AutoRG：一种大小模型协同的自动报告生成框架

张京[1,2], 舒江明[1,2], 张宇翔[1,2], 吴斌[3], 王巍[3], 于剑[1,2], 桑基韬[1,2]*
[1]北京交通大学计算机科学与技术学院
[2]交通数据分析与挖掘北京市重点实验室
[3]深圳素问数据智能科技有限公司
{zhangj_, 20281016, yuxiangzhang, jianyu, jtsang}@bjtu.edu.cn
{wubin, wangwei}@suwen.ai

摘要

自动报告生成技术在提高工作效率和节约人力资源方面具有显著潜力。大语言模型的出现使得报告流畅度与可解释性得到提升。然而，现有工作仍依赖人工，缺乏灵活性和丰富度。同时，小模型错误或冗余的输出与大模型自身的随机性会导致报告质量不稳定。本文提出大小模型协同的自动报告生成框架AutoRG，通过大模型的工具理解与规划能力减少人工干预，提升报告丰富度，并通过信息修正与报告迭代机制提高报告的稳定性。本文以自动专利报告生成为场景，从多个维度对AutoRG进行全面测试。结果表明，该框架在提高报告生成的丰富度和质量稳定性方面具有显著优势。

关键词： 自动报告生成； 大语言模型； 智能体； 大小模型协同

AutoRG: An automatic report generation framework for Large and small model collaboration

Jing Zhang[1,2], Jiangming Shu[1,2], Yuxiang Zhang[1,2], Bin Wu[3], Wei Wang[3], Jian Yu[1,2], Jitao Sang[1,2]
[1]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
[2]Lab of Traffic Dala Analysis and Mining, Beijing, China
[3]Suwen Data Intelligent Technology Co., LTD., Shenzhen, China
{zhangj_, 20281016, yuxiangzhang, jianyu, jtsang}@bjtu.edu.cn
{wubin, wangwei}@suwen.ai

Abstract

Automatic report generation technology has demonstrated enormous potential in saving human resources and enhancing work efficiency. With the introduction of large-scale artificial intelligence models, the fluency and interpretability of generated reports have been significantly improved. However, current automatic report generation still requires manual completion, lacking flexibility and richness. Moreover, errors or redundancies in the outputs of small models, along with the inherent randomness of large models, can result in the lack of stability in report content. This paper introduces AutoRG, an integrated framework designed for automated report generatio, to reduce human intervention by leveraging the understanding and planning capabilities of large models, consequently enriching the depth of generated reports. Additionally, it incorporates an information correction and content iteration mechanism to diminish reliance on the outputs of small models and mitigate the impact of randomness inherent in large models. To evaluate the efficacy of the AutoRG framework comprehensively,

*通讯作者

基金项目：中央高校基本科研业务费专项资金(62172094)；国家自然科学基金(62172094)

we conducted experiments focusing on the automated generation of patent technology reports. Our evaluation approach across various dimensions, ensuring a thorough examination of AutoRG's performance. The results indicate significant advantages of the framework in improving the richness and stability of report generation.

Keywords: Automatic report generation , Large language model , Agent , Collaboration between large and small models

1 引言

自动报告生成是通过计算机算法自动提取原始数据或文本中的关键信息，按照既定格式和结构生成具有特定目的和内容的报告的过程。该技术已广泛应用于多个领域，例如医学影像报告(Nakaura et al., 2024)、社会金融报告(Ren et al., 2021)、企业财务报告(**?**)、信息检索报告(Babour and Khan, 2021)等。自动报告生成的优势在于，它能够减轻人们在报告编写任务上的负担，节省阅读文献、分析数据和重复文本编写工作所需的大量时间和精力。此外，自动生成的报告具有一定的参考价值，能够为行业专家提供撰写报告的参考，甚至作为内容补充。

传统自动报告生成方法主要分为抽取式和生成式两大类。抽取式方法主要通过自然语言处理的小模型，通过规则筛选(Gong et al., 2017)、主题聚类(Wang et al., 2019)、数据统计(Babour and Khan, 2021)等技术手段直接从原始数据中提取相关信息并填充到预设的模板中。这种方法的优势在于能够迅速获取基础信息，但其灵活性较差，难以适应多变的报告需求，同时也缺乏对内容进行深入分析的能力。另一方面，生成式方法通过端到端训练的模式，如编码器-解码器模型，试图模仿人类编写报告的过程。这类方法通过学习原始报告中的潜在变量分布(Li et al., 2023; Tsaniya et al., 2024)，并结合知识图谱等先验知识(Li et al., 2019; Li et al., 2023)，从而生成更加自然和连贯的报告文本。然而，尽管生成式方法在模仿人类写作风格方面取得了一定进展，但它们在生成高价值报告方面仍面临挑战。这是因为，一篇高价值的报告不仅需要信息的准确性和内容的丰富性，更关键的是要具备高度的可解释性和透明的推理过程。报告应当能够清晰地阐述其分析结果、结论和建议的依据，使得读者充分理解报告的核心观点和逻辑基础，从而建立起对报告内容的信任。端到端训练的方法虽然能够捕捉到一定程度的语言规律，但在确保报告的可解释性和推理透明度方面仍有所不足。

随着大型语言模型的出现，大模型凭借其强大的语义理解和生成能力，有望在保持报告内容自然流畅的同时，提供更加详尽的解释和透明的推理路径，生成易于理解和信任的报告。一些工作探究了大模型在自动报告生成上的性能。面对数据量较小的场景下，Ding(2024)直接将有限的原始文本输入到大型模型中，直接生成报告。这种方法简单直接，但真实报告场景下往往需要处理大量文档。这时则倾向于人为规定某些小型模型用于数据处理，如信息提取(Wang et al., 2019)、主题聚类(Wang et al., 2019)等，大模型则根据处理后的数据进行指定内容的推理和生成。这种人为设计的半模板式大模型报告生成方法利用小模型在特定任务上的专业性和高效性，解决了大模型的输入长度问题，但同时也存在一些局限性。首先，这些方法并没有充分发挥大型模型在工具使用和自主规划方面的能力，在生成报告时依然需要依赖人工设定报告模块和对应的小模型类型，这限制了报告内容的全面性和多样性。其次，这种输入提示并直接输出报告的工作流程会由于小模型的错误或冗余的输出和大模型自身生成过程的随机性而不稳定，导致部分报告的质量有所下降。

基于上述存在的两个问题，本文提出了一种大小模型协同的自动报告生成框架AutoRG，旨在通过充分发挥大型语言模型与小型专业模型的互补优势，在确保报告质量的前提下，将人从报告生成的过程中解放（见图1）。AutoRG框架将报告生成任务划分为"大纲规划与评估"、"工具调用与内容修正"和"报告生成与优化"三个阶段。"大纲规划与评估"阶段旨在通过大模型的规划能力，定制化生成和筛选报告所需的子模块，确保每个子模块都能有合适的小模型辅助完成；"工具调用与内容修正"阶段利用大模型的工具理解能力选择性挑选小模型，结合优秀的摘要生成能力对小模型输出进行修正，得到价值密度较高的分析数据；而"报告生成与优化"阶段旨在通过多智能体协作工作流的方式，迭代生成精确、可解释性强且具有深度的报告内容。总结来说，本文的主要贡献可以概括为：

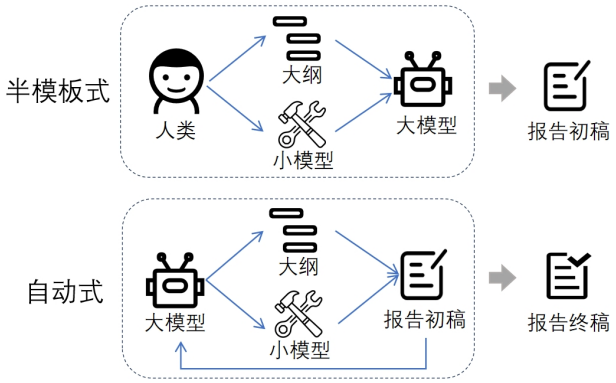

Figure 1: 半模板式与自动式大模型报告生成框架的对比

（1）本文提出了一个大小模型协同的自动报告生成框架AutoRG，有效利用大模型的工具理解和自主规划能力，自动化生成更灵活、全面的报告，无需过多的人工干预。

（2）本文通过引入信息修正和报告迭代机制，降低对小模型输出和大模型随机性的依赖，提高了报告质量的稳定性。

（3）本文以自动专利技术报告生成为例，采用主客观评估结合的方式，从多个维度验证了框架的有效性。

2 相关工作

2.1 大语言模型

大语言模型（Large Language Model, LLM）是一类在超大规模数据集上进行预训练的大型预训练语言模型，旨在有效地理解和生成人类语言。与小型语言模型相比，大模型因其庞大的参数量和数据训练量，展现出了卓越的涌现能力，特别是在指令跟随和多步推理方面(Zhao et al., 2023)。指令跟随能力允许大模型根据给定的自然语言指令执行任务，即使在缺少具体示例的情况下也能表现出良好的适应性和灵活性。这种能力使得大模型能够在无需额外训练的情况下，理解和执行多样化的任务，如文本生成(Lu et al., 2023)、问题回答(Shi et al., 2023)和代码生成(Chen et al., 2021)等。多步推理能力则使得LLM能够处理涉及多个推理步骤的复杂问题。在适当的提示策略下，大模型能够模拟人类的推理过程，逐步解决如数学问题(Azerbayev et al., 2023)、逻辑谜题(Dua et al., 2022)等任务，显示出对深层次语义和逻辑关系的理解。这些能力使得大模型能够在缺乏直接示例的情况下，根据新任务的指令进行有效的执行，并在处理复杂问题时展现出卓越的性能。

智能体是人工智能领域中的一个核心概念，指能够感知自身所处环境、自我决策并采取行动的人工智能实体(Xi et al., 2023)。大模型的卓越表现在智能体中展现出了巨大的潜力。在大模型智能体框架中，利用大模型强大的零样本指令跟随能力，智能体可以通过根据工具的描述学会如何理解和使用工具，完成更复杂的任务，例如通过搜索工具从外部查询更丰富的信息(Schick et al., 2024)。同时，多步推理能力促进了智能体对复杂任务的规划能力(Xi et al., 2023)。规划使智能体能够将复杂任务分解为子任务(Shen et al., 2024)，并制定合理的行动计划以实现目标。还有一些工作(Hong et al., 2023)合理利用多个智能体之间的协作，使其在各种应用场景中更加有效和可靠。

2.2 自动报告生成

2.2.1 抽取式自动报告生成

抽取式报告生成方法通过不同的信息处理小模型获得相关分析结果。这类方法通常需要人工预设好模板，然后使用数据统计(Babour and Khan, 2021)、文本摘要与关键词提取(Wang et al., 2019)等模型，对原始数据进行相关信息的提取，并填入对应的模板中。Babour(2021)提出一个自动生成指定国家科学计量研究分析报告的模型，通过使用数据统计模型对科学出版物数据进行详细的统计分析，填入模板并交由评审员进行评审修改得到分析报告。Noh(2020)提出一种名为Wise Issue Report的自动报告生成系统，利用主题生成模型生成不同主题，通过检索

和多文档摘要为每组新闻生成摘要，结合时间事件摘要模型生成不同的摘要内容，组合得到完整报告。

一方面，抽取式报告极度依赖模板，灵活性差；另一方面，抽取式获得的报告只能呈现浅层信息，无法进行深入分析，不适用于专业分析。

2.2.2 生成式自动报告生成

传统的生成式报告生成方法主要通过解码器模型在相关报告数据集上的端到端训练，获得报告生成的能力。Wang(2024)提出了一种基于条件变分自动编码器的方法，使用Bi-GRU编码输入的新闻，通过知识蒸馏和教师-学生网络结构来细化解码器组件的输出，实现金融报告的自动生成。Ren(2021)提出了一种新的混合深度生成神经模型，使用具有注意力机制的指针生成器网络学习输入新闻的大纲，通过改进的变分自动编码器生成宏观金融报告。Li(2023)为了提高放射学报告自动生成的质量，提出知识增强注入框架，通过结合医学概念和相似报告中的临床信息三元组，有效提高了报告的准确性和流畅性。

端到端训练的生成式方法虽然能够使模型掌握一定的报告语言规律并生成连贯的文本，但也往往导致模型过分关注于模仿训练数据中的表面特征，忽视深层语义逻辑。生成的报告可能在形式上看似合理，但在内容深度和逻辑性上却难以达到要求，可解释性不足。此外，高质量的专业报告资源相对稀缺，也限制了其对新领域或新主题报告的泛化能力。

2.2.3 基于大模型的自动报告生成

大模型在自动报告生成领域中正逐渐成为一种趋势。得益于大模型在语义理解与文本生成方面的强大能力，它为生成具有深度和广度的报告内容提供了新的可能性。这类方法结合抽取式和生成式报告生成的优点，采用分工合作的策略，其中大模型主要负责生成连贯的文本内容，而小模型则专注于原始数据的信息抽取，以辅助大模型生成更为详尽和信息丰富的报告。Colverd(2023)提出了一个利用LLM自动生成洪水灾害报告的系统，小模型负责检索关键词相似文档，大模型对文档进行评估并提取与洪水事件相关的信息，通过提示的方式让大模型回答预设的问题，从而生成一个内容全面，结构完整的洪水灾害报告。Reddy(2023)利用大模型自动生成态势报告，使用层次聚类进行新闻分类，利用小型生成模型生成类别标题，使用GPT3基于生成的类别标题进行战略子标题和内容的生成。

这种分工合作的方法虽然能够弥补传统生成式的不足，但仍然存在一些局限性。首先，生成报告的初步阶段通常需要人工设定报告的内容与使用的小模型，这可能会导致报告结构或内容的不全面。例如，在行业趋势预测中，除了行业标签和统计数据外，可能还需考虑龙头企业和专家的发展信息，以增加报告的丰富度和深度。此外，大型语言模型在生成报告时可能会因为回复的不稳定性而产生质量波动。这种不稳定性一方面来自小模型输出的错误与冗余信息，另一方面来自大模型生成文本时的随机性。与人类不同，大模型缺乏自主修正和完善的能力，这也限制了报告内容的优化。

3 AutoRG

在本节中，本文将详细介绍AutoRG框架，如图2所示。框架将报告生成的过程分为三个阶段。第一阶段，提供预设的一级题目和可利用的小模型集合，利用大模型生成尽可能丰富且可实现的子模块集合，将其作为报告的二级目录（第3.1节）；第二阶段，利用大模型的工具理解与摘要生成的能力，为每个子模块生成高信息量且更精准的分析数据（第3.2节）；第三阶段，以多智能体协作的方式对报告内容进行迭代优化，获得最终的报告（第3.3节）。

3.1 大纲规划与评估

大纲规划阶段的目标是为报告创建一个详尽可实现的目录结构。算法流程如1所示。

大纲规划：首先，将预定义的题目模板和可用的小模型集合交给大模型。通过特定的提示，引导大模型为每个题目模板规划生成一系列候选的子模块。这些子模块将作为报告的初始二级目录。在生成子模块时，要求大模型不仅提供模块的标题，还需要预测可能需要的数据类型，以便后续的信息收集。

可行性评估：尽管大模型能够生成丰富的子模块集合，但并非所有生成的子模块都能通过现有的小模型集合来实现。因此，需要对这些子模块进行筛选，确保它们与现有的小模型集合相匹配。大模型需要对每个生成的子模块进行反思并做出两个关键决策：确定哪些子模块是可

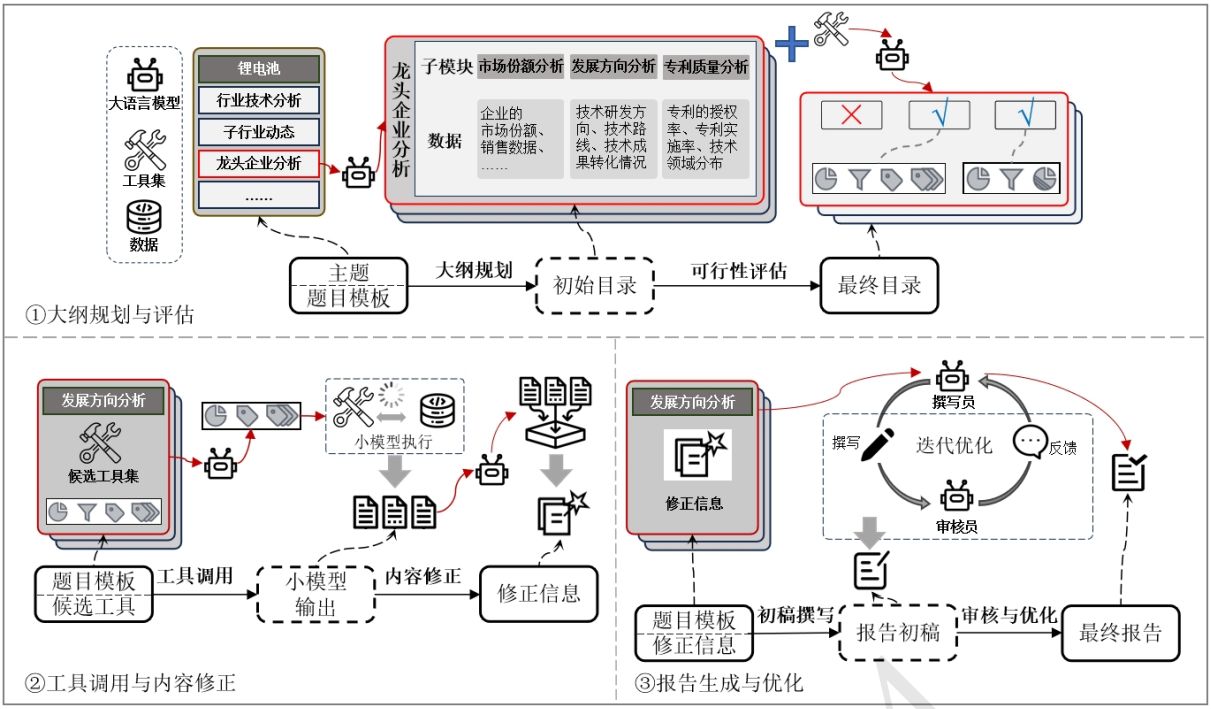

Figure 2: AutoRG报告生成框架

以执行的，以及哪些小模型可能适用于这些子模块。这是小模型首次被引入决策过程，目的是为了简化后续工具调用任务。具体来说，需要向大模型提供了小模型的列表以及完成子模块所需的数据类型。然后，大模型根据提示来判断每个子模块的可执行性，并在必要时从给定的小模型集合中选择合适的小模型子集。如果大模型评估某个子模块无法完成，则该子模块将被排除在最终目录之外。

3.2 工具调用与内容修正

工具调用与内容修正阶段旨在大模型完成小模型的自主选择，为每个子模块获取丰富且高质量的分析数据，确保报告的准确性和深度。具体来说，包括以下几个步骤:

工具调用：在工具调用步骤中，大模型需要根据当前子模块的任务要求，以及可用的小模型集合，进行精确的调用规划。小模型被视为大模型可利用的工具，大模型需要根据任务的具体需求，理解并决定调用哪些小模型，以获得最佳的输出结果。通过有效的工具选择，使得每个小模型都能在其所擅长的领域内发挥最大的效能，确保信息收集的全面性。

内容修正：内容修正的目的是对小模型的输出进行修正优化。大模型在规划步骤之后，并非直接生成报告内容，而是进行信息的过滤和提炼，识别并保留长文本中关键且注意力较大的信息，去除冗余和不相关的部分，降低小模型错误或有干扰的输出信息占比。本文使用直接提示的方式让大模型对小模型的输出进行审查，抽取出最有价值的信息，并将其转化为连贯、流畅的文本段落。这个过程有助于减少信息的冗余，并确保分析数据的高度关联与准确。

3.3 报告生成与优化

报告生成与优化部分旨在通过模拟真实人类书写报告时的交互，将修正后的分析数据转化为报告。这一阶段的目标是确保报告不仅表达清晰、逻辑性强，而且信息丰富、分析深入，能够为目标读者提供易于理解且实用的信息。在这一过程中，本文利用大模型分别模拟人类的专家审核员和报告撰写员的角色，通过协作的方式进一步提升报告的质量和准确性。

初稿撰写：在生成过程中，审核员将每个子模块的标题和整合后的数据提供给撰写员，要求其构建相应的初稿。在这里，审核员仅指定子模块的标题作为二级标题，而子模块内的报告格式则由撰写员，即由大模型自主决定。这种自主性赋予了大模型灵活性，使其能够根据每个子模块的具体内容来设计最合适的报告结构，确保报告内容的针对性和实用性。

Algorithm 1 大纲生成算法

输入: 题目模板$H = \{h_1, h_2, ..., h_n\}$, 报告主题$domain$, 小模型集合$T = \{t_1, t_2, ..., t_k\}$，大模型$M$
输出: 子模块信息矩阵$S = \{s_{ij}\}$，候选小模型信息矩阵$C = \{c_{ij}\}$，其中s_{ij} 是第i 个模板的第j 个子模块，c_{ij} 是s_{ij} 对应的候选小模型列表
过程:
for $h_i \in H$ **do**
    $S_i = M.generate(h_i, domain)$ ▷ 大纲规划
    for s_{ij} in S_i **do**
        if $M.judgement(s_{ij})$ **then** ▷ 可行性评估
            $c_{ij} = M.choose(T, s_{ij})$
            保存s_{ij} 与c_{ij}
        else
            删除s_{ij}
        end if
    end for
end for

Algorithm 2 工具调用与内容修正算法

输入: 第i 个模板的第j 个子模块信息s_{ij}，候选小模型列表c_{ij}，大模型M
输出: 优化后的分析数据$data_{new}$
过程:
$data_{old} =$ ""
while True and $c_{ij} \neq \emptyset$ **do**
    $c_k = M.choose(c_{ij}, s_{ij})$ ▷ 工具调用
    if 没有选择c_k **then**
        跳出循环
    end if
    $output_c = c_k.run()$
    $c_{ij} = c_{ij} \setminus \{c_k\}$
    $data_{old} = data_{old} + output_c$
end while
$data_{new} = correct\_func(data_{old})$ ▷ 内容修正

审核与优化：正如一篇优秀的报告往往需要经过审核和反复修改，自动报告生成同样需要经历一个细致的迭代过程，以确保最终产出的报告达到高质量标准。在生成初稿之后，为了减少大模型生成内容的不稳定性，将对报告内容进行审核与优化。具体来说，审核员将对撰写员返回的初稿进行细致的检查，并提出改进意见，这些意见可能涉及格式、内容相关度、分析深度等多个方面。撰写员将根据审核员的反馈，在初稿的基础上进行近一步的优化和完善，从而得到最终的、高质量的报告。

4 实验

4.1 实验设置

本节以专利技术报告的生成为例，具体阐述相关实现细节。

数据：由于资源的限制，本文专注于生成2023年的专利技术分析报告，从特定的在线资源[0]中收集了大量2023年专利数据，包括专利标题、摘要、发明人、申请公司、IPC分类号等信息，并存储在数据库中，以便报告生成使用。为了及时捕捉技术信息，本文以双月作为一个生成周期。在这个周期内，报告覆盖了从宏观到微观、从广泛到具体的不同粒度主题，以确保进

[0]https://analysis.iprdb.com/, https://www.uyanip.com/

Algorithm 3 报告迭代算法

输入: 第i 个模板的第j 个子模块信息s_{ij}，修正后的数据$data_{new}$，审核员A，撰写员M
输出: 第i 个模板的第j 个子模块的报告内容$report_{ij}$
过程:
$draft = M.generate\_draft(s_{ij}, data_{new})$ ▷ 初稿撰写
$feedback = A.review\_draft(draft)$ ▷ 审核
if 没有新的$feedback$ **then** 不做操作
else
    $draft = M.optimize\_draft(draft, feedback)$ ▷ 优化
end if
$report_{ij} = draft$

行全面而深入的实验。最终，共计生成23篇报告，138个子模块。

小**模型**：本文设计了报告生成中常见的小模型，包括：标签抽取模型、聚类模型、质量评分模型、统计模型。其中，标签抽取模型负责从专利文档中提取公司、专家和行业的相关技术标签，快速识别专利的技术领域和相关性；聚类模型将相似的专利归类到相应的子类别中，有助于在报告中展示专利的分布情况；质量评分模型通过评估专利的权利要求数量、法律状态等因素，使用岭回归模型对专利进行评分；统计模型对专利数据进行量化分析，包括计算专利的IPC分布、专利类型比例等，为报告提供宏观视角。

大模型：大模型在自动报告生成中扮演着重要的角色，负责规划和处理从小模型中获得的数据，以及最终生成连贯的报告文本。考虑到实际应用的可行性和成本效益，本文选择智谱AI作为生成报告的大模型。智谱AI基于GLM构建，通过大量的中文数据训练和对齐，是一个在中文自然语言处理领域表现出色的生成式语言模型。

对比基线：AutoRG旨在成为一个通用的报告生成框架，不局限于特定类型的报告，其生成的报告目录由大模型动态决定，这导致在报告结构上与现有相关工作不同。同时，现有的报告生成方法通常专注于特定类型的报告，生成流程和结构目录高度定制化，使得跨方法的直接对比变得复杂。因此，本文从现有工作的工作流，即半模板式大模型报告生成方法的角度进行对比。为确保比较的一致性，本文以大纲生成阶段输出的目录作为报告的标题。在半模板式大模型生成方法中，报告内容的生成使用人工判定的方式，选择最适合当前子模块的一个小模型，并将输出直接交给大模型进行报告生成。

消融：同时，为验证方法中不同模块的贡献，本文设置了消融实验。分别通过与不进行工具选择、不进行工具调用规划、不做内容修正以及不进行审核优化下生成的报告进行对比，直观分析这四个部分对最终报告质量的影响。

4.2 评估方法

考虑到一篇完整的报告内容篇幅可能过长，导致评估过程变得复杂和耗时，本文选择了以模块为单位进行评估。具体来说，根据二级标题将报告划分为多个独立的部分，每个部分都对应一个特定的报告内容。这样的划分可以使得评估过程更加高效。

主观评估：主观评价指标体系的构建参考了以往报告生成质量评估(Babour and Khan, 2021)的工作，包括以下四个维度：事实性、一致性、充分性和整体质量。其中，事实性指标旨在评估报告内容的准确性和真实性；一致性指标关注报告内部逻辑的连贯性以及与既定主题的相关性；充分性指标则评估报告内容是否全面，是否存在信息的遗漏或冗余；整体质量指标综合考量报告是否为读者提供了新的见解和知识。围绕上述评价体系，本文分别用模型评估和人为评估的方式从主观层面对报告生成的质量进行评价。

模型评估为每个评价指标设定独立的投票标准，以确保从不同维度对报告的质量进行细致的考量。本文使用Sparkv3.5[1]和Kimi[2]作为评估模型，通过两两比较的方式让评估模型进行投票。为了避免评估过程中位置造成的影响，本文随机安排了待评估段落的顺序。以半模板式报告作为基准，通过计算胜率来评估不同方法相对于半模板式报告的表现。胜率的计算公式

[1]https://xinghuo.xfyun.cn/
[2]https://kimi.moonshot.cn/

	事实性			一致性			充分性			整体质量		
	Spark	Kimi	avg	Spark	Kimi	avg	Spark	Kimi	avg	Spark	Kimi	avg
AutoRG	**61.59**	**61.59**	**61.59**	**50.72**	60.14	**55.43**	**64.49**	64.49	**64.49**	**60.14**	**63.77**	**61.96**
w/o 可行性评估的工具选择	58.54	56.91	57.72	39.84	58.54	49.19	57.72	63.41	60.57	56.10	59.35	57.72
w/o 工具调用规划	60.98	54.47	57.72	47.15	52.85	50.00	64.23	58.54	61.38	57.72	57.72	57.72
w/o 自主信息修正	60.98	60.16	60.57	49.59	60.16	54.88	59.35	**65.04**	62.20	55.28	62.60	58.94
w/o 内容审核迭代	53.03	60.61	56.82	43.94	**60.61**	52.27	53.79	62.12	57.95	49.24	62.12	55.68

Table 1: 不同报告生成方法下胜率的对比

如1，k为报告数量,n_i为第i篇专利比较的次数，也是子模块的数量，a_i为被选择的数量。

考虑到完整报告长度与专业性可能导致评估困难，因此本文采取了以模块为单位的整体评价的方式进行人为评估。每个评审员将随机负责3篇报告的对比，根据评分标准进行综合投票。

$$win = \sum_{i=1}^{k} \frac{a_i}{n_i} \tag{1}$$

客观评估：为降低主观性对评估结果的影响，本文进一步设置了不同的客观评估指标。这些指标涵盖了词汇多样性、内容联系度、领域独特性和句法复杂性四个方面，旨在从更丰富的维度对生成报告进行客观对照。词汇多样性的评估采用了类符-形符比（Type-Token Ratio, TTR）作为衡量指标。TTR通过计算文本中不同词汇类型（type）与总词数（token）的比率，来量化文本的词汇丰富度。这一指标能够有效反映报告中词汇的使用广度和多样性。

内容联系度和领域独特性的评估则采用了self-BLEU方法。其中，内容联系度旨在衡量一篇报告内部各个模块之间的联系是否紧密，通过比较报告内部模块与其他模块之间的相似性，来衡量报告整体的联系度。领域独特性旨在评估不同领域之间报告与报告的相似性，如果对于同一个模块，不同领域报告之间的内容过于相似，这样的报告在领域深度上有较大的提升空间。

句法复杂性的评估基于句子嵌套深度的计算。通过测量句子中嵌套句法结构的最大深度，可以评估句子的句法复杂度，进而反映句子结构的复杂性，以及生成文本的句法多样性。

$$TTR = \frac{\#\text{不同类型的词（Type）}}{\#\text{全部词（Token）}} \tag{2}$$

$$self - BLEU = exp(-\sum_{i=1}^{n} w_i \cdot \log \frac{C(i)}{Ni}) \tag{3}$$

4.3 实验结果分析

4.3.1 主观评估

（1）AutoRG生成的报告在内容上更加充分、全面。如表1所示，模型评估结果揭示了AutoRG与半模板式报告方法相比的胜率，其中高亮显示的部分表示AutoRG的平均胜率超过了半模板式大模型报告生成方法。特别是在报告的充分性这一关键指标上，AutoRG展现出了显著的优越性。这归功于AutoRG利用大模型的自主规划能力，能够自主选择并整合多种小模型，生成内容更为全面和详尽的报告，证实了AutoRG框架在增强报告内容丰富度方面的显著效果。

（2）AutoRG生成的报告在质量上更稳定，效果更好。评估结果显示，AutoRG在事实性、一致性以及整体质量方面均超越了传统的半模板式报告方法，在报告质量上有着一定的优势。此外，通过箱型图（见图3）对生成报告的胜率分布进行可视化分析，可以观察到AutoRG在所有评估指标的胜率分布上均显著高于半模板式报告，并且在一致性、充分性和整体质量中均呈左偏分布，这表明在多数情况下，AutoRG能够产生质量较高的报告。这一统计特性凸显了AutoRG在维持报告质量方面的出色能力。

同时，本文将人为评估结果通过网页形式进行了直观展示[3]。具体来说，图4展示了评审员对投票结果的统计分析。在这项评估中，每位评审员均随机针对评估规则对一篇报告的不同模

[3]http://8.130.143.149:5000/

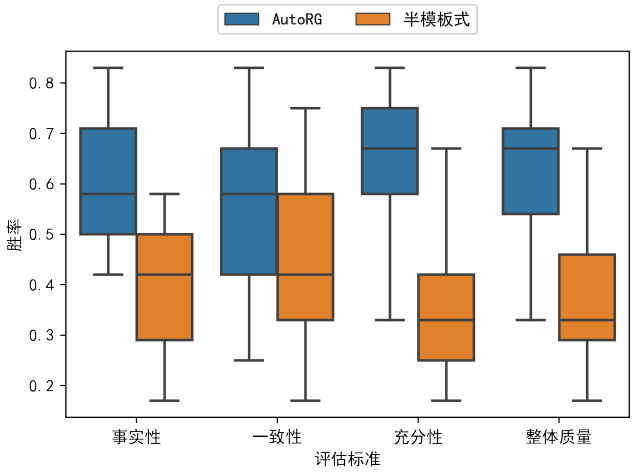

Figure 3: 不同维度下箱型图对比

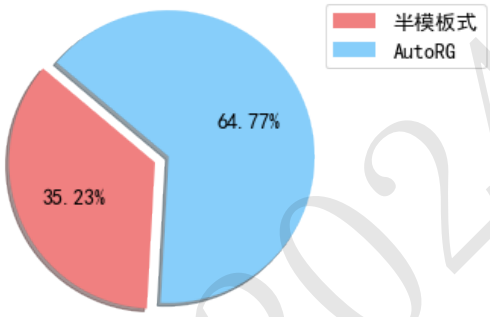

Figure 4: 人为评估投票结果

块分别进行了投票。图中可以清晰地观察到AutoRG在人为评估中获得了压倒性的优势，这一结果进一步证实了AutoRG的有效性和优越性。

4.3.2 消融实验

为了更清晰对比不同模块对最终报告生成结果的不同，本文在表2统计了消融实验下各版本与AutoRG平均胜率的差值，并有如下几点结论：

（1）多轮工具选择对提升报告一致性有积极作用。在可行性评估阶段，工具选择和工具调用规划对一致性的影响尤为关键。直接使用所有小型模型作为候选工具集会导致一致性下降6.24%，而将初轮工具选择的结果直接作为候选工具集则导致5.43%的下降。这归结为对庞大工具集的单轮选择虽然一定程度确保了信息的多样性，但也降低了信息的联系度，削弱了子模块与分析内容之间的相关性。而本文采用先粗后精的方式，通过初步筛选和基于详细描述的进一步筛选，提升了报告的一致性和连贯性。

（2）输入数据的质量影响报告的整体质量。自主信息修正模块的作用是对小模型的原始输出做信息的过滤与整合。从表2的结果可以发现，使用初始小模型的输出作为输入数据会导致报告在分析层面的质量下降，这一点在报告整体质量的显著降低中得到了体现。在这一过程中，过多的非相关信息会干扰分析过程，造成输入的冗长和不集中。自主信息修正模块可以进行有效的数据净化，提高报告的分析深度。

（3）迭代对报告内容的优化至关重要。与其他模块相比，当取消内容审核迭代阶段，生成报告的四个指标均有更显著的下降，这揭示了迭代对报告内容优化的重要性。迭代提供了精炼报告内容的机会，允许对初版中可能被忽视的细节和未充分探讨的分析进行深入审视。这种修正和完善的过程，确保了最终报告能够全面且准确地反映分析结果，提高了报告的质量。

第二十三届中国计算语言学大会论文集，第539页-第552页，太原，中国，2024年7月25日至28日。
卷1：主会论文

	事实性	一致性	充分性	整体质量
AutoRG	-	-	-	-
w/o 可行性评估的工具选择	↓3.87	**↓6.24**	↓3.92	↓4.24
w/o 工具调用规划	↓3.87	↓5.43	↓3.11	↓4.24
w/o 自主信息修正	↓1.02	↓0.55	↓2.29	↓3.02
w/o 内容审核迭代	**↓4.77**	↓3.16	**↓6.54**	**↓6.28**

Table 2: 消融实验下胜率的变化

	TTR	报告内self-BLEU	报告间self-BLEU	句法深度
半模板式	0.4805	0.0216	0.0889	4.7793
AutoRG	0.6066	0.0229	0.0830	4.7889

Table 3: 客观指标下不同报告生成方法的对比

4.3.3 客观评估

（1）AutoRG生成的报告在词汇运用上更加多样。如表3所示，AutoRG在TTR指标上明显超越了传统的半模板式方法。这一结果揭示了AutoRG在词汇选择上的丰富性，更有效地传达了信息。这种词汇多样性的实现主要归功于在工具调用、内容修正与迭代阶段的设计。AutoRG不仅能自主选择与任务最匹配的多样化小模型，还能通过信息修正精准提炼关键信息，运用更为丰富和多变的词汇进行表达。这一过程增强了报告的准确性，也提升了其可理解性。而迭代过程的优化则进一步增强了报告的词汇丰富度与表达能力。

（2）AutoRG生成的报告在模块间建立了更紧密的联系。实验数据显示，AutoRG生成的报告在报告内self-BLEU指标上表现出更高的相似度。这得益于规划与工具调用，使得相同的小模型信息在多个模块之间流畅传递。这意味着报告中的各个模块不再是孤立的信息点，而是相互关联、相互支撑的整体。这样的结构使得读者在阅读报告时能够更容易地跟踪信息的脉络，理解各个部分之间的逻辑关系，从而更全面地把握报告的核心内容。

（3）AutoRG生成的报告在不同领域间展现出更高的适应性与差异性。通过对比报告间self-BLEU指标，可以发现，AutoRG在领域多样性方面具有一定的优势。报告之间更高的差异性表明能够根据不同领域的需求，提供更为精准和独特的分析报告。这种差异性依然得益于信息修正模块与报告迭代过程中的细化调整。通过模拟人类专家的审核与优化过程，自动式生成能够对报告进行定制化调整，提高了报告的专业度，也确保了其在不同领域间的差异性和适用性。

（4）AutoRG生成的报告更具专业性。表3中，通过测量句法深度，可以发现AutoRG生成的报告在句法结构的复杂性上表现出细微的优势。这种相对复杂的句子表达说明AutoRG在分析内容上相比半模板式在内容上更为透彻和深入，使得生成的报告在专业性上可能更胜一筹。

5 局限与未来方向

在分析AutoRG报告生成优势的同时，本文也关注到其存在的局限，尤其是在事实性与一致性方面。通过评估模型的反馈，发现针对事实性的评估主要集中在真实性方面，无从判断生成报告的准确性与幻觉程度，对于错误信息的处理也依然有较大的改进空间。随着资源的积累和研究的深入，我们也将在未来的工作中继续开展这一问题。此外，AutoRG在报告生成中采用了固定的模块结构，但在实际报告编写过程中，可能需要对报告的目录进行反复调整。本文计划探索大模型在自动调整报告结构方面的潜力，以更好地适应报告编写的动态性和复杂性，为用户提供更加全面和高质量的报告服务。

6 总结

本文提出了一个大小模型协同的自动报告生成框架AutoRG。通过有效利用大模型的工具理解与自主规划能力，自动化实现更灵活和全面的报告。同时，通过引入信息修正机制与报告迭代机制，降低对小模型输出和大模型生成随机性的依赖，提高了报告的质量与稳定性。本文在专利场景下以技术报告生成为例，通过模型主观评估和客观评估结合的方式，从多个维度证明，AutoRG生成的报告有着更优的效果。

第二十三届中国计算语言学大会论文集，第539页-第552页，太原，中国，2024年7月25日至28日。
卷1：主会论文

参考文献

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints. *arXiv preprint arXiv:2302.09185*.

Amal Babour and Javed I Khan. 2021. Automatic Generation of a Metric Report: A Case Study of Scientometric Analytics. *IEEE Access*, 10:3923–3934. IEEE.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, 33(01):6666–6673.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*.

Grace Colverd, Paul Darm, Leonard Silverberg, and Noah Kasmanoff. 2023. FloodBrain: Flood Disaster Reporting by Web-based Retrieval Augmented Generation with an LLM. *arXiv preprint arXiv:2311.02597*.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. Qadynamics: Training dynamics-driven synthetic qa diagnostic for zero-shot commonsense question answering. *arXiv preprint arXiv:2310.11303*.

Hilya Tsaniya, Chastine Fatichah, and Nanik Suciati. 2024. Automatic radiology report generator using transformer with contrast-based image enhancement. *IEEE Access*. IEEE.

Jingwen Wang, Hao Zhang, Cheng Zhang, Wenjing Yang, Liqun Shao, and Jie Wang. 2019. An effective scheme for generating an overview report over a very large corpus of documents. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–11.

Junpeng Gong, Wen Ren, and Pengzhou Zhang. 2017. An automatic generation method of sports news based on knowledge rules. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 499–502. IEEE.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2023. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1):253–270. Springer.

Qingqiu Li, Jilan Xu, Runtian Yuan, Mohan Chen, Yuejie Zhang, Rui Feng, Xiaobo Zhang, and Shang Gao. 2023. Enhanced Knowledge Injection for Radiology Report Generation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2053–2058. IEEE.

Revanth Gangi Reddy, Yi R Fung, Qi Zeng, Manling Li, Ziqi Wang, Paul Sullivan, and Heng Ji. 2023. Smartbook: Ai-assisted situation report generation. *arXiv preprint arXiv:2303.14337*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Takeshi Nakaura, Naofumi Yoshida, Naoki Kobayashi, Kaori Shiraishi, Yasunori Nagayama, Hiroyuki Uetani, Masafumi Kidoh, Masamichi Hokamura, Yoshinori Funama, and Toshinori Hirai. 2024. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, 42(2):190–200. Springer.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 36.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Yujuan Ding, Yunshan Ma, Wenqi Fan, Yige Yao, Tat-Seng Chua, and Qing Li. 2024. FashionReGen: LLM-Empowered Fashion Report Generation. *arXiv preprint arXiv:2403.06660*.

Yunpeng Ren, Wenxin Hu, Ziao Wang, Xiaofeng Zhang, Yiyuan Wang, and Xuan Wang. 2021. A hybrid deep generative neural model for financial report generation. *Knowledge-Based Systems*, 227:107093. Elsevier.

Yunseok Noh, Yongmin Shin, Junmo Park, A-Yeong Kim, Su Jeong Choi, Hyun-Je Song, Seong-Bae Park, and Seyoung Park. 2020. WIRE: an automated report generation system using topical and temporal summarization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2169–2172.

Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan Wang. 2024. Generating long financial report using conditional variational autoencoders with knowledge distillation. *IEEE Transactions on Artificial Intelligence*. IEEE.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

附录A.大纲规划与评估的大模型提示:

大纲规划	如果生成“{domain}专利的技术分析报告”，对于其中的”{outline}”模块，可能包括哪些部分?分别需要哪些数据? 要求的输出：1. 分点回答；2. 格式是“1. xx分析
可行性评估	# 你的任务是生成一个名为”{domain}专利技术分析报告” 的{task_name}模块。 # 完成该模块必须需要以下数据： {task_need_data} # 有以下候选模型： {model_info} # 请判断： 1.根据候选模型的输出，是否可以获得完成该模块需要的数据？（{task_need_data}） 2. 如果不可以，或者需要超过四个模型，那么请直接告诉我“不可以”。（例如，信息统计模型无法获取研发投入信息） 3. 如果可以，你应该如何选择和组合这些模型？请注意，最多只能选择四个模型，如果超过四个，参考第二点。（例如，信息统计模型可以获取专利数量信息） # 返回的格式有两种： - “不可以。因为无法获取xx信息” - “可以。流程是：1. 使用xxx进行xxx\n2. 使用xxx进行xxx”

附录B.工具调用与内容修正的大模型提示:

工具调用	### 任务 工具判断。判断以下是否存在符合条件的工具。如果没有，则直接回答‘无法完成‘；如果有，则请从中选择合适的工具，用于完成当前需求。 ### 需求 {domain}技术报告-{title}中包含以下模块：{tool_list}，而其中你需要完成的是{tool_key}模块。 ### 工具 {tools_info} ### 要求 1. 选择的工具必须能够提供完成{tool_key}需求的全部或部分需要的数据，可以提供地不完整，但是不能完全无关； 2. 选择的时候需要考虑与其他模块间的重合度，根据情况可以舍弃重合度高的非关键工具； 3. 选择的时候务必带有客观性和严肃性，不要有情感倾向； 4. 请严格遵守工具输出的信息，请注意前后回答、逻辑的一致性； ### 返回格式 1. 需求分析。 2. 工具分析。包括分析该工具可以提供什么信息? 这个信息和需求有关联吗? 3. 综上所述，我的回答是：“无需再选择” 或者综上所述，我的回答是：“xx模型”
内容修正	# 任务 我将给你一段很长的信息文本，请理解背景并耐心阅读文本内容，然后以长文的形式进行信息的过滤与整合。 # 背景 {domain}{tool_key} # 信息: {text} # 要求 1. 保留的内容和{domain}{tool_key}相关；2. 以长文的形式进行总结，保留的内容尽可能丰富、细致；3. 内容尽可能流畅，不要有任何分点的格式；4. 不要过度忽略重要细节，重要细节可能对后续分析有极大帮助；5. 直接回答整合后的长文即可。

附录C.报告生成与优化的大模型提示:

初始撰写	### 任务 我将给你一些对撰写{domain}{tool_key}有帮助的信息，它们可能比较长，请你耐心阅读并撰写{domain}行业技术分析报告——“{tool_key}”模块的内容。 ### 信息 {text} ### 要求 1. 第一步先简要梳理关键信息；2. 第二步重点结合{tool_key}，对这些信息分析彼此联系；3. 最后，扩展思维，围绕{tool_key}进行技术层面的深入思考；4. 以“### {tool_key}”开头，直接输出报告即可。
审核	以下是一段{domain}领域技术报告对{subtitle}相关的内容，请你围绕{subtitle}，阅读并审查内容，返回作为专家2 3条最核心的批改意见。 ### 内容 ”{agent_1}” 要求返回格式：“1. 意见1；2. 意见2；3. 意见3”
优化	user: 初始撰写提示 assistant: {draft} user: {advice} 当然这只是我的意见，因为我并不了解原信息的全部内容，你可以根据自己的理解对初稿进行润色，使它质量更高、更流畅。 assistant: