3.5-turbo与gpt-4-turbo两类模型及多种提示策略下进行实验验证。研究结果显示，基础信息抽取环节存在显著的结构性偏见，表现为对静态实体(如技术、组织、发表期刊)的识别效果较好，而对动态要素(如人物、对象、效果)的识别能力较弱。在引入反思机制与多轮问答策略后，各要素的识别均衡性与整体性能均得到显著提升，但“对象”等特定要素的识别仍存在滞后，提示模型在领域语义理解上存在盲区。本研究通过信息抽取方法有效识别并缓解了大模型在基因编辑技术领域的认知偏见，为发展可信赖的信息处理技术、支持生物安全治理提供了方法依据与实证参考。

关键词

基因编辑，信息抽取，大语言模型，检索增强生成(RAG)，偏见评估

Bias Assessment of Large Language Models in Gene Editing Technology through Information Extraction

Mei Chen[1,2*], Yang Gao[1,2], Lingang Shi[1,2], Mingjing He[1,2], Xiaoying Liu[1,2]

[1]Key Laboratory of Intelligent Analysis and Security Governance of Ethnic Languages, Ministry of Education, Minzu University of China, Beijing
[2]School of Information Engineering, Minzu University of China, Beijing

[*]通讯作者。

文章引用: 陈梅, 高扬, 石林钢, 何明净, 刘小英. 基于信息抽取的基因编辑技术领域大模型偏见评估研究[J]. 人工智能与机器人研究, 2025, 14(6): 1398-1409. DOI: 10.12677/airr.2025.146131

Received: October 23, 2025; accepted: October 31, 2025; published: November 10, 2025

Abstract

Gene editing technology represents a rapidly advancing field with substantial application potential and significant biosecurity risks. Large language models (LLMs) are increasingly used for scientific research and communication in this domain. If such models exhibit cognitive bias in gene editing technology, they may exacerbate biosecurity risks. To systematically assess such bias, we propose an event-centric information extraction framework. The framework targets six core event elements—Person, Organization, Technology, Object, Effect, and Publish—and employs a cascaded evaluation pipeline ("Basic-Rethink-Multi-Query"). We evaluate gpt-3.5-turbo and gpt-4-turbo under multiple prompting strategies. Results reveal significant structural bias in the basic extraction stage: the models perform better on relatively static entities (Technology, Organization, Publish) compared to dynamic elements (Person, Object, Effect). Incorporating Rethink mechanisms and Multi-Query strategy substantially improves both inter-category balance and overall extraction performance. However, extraction of certain elements (e.g., Object) remains comparatively weak, indicating gaps in domain-specific semantic understanding. This study applies information extraction methods to identify and mitigate LLM cognitive bias in the domain of gene editing technology, thereby providing a methodological foundation and empirical evidence for trustworthy information processing technologies and biosecurity governance.

Keywords

Gene Editing, Information Extraction, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Bias Assessment

1. 引言

基因编辑技术通过对基因组进行精确修饰以赋予生物新特征或功能，是一类具有重大应用潜力且伴随显著生物安全风险的前沿生物技术。以 CRISPR/Cas9 为代表的基因编辑工具近年来发展迅速，在推动生物经济与科研创新的同时，也引发了伦理争议与生物安全等多重挑战。随之而来的是海量的相关文本(新闻、论文、元数据库等)，如何从中高效、可靠地抽取结构化信息以支撑风险评估与决策，已成为亟待解决的任务。

大语言模型(Large Language Models, LLMs)凭借其卓越的自然语言理解与生成能力，在信息抽取(Information Extraction, IE)任务中展现出显著的应用潜力(例如 ChatGPT [1])。然而，该类模型基于“大规模无监督预训练与指令微调”的范式，不可避免地习得了训练语料中所蕴含的社会偏见(如 Stereotypes)及误导性陈述[2]。此类偏见可能进一步影响模型在语义解析与实体抽取过程中的表现，导致其对特定语义要素产生系统性偏好或忽视。当前，这一风险在具有重大社会与伦理意义的基因编辑技术领域尚未得到充分检验。具体而言，LLM 在基因编辑技术领域的信息抽取中存在何种认知偏见、这些偏见的具体表现与形成机制是什么、以及有哪些有效的缓解策略，仍是当前缺乏实证研究的关键问题。

在大语言模型信息抽取领域，国内研究多聚焦于领域适配与任务导向的模型优化，例如本地化部署

与提示工程以应对低资源场景[3]、法律领域的专项微调以提升领域任务性能[4]、指令微调与思维链增强以实现多任务统一建模[5]，以及自监督学习技术用于深层语义挖掘[6]。国外则在方法论与跨模态扩展方面取得进展，如 LMDX 框架在视觉丰富文档(VRDU)上的实体抽取与定位[7]、Text2DB 任务与 OPAL 框架对 IE-数据库集成范式的重塑[8]、Prompt Chaining 在低资源任务中的分解策略[9]、以及通过模板多样化提升零样本鲁棒性的 K2Q 数据集构建[10]。然而，上述工作多集中于提升抽取效率或跨模态能力，针对基因编辑技术语料中的信息抽取偏见(即模型对不同类别要素的系统性能力差异)尚缺乏专门评估。

针对这一研究空白，本研究提出一种面向基因编辑领域的 LLM 偏见评估方法。通过构建“Basic-Rethink-Multi-Query”级联评估流程，聚焦基因编辑事件六类核心要素(人物、组织、技术、对象、效果、发表期刊)，系统评估不同模型与提示策略下的信息抽取能力与偏差表现。本研究旨在识别与评估 LLM 在基因编辑技术领域的认知偏见，并为其在前沿生物技术领域的信息处理、风险预警与治理策略提供实证依据与方法支持。

2. 材料与方法

为系统性评估大语言模型在基因编辑技术领域的信息抽取能力与认知偏见，本研究构建了一个基于事件要素的信息抽取分析框架(图 1)。该框架基于经过大模型辅助标注与人工校对的数据集，采用“Basic-Rethink-Multi-Query”级联评估流程：首先，在四种提示策略下对六类关键事件要素进行初步抽取；继而，通过 Rethink 模块与基于检索增强生成(RAG, Retrieval-Augmented Generation)的 Multi-Query 模块对结果进行迭代优化；最终，采用精确度(P)、召回率(R)、F1 值(F1)及变异系数(CV)等指标，从准确性、鲁棒性和一致性方面对输出结果进行全面评估。

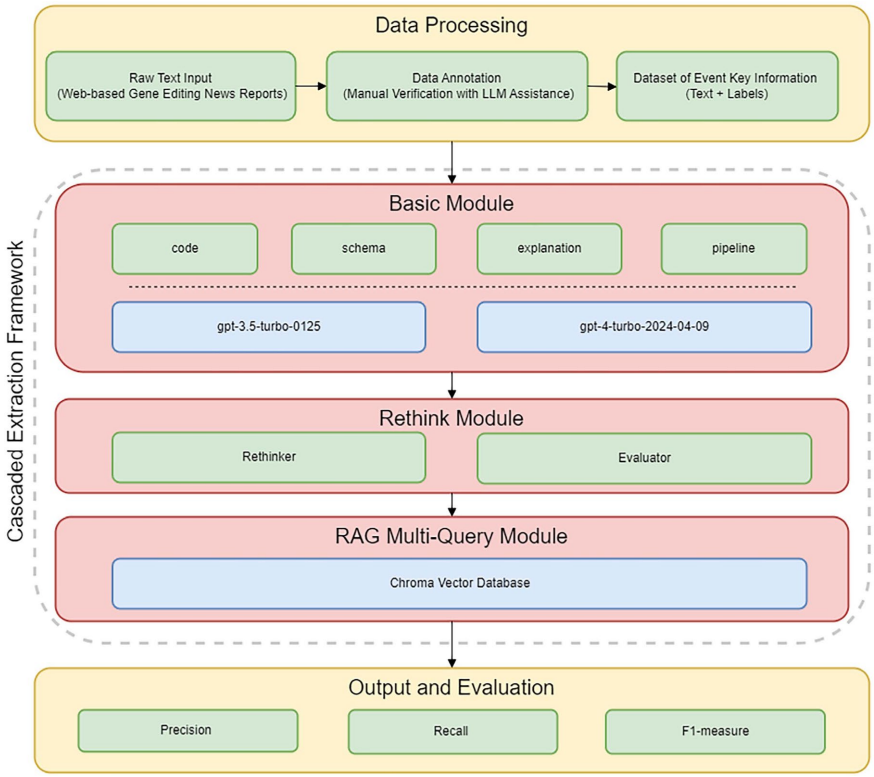

Figure 1. Information extraction framework
图 1. 信息抽取分析框架

2.1. 数据集

2.1.1. 基因编辑事件定义

本文基于事件对基因编辑技术文本进行处理。事件指文本中由特定触发词标识的具体事实或状态变化(如“发布”或“突破”)，其核心由事件类型及若干论元(即构成该事件的参与者与细节要素，如人物、时间、地点等，承担特定角色)构成。与技术本身相比，事件蕴含复杂时空情境和丰富社会属性，能更全面地反映技术应用的实际情况。例如，“基因编辑”是一项技术，而“基因编辑水稻”则是一个事件。本文将基因编辑事件定义为包含以下六个核心要素的结构化单元：人物(Person)、组织(Organization)、技术(Technology)、对象(Object)、效果(Effect)及发表期刊(Publish)，其中各要素名称即为基因编辑事件中的论元。

2.1.2. 数据来源

基因编辑技术领域缺乏研究所需的“文本信息抽取”标签数据。本研究以“中国科技网(stdaily.com)”热点板块中检索“基因编辑”所得文档作为主要实验数据来源，该网站内容专业且可读性高，符合事件关键信息抽取数据集的构建需求。经关键词筛选、时间范围限定及人工质量评估，最终获得近年来“生物医学–基因编辑”相关文档百余篇。

2.1.3. 事件关键信息数据标注

对于“关键信息抽取”数据的标注，在每一篇基因编辑技术报道文本中设定 6 类关键信息标签，标签含义如表 1 所示。采用大语言模型辅助自动标注，并经人工校对修正，最终对 107 篇文档逐条完成 6 类标签的标注工作。

Table 1. Key information tags and descriptions
表 1. 关键信息标签及含义

tag	标签含义	示例
person	参与实验的相关人物名字	赵开军
organization	参与实验的相关组织团队的名称	中国农业科学院作物科学研究所水稻抗病基因挖掘与利用创新组
technology	基因编辑技术的全名	CRISPR/Cas9 介导的同源重组技术
object	技术的实验对象	感病品种水稻
effect	实验的效果	成功将感病品种水稻转变成广谱高抗白叶枯病的水稻品种，创制了广谱高抗白叶枯病水稻培育的新途径，为有效利用重要的基因核心元件提供了模型。
publish	发表的期刊名称	《分子植物(Molecular Plant)》

2.2. 级联评估流程

2.2.1. 大语言模型选择

当前主流大语言模型普遍建立在 Transformer 架构基础之上。该架构由谷歌于 2017 年提出[11]，其核心创新在于采用 Self-Attention 机制，完全摒弃了循环网络结构，使模型能够在每个位置同时关注输入序列中的所有其他位置，从而有效捕捉全局上下文信息，并显著提升长序列建模的效率。Transformer 凭借其高度并行化的计算模式与多层堆叠设计，成为 BERT、GPT 等代表性模型的实现基础。

近年来，ChatGPT 等先进模型进一步在 Transformer 基础上引入基于人类反馈的强化学习(RLHF)[12]，通过监督微调、奖励模型训练和策略优化算法，使生成文本更贴合人类偏好与实用需求，推动了自然语言生成技术在效果与适用性方面的显著进步。

本研究在关键信息抽取部分选择了 gpt-3.5-turbo-0125 [12]和 gpt-4-turbo-2024-04-09 [13]两个模型，

Temperature 值设置为 0，使输出更确定，更适合结构化信息抽取。模型参数如表 2 所示。

Table 2. Models and parameter settings
表 2. 模型及其参数

模型名称	Temperature 参数设置	上下文长度
gpt-3.5-turbo-0125	0	16 k
gpt-4-turbo-2024-04-09	0	128 k

2.2.2. 提示学习方法

(1) 零样本提示学习[14]：一种让 LLM 在没有任务特定训练数据的情况下执行任务的技术。通过设计提示(如模板、描述等)引导模型生成特定输出，适用于文本分类、生成等任务，优势在于泛化能力和语义理解，无需大量标注数据。

(2) 少样本提示学习[15]：在少量标注数据的支持下提升复杂任务表现，通过提供少量示例增强模型的推理能力。其核心是设计简洁提示，帮助模型学习任务模式，减少对大量标注数据的依赖，降低数据成本，提升模型适用性。

2.2.3. 提示词模板

本研究参考 Sun *et al.* (2024) [16]利用 ChatGPT 进行药物警戒事件的提取研究，构建了四种提示词模板(code、schema、explanation、pipeline)。code 通过将描述文本与代码片段组合来制定抽取指令和输出格式；schema 提供抽取指令并枚举事件论元类型、指定事件类型；explanation 基于 schema 对各事件论元类型和事件类型进行详细解释；pipeline 以流水线的方式对每一个事件论元参数进行相关的问题查询。各个模板结构特点及对比情况如表 3。

Table 3. Comparison of the four prompt templates
表 3. 四种提示词(Prompt)对比

模板	code	schema	explanation	pipeline
结构	Python 代码结构，提示语融入代码中	描述事件类型和参数类型的 JSON 结构	Schema + 自然语言解释	问答指令 + 问题列表
优点	代码的逻辑结构	清晰，适合模型生成结构化输出	详细，适合复杂场景	精确，适合问答形式抽取
适用场景	代码解析能力强的模型	模型生成结构化输出的场景	复杂事件抽取任务	精确提取每个参数信息的场景

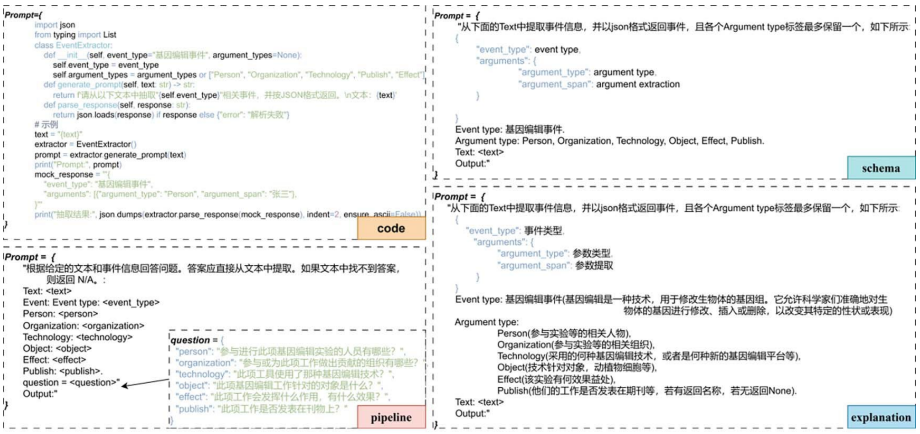

Figure 2. Example prompts for each template
图 2. 各模板示例

构建多种模板旨在从不同角度评估模型效果，测试其在结构化输出、参数理解和推理能力上的表现，进而找到最佳抽取方法，增强信息准确性与适用性。各模板样例如图 2 所示。

2.2.4. 反思(Rethink)模块

为修正基础信息抽取模块可能存在的错误或遗漏，本研究引入了一个反思(Rethink)模块。该模块基于提示工程方法，通过设计针对性反思提示词(如图 3)，引导大语言模型对基础模块的抽取结果进行自我评估与修正[17]。

具体流程如下：首先，将基础模块中性能最优的 explanation 提示模板所得结果作为反思对象的初始输入。其次，在反思提示中增设限制性条件(如：若文本中无相应信息则返回“None”；技术名称须完整抽取)，以约束生成内容、减少虚构；最终，形成“初始输入－反思提示－反思结果”的三元组输出。

该模块由同一语言模型担任反思器(Rethinker)与评估器(Evaluator)，对初始抽取信息的适当性进行评估与反思，最终得到反思结果。

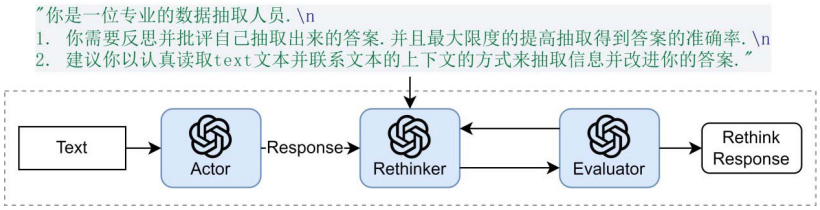

Figure 3. Structure of the rethink module
图 3. Rethink 模块结构图

2.2.5. RAG 多问答(Multi-Query)模块

为克服反思模块在基因编辑技术等专业领域存在的知识局限性，本研究引入检索增强生成(RAG)模型[18]。该技术已成功应用于生物医学文本处理任务，如从社交媒体中提取药物不良事件[19]，证明了其在不依赖大规模训练数据的情况下提升专业领域任务表现的潜力。本研究在此基础上，进一步采用多问答(Multi Query)策略以提升信息抽取的准确性与泛化能力。本研究基于 Chroma 向量数据库构建领域知识库。通过爬取 PubMed 等专业文献源的文本，经切分与编码后存储，形成面向基因编辑技术领域的专用外部知识源。

Multi-Query 方法的实现流程(图 4)如下：首先，将反思模块的输出结果构建为初始查询；随后，利用大语言模型生成多个语义一致但表述多样的补充查询问题(如针对同一技术概念的不同术语形式)；最后，用所有查询语句分别检索知识库，并将所得相关文档整合后输入生成模型，以产生最终答案。

该设计主要基于两方面原因：一是可覆盖专业术语的不同表述变体，减少遗漏；二是可降低因单一查询表述偏差或初始抽取错误导致的检索失效风险，提高检索结果的相关性与鲁棒性。

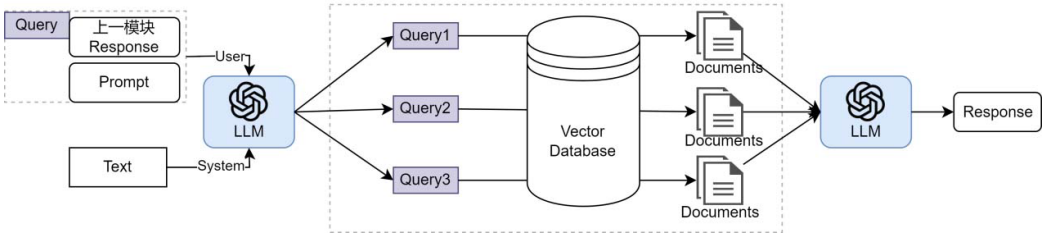

Figure 4. Structure of the Multi-Query Module
图 4. Multi-Query 模块结构图

2.3. 评估指标

本研究的评估体系包含性能与偏见两个层面。在性能评估上，采用精确率(Precision, P)、召回率(Recall, R)与 F1 值(F1-measure, F1)作为核心指标[20]。为解决诸如“自然”与“《自然》”等表面形式不一致而导致的严格字符串匹配失败问题，本文采用 BERTScore [21]评估模型抽取标签与标准标签之间的语义相似度(取值范围 0~1)。当两者的 BERTScore 不低于阈值 0.8 时，视为匹配；基于该匹配规则计算 P、R、F1，计算公式见式(1)~(3)。F1 值越高表示抽取效果越好。阈值 0.8 经实验比对确定，以在保证信息完整性与正确性的前提下取得较好平衡。

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2PR}{P + R} \tag{3}$$

其中，TP (True Positive)表示将正类抽取为正类的数量，指的是模型在信息抽取任务中，对输入文本对应的标签内容抽取正确的实例数量。相应地，FP (False Positive)为将负类抽取为正类数，FN (False Negative)为将正类抽取为负类数。

在偏见评估上，通过计算六类标签 F1 值的变异系数(Coefficient of Variation, CV)来衡量模型表现的离散程度，其计算公式见式(4)。

$$CV = \frac{\sigma}{\mu} \times 100\% \tag{4}$$

CV 值越低，表明模型在不同类别上的性能越稳定，即认知偏见越小。该无量纲指标消除了均值量纲的影响，使得在不同平均性能水平(例如某模型平均 F1 为 50%，另一为 70%)之间可直接比较离散程度。

3. 实验结果与分析

本研究旨在评估 LLM 在基因编辑技术领域信息抽取任务中的表现，着重比较不同提示词模板下的性能差异，并评估引入反思(Rethink)与 RAG Multi-Query 机制后对抽取精度及潜在认知偏见的缓解效果。实验在 zero-shot (零样本)设定下展开，基于“Basic-Rethink-Multi-Query”级联评估流程，逐步量化 Basic、Rethink 与 Multi-Query 各模块的贡献与局限。评估指标包括要素级精确率、召回率与 F1 值，及用于衡量类别间不均衡性的变异系数(CV)。

3.1. 基础(Basic)模块评估

本部分旨在评估不同提示模板在基因编辑语料上对六类信息标签(person、organization、technology、object、effect、publish)的抽取能力。实验采用了四类提示模板(code、schema、explanation、pipeline)，在 gpt-3.5-turbo-0125 和 gpt-4-turbo-2024-04-09 模型上分别进行测试，结果如图 5、图 6 所示。

实验数据显示(图 5)，模型在信息抽取任务上存在明显的类别偏好。无论是 gpt-3.5-turbo 还是 gpt-4-turbo，对“技术、组织、发表期刊”等具体、常见类别抽取的 F1 值显著高于对“人物、对象、效果”等类别的表现。例如，在使用解释型提示(explanation prompt)时，gpt-4-turbo 对“组织”(91.59%)、“技术”(80.37%)和“发表期刊”(79.44%)的识别效果远优于“人物”(59.81%)、“对象”(40.19%)和“效果”(47.66%)。这种差异在不同提示词模板和不同模型之间保持一致。值得注意的是，相较于 gpt-4-turbo 或使用 explanation 提示模板，gpt-3.5-turbo 在 code、schema 与 pipeline 三类提示下对“组织”和“发表期刊”的

F1 值明显偏低。该现象既可能反映模型能力的固有差距，也可能由 gpt-3.5-turbo 对提示词模板的敏感性引起。

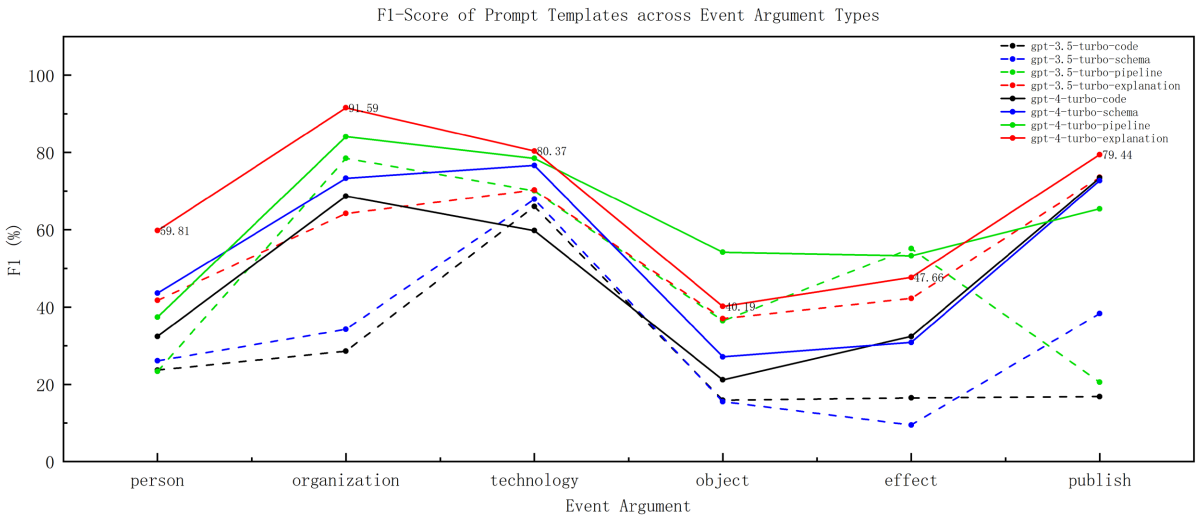

Figure 5. Extraction performance (F1, %) of the four prompt templates across event arguments
图 5. 四种提示词模板在各事件论元上的抽取性能(F1, %)

总体而言，gpt-4-turbo 的性能显著优于 gpt-3.5-turbo，尤其在“组织”和“发表期刊”等类别上优势明显。然而，尽管绝对性能提升，不同实体类别间的相对性能差距依然存在。导致这一差异的主要原因可归纳如下：

(1) Intrinsic Bias：大型语言模型在海量通用语料上预训练，导致其知识表征存在固有偏差。诸如“技术”、“组织”和“发表期刊”等术语在训练语料中出现频率高、表述规范，模型对其建模充分；相比之下，“人物”、“对象”(尤其是特定的生物模型)和“效果”等信息在语料中分布稀疏且表述多样，模型表征相对薄弱。因此，模型本能地倾向于识别和生成高频、规范的类别，而对低频、长尾信息的识别能力不足。

(2) Extrinsic Bias：本研究的零样本信息抽取任务主要依赖模型的提示匹配与浅层推理能力。在此设定下，“技术”“组织”“发表期刊”通常以显式的专有名词形式出现(如“CRISPR-Cas9”“哈佛大学”“《科学》”)，与提示词能形成直接、精确的表层匹配。反之，“效果”需要模型理解因果描述并进行推断；“人物”和“对象”则常以代词、职位或缩写等隐晦形式表达，对模型的深层语义理解与跨句推理能力提出了更高要求，从而导致识别性能差异。

(3) Inherent Difficulty Disparity among Classes：即使不考虑模型与任务因素，不同实体类别本身在语言表述上就存在固有难度梯度。“技术”“组织”等类别对应规范的命名实体，边界清晰；而“效果”作为抽象概念，其表述高度依赖上下文；“对象”的命名方式则极具多样性。这种固有的表述复杂性，在零样本条件下被放大，使得模型更易捕获显式信息，而难以处理隐晦和多样的表述。

上述结果揭示了大语言模型在基因编辑技术领域存在一种“重静态实体、轻动态要素”的固有偏见：模型更倾向于关注事件中的静态实体(如 CRISPR-Cas9)，而相对忽视了基因编辑事件的执行者、被编辑对象、实验效果等动态要素。

图 6 结果显示，explanation 模板显著提高信息抽取的准确性和稳定性。采用提供详细注释和上下文的“explanation”提示模板后，模型对基因编辑事件的整体识别效果显著提升(gpt-4-turbo 和 gpt-3.5-turbo

的 F1 分别达到 66.51%和 54.85%)，同时有效降低了输出结果的认知偏见(CV 值分别降至 30.68%和 29.70%)。这表明，通过增强提示的语义信息，能够改善模型对专业术语和复杂关系的理解能力，从而不仅提高抽取准确率，也在一定程度上缓解了模型的结构性偏差。

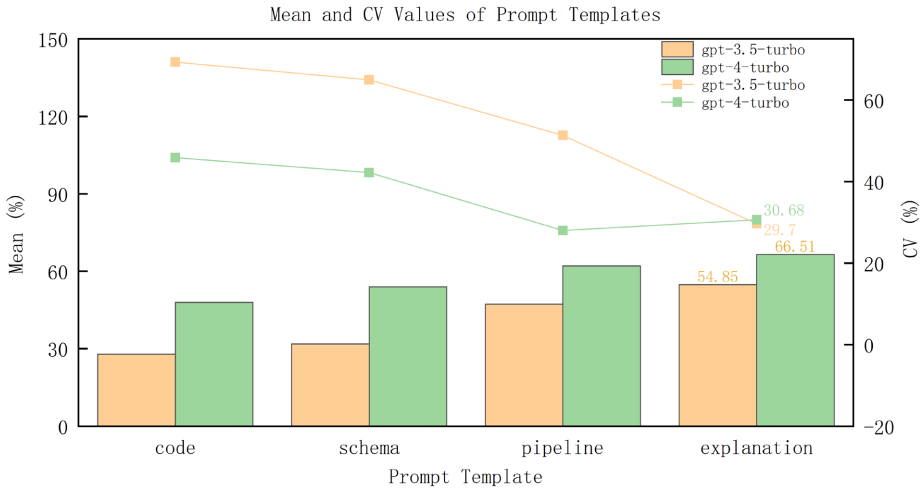

Figure 6. Macro-F1 score and Coefficient of Variation (CV) of four prompt templates for various event arguments
图 6. 四种提示词模板在各事件论元上的宏平均 F1 分数与性能变异系数(CV)

基于以上分析，本研究选定“explanation”提示模板作为后续级联评估流程的基准策略，旨在通过语义提示优化，全面提升信息抽取的全面性、准确性与公平性。

3.2. 反思(Rethink)及多问答(Multi-Query)模块评估

为缓解基础提示可能导致的遗漏与低置信度输出，本研究引入了两类改进机制：一是 Rethink (反思)，促使模型回顾并补充初步抽取结果；二是基于检索增强生成(RAG)的 Multi-Query (多问答)，通过从 PubMed 等权威来源构建的专用 Chroma 知识库进行多轮检索，以覆盖术语变体并补充上下文信息。图 7 与图 8 分别展示了引入上述机制前后，各标签的抽取性能及整体稳定性变化。

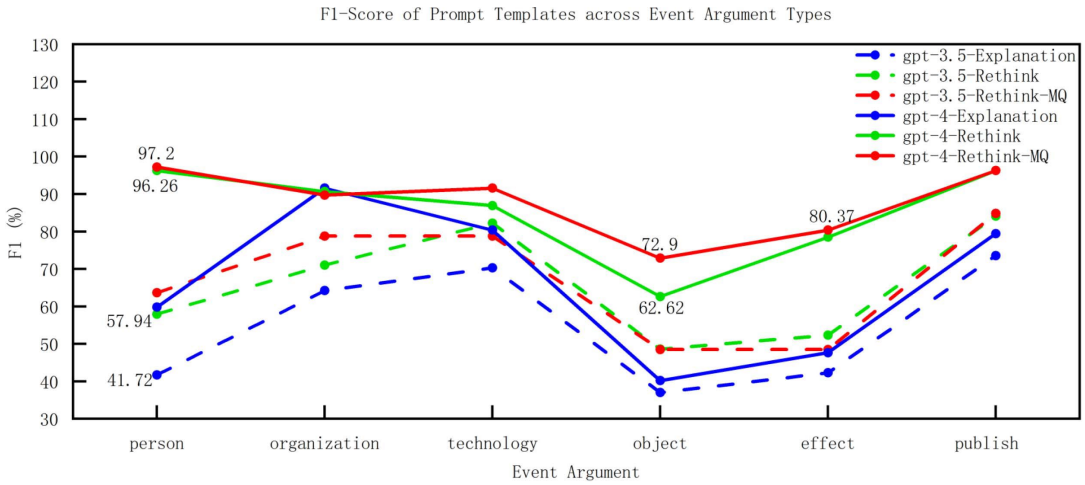

Figure 7. Extraction performance (F1, %) of different extraction mechanisms across event arguments
图 7. 不同抽取机制在各事件论元上的抽取性能(F1, %)

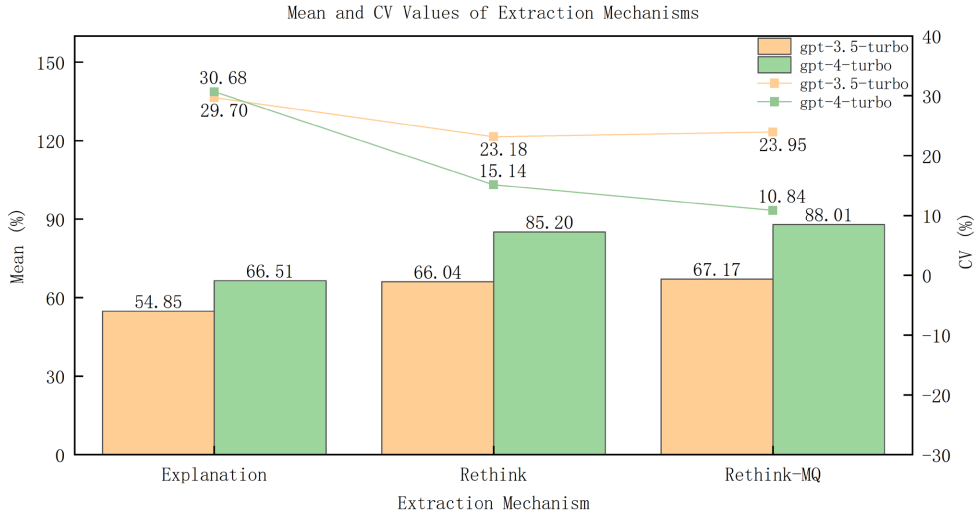

Figure 8. Macro-F1 score and coefficient of variation (CV) across event arguments for different extraction mechanisms
图 8. 不同抽取机制在各事件论元上的宏平均 F1 分数与性能变异系数(CV)

主要发现如下：

(1) 整体性能显著提升：在加入 Rethink 或多查询模块(Rethink-MQ)后，所有类别的 F1 值普遍上升且类别间差距缩小(图 7)，显示改进机制对缓解基因编辑事件要素的识别不均衡具有明显效果。总体而言，gpt-3.5-turbo 与 gpt-4-turbo 的总 F1 分别从基础版本的 54.84%、66.51%提升至 Rethink-MQ 下的 67.17% (提升 22.5%)与 88.01% (提升 32.3%)，CV 值分别从 29.70%、30.68%降至 23.95% (下降 19.4%)和 10.84% (下降 64.7%) (图 8)。值得注意的是，Rethink 模块在 gpt-4-turbo 的 organization 标签上略有下降，表明其总体有效，但过度反思可能适得其反；Multi-Query 模块在模型中个别标签(如 gpt-3.5-turbo 的 technology 与 effect、gpt-4-turbo 的 organization)比 Rethink 模块效果略有降低，但整体表现最佳，个别标签可能因 LLM 检索歧义受影响，但总体优势明显。

(2) 反思对薄弱类别的识别增强：反思机制对原本性能较弱的类别(如“人物”)提升尤为显著。以 gpt-4-turbo 为例，经 Rethink 处理后“人物”F1 值提升至 96.26%，几乎消除了与“技术”等类别之间的差距；gpt-3.5-turbo 的“人物”F1 值也从 41.72%提升至 57.94% (图 7)。说明反思机制能够明显强化模型原本识别能力较弱的要素类别，有效补偿基础模型中的认知偏差。

(3) 多查询检索的补充作用：RAGMulti-Query 通过多轮检索扩展术语覆盖与上下文来源，主要解决术语变体与上下文缺失问题，从而提升识别一致性与稳定性。例如在 gpt-4-turbo 中，加入 Multi-Query 后“技术”“对象”“效果”类别的 F1 上升，其中“对象”类别由 Rethink 的 62.62%提升至 Multi-Query 的 72.90% (图 7)。该策略对处理如“CRISPR-Cas9”/“Cas9”等变体尤为有效。

(4) 残留偏差问题：尽管多数类别显著改观，但偏见未被完全消除。gpt-4-turbo 在最优 RAGMulti-Query 设置下，“人物”类别 F1 已提升至 97.20%，而“对象”“效果”仍相对滞后，分别约 72.90%和 80.37% (图 7)。gpt-3.5-turbo 在部分类别上的提升有限，反映出该模型在知识覆盖与泛化能力上的局限。

综上所述，在引入 Rethink 机制与 Multi-Query 策略后，模型通过链式推理与外部知识补充不仅显著提升了整体性能，还调整了原有输出偏好，明显提升了对“人物”“对象”等先前薄弱要素的识别能力，部分缓解了“重静态实体，轻动态要素”的结构性倾向。但仍有若干语义盲区未被完全覆盖，表明仅靠提示优化与检索增强无法彻底根除所有偏差，后续需结合更细粒度的数据增强、领域微调及更丰富的知识源以进一步提升对复杂动态要素的抽取性能。

4. 结论与展望

本研究系统揭示了 LLM 在基因编辑技术领域信息抽取任务中存在的结构性认知偏差，即模型普遍表现出“重静态实体、轻动态要素”的倾向——其对“技术、组织、发表期刊”等静态实体的识别能力显著优于“人物、对象、效果”等动态要素。这种偏差既源于预训练语料的内在分布不均衡，例如某些动态要素相关语料在预训练集中占比极低，导致模型学习不足；也受到提示词构建与抽取机制等外在因素的制约，比如提示词设计未能充分涵盖动态要素的多样表达形式。同时，实验过程中也发现了一些异常数据点，这可能与该类文献中对这些内容表述的复杂性和专业性更强有关。从应用角度看，此类偏差也可能因忽视研究者、实验对象及伦理信息而影响生物安全评估与治理决策的完整性。

针对上述问题，本文创新性地提出了“Basic-Rethink-Multi-Query”级联评估流程。实验结果表明，该机制不仅能显著提升信息抽取的整体效能，还可有效增强对“人物、对象、效果”等薄弱要素的识别能力，改善多类别间的均衡性，从而在相当程度上修正模型的初始输出偏差，提升输出结果的全面性与可用性。

总体而言，本研究不仅为理解 LLM 在专业领域中的局限性提供了重要的理论与实证依据，也为应对领域自适应挑战提供了一条低成本、高效率的技术路径。所提出的级联评估流程具备良好的可迁移性，为构建更可靠、全面的生物医学信息处理系统提供了有效支持，也为后续实现系统化的生物技术安全治理与伦理审查提供了实践参考。

尽管取得积极进展，研究仍存在若干限制：其一，部分动态事件要素(尤其“对象”)的识别性能仍需提升；其二，实验基于有限规模的新闻语料，缺乏更大规模、多语种及多来源的验证；其三，未能涵盖全部模型架构与推理范式。为此，建议未来工作重点如下：

1) 深入剖析残留偏差成因并量化其影响；
2) 系统评估不同类型模型(闭源/开源)在性能与成本间的折中；
3) 通过领域微调、跨语料迁移与检索增强(RAG)强化知识融合；
4) 将偏见评估纳入生物安全与伦理预警流程，开发定量与定性相结合的治理工具；
5) 在更多文献类型与来源上复现实验并推动评测基准的标准化。

基金项目

国家社会科学基金资助项目(25CKX003)。

参考文献

[1] Dhaini, M., Poelman, W. and Erdogan, E. (2023) Detecting ChatGPT: A Survey of the State of Detecting ChatGPT-Generated Text. *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, Shoumen, 1-12.

[2] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., *et al.* (2024) Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, **50**, 1097-1179. https://doi.org/10.1162/coli_a_00524

[3] 时宗彬, 朱丽雅, 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, 2024, 8(7): 23-31.

[4] 沈晨晨, 岳圣斌, 刘书隽, 等. 面向法律领域的大模型微调与应用[J]. 大数据, 2024, 10(5): 12-27.

[5] 赵勤博, 王又辰, 陈荣, 等. 面向开源情报的信息抽取大语言模型[J]. 计算机工程与设计, 2024, 45(12): 3772-3778.

[6] 孙亚伟. 基于多维度语义挖掘的情绪信息抽取技术研究[D]: [博士学位论文]. 北京: 北京邮电大学, 2024.

[7] Perot, V., Kang, K., Luisier, F., Su, G., Sun, X., Boppana, R.S., *et al.* (2024) LMDX: Language Model-Based Document Information Extraction and Localization. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 15140-15168. https://doi.org/10.18653/v1/2024.findings-acl.899

[8] Jiao, Y., Li, S., Zhou, S., Ji, H. and Han, J. (2024) Text2DB: Integration-Aware Information Extraction with Large Language Model Agents. *Findings of the Association for Computational Linguistics ACL* 2024, Bangkok, 11-16 August 2024, 185-205. https://doi.org/10.18653/v1/2024.findings-acl.12

[9] Kwak, A., Morrison, C., Bambauer, D. and Surdeanu, M. (2024) Classify First, and Then Extract: Prompt Chaining Technique for Information Extraction. *Proceedings of the Natural Legal Language Processing Workshop* 2024, Miami, 16 November 2024, 303-317. https://doi.org/10.18653/v1/2024.nllp-1.25

[10] Zmigrod, R., Shetty, P., Sibue, M., Ma, Z., Nourbakhsh, A., Liu, X., *et al.* (2024) "What Is the Value of Templates?" Rethinking Document Information Extraction Datasets for LLMs. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, 12-16 November 2024, 13162-13185. https://doi.org/10.18653/v1/2024.findings-emnlp.770

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 6000-6010.

[12] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., *et al.* (2022) Training Language Models to Follow Instructions with Human Feedback. *Proceedings of the* 36*th International Conference on Neural Information Processing Systems*, New Orleans, 27730-27744.

[13] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., *et al.* (2023) GPT-4 Technical Report. arXiv: 2303.08774.

[14] Wang, W., Zheng, V.W., Yu, H. and Miao, C. (2019) A Survey of Zero-Shot Learning. *ACM Transactions on Intelligent Systems and Technology*, **10**, 1-37. https://doi.org/10.1145/3293318

[15] Song, Y., Wang, T., Cai, P., Mondal, S.K. and Sahoo, J.P. (2023) A Comprehensive Survey of Few-Shot Learning: Evolution, Applications, Challenges, and Opportunities. *ACM Computing Surveys*, **55**, 1-40. https://doi.org/10.1145/3582688

[16] Sun, Z., Pergola, G., Wallace, B. and He, Y. (2024) Leveraging ChatGPT in Pharmacovigilance Event Extraction: An Empirical Study. *Proceedings of the* 18*th Conference of the European Chapter of the Association for Computational Linguistics* (*Volume* 2: *Short Papers*), St. Julian's, 17-22 March 2024, 344-357. https://doi.org/10.18653/v1/2024.eacl-short.30

[17] Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E. and Fung, P. (2023) Towards Mitigating LLM Hallucination via Self Reflection. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, 6-10 December 2023, 1827-1843. https://doi.org/10.18653/v1/2023.findings-emnlp.123

[18] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, Vancouver, Article 793.

[19] Berkowitz, J., Srinivasan, A., Cortina, J. and Tatonetti1, N. (2024) TLab at #SMM4H 2024: Retrieval-Augmented Generation for ADE Extraction and Normalization. *Proceedings of the* 9*th Social Media Mining for Health Research and Applications* (*SMM4H* 2024) Workshop and Shared Tasks, Bangkok, 15 August 2024, 153-157. https://doi.org/10.18653/v1/2024.smm4h-1.36

[20] Yacouby, R. and Axman, D. (2020) Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Online, 20 November 2020, 79-91. https://doi.org/10.18653/v1/2020.eval4nlp-1.9

[21] Zhang, T.Y., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y. (2020) BERTScore: Evaluating Text Generation with BERT. https://doi.org/10.48550/arXiv.1904.09675