

大语言模型在天气预报中的应用探讨*

代刊 高嵩 孟宏欣 唐健

国家气象中心，北京，100081

提要：本文旨在探讨大语言模型（LLM）在天气预报中的应用潜力及其面临的挑战。文章分析 LLM 在气象及相关行业应用，包括知识检索、基座模型、诊断分析、工具调用及文字生成等应用场景，指出 LLM 在提升天气预报的精准度和业务智能化水平方面有巨大潜力。LLM 通过高效处理海量气象知识、整合跨领域多源信息、生成定制化预报产品等，为预报员提供了强大的辅助工具。通过构建高质量的气象语料库、优化基准测试框架、结合外部工具等，可进一步提升 LLM 在天气预报中的应用效果。LLM 为气象领域带来了新的技术机遇，但其广泛应用仍需在语料质量、模型优化及人机协作等方面持续探索与完善。LLM 在大气运动时空理解、偏见与幻觉等方面仍存在局限性，需通过数据清洗、去偏见及微调、检索增强生成等技术加以改进。

关键词：大语言模型，天气预报，检索增强生成，气象智能体

DOI: 10.7519/j.issn.1000-0526.2025.061901

Exploration of the Application of Large Language Models in Weather Forecasting

DAI Kan GAO Song MENG Hongxin TANG Jian

National Meteorological Centre, Beijing 100081

Abstract: This article aims to explore the potential applications and challenges of Large Language Models (LLMs) in weather forecasting. Through analyzing LLMs' applications in meteorology and related industries, including knowledge retrieval, foundation models, diagnostic analysis, tool calling and text generation, the article demonstrates that LLMs have tremendous potentials in improving weather forecast accuracy and business intelligence. LLMs serve as powerful

assistive tools for forecasters by efficiently processing massive meteorological knowledge, integrating
*国家气象中心重点创新团队智能预报技术项目（CA2022Z000）
第一作者：代刊，主要从事智能预报技术、统计后处理技术的研究，Email: daikan1998@163.com.
通讯作者：高嵩，主要从事天气预报业务系统及平台研发，Email: gaos@cma.gov.cn

cross-domain multi-source information, and generating customized forecast products. The effectiveness of LLMs in weather forecasting can be further enhanced by building high-quality meteorological corpora, optimizing benchmark testing frameworks, and incorporating external tools. While LLMs bring new technological opportunities to the meteorological field, their widespread application still requires continuous exploration and improvement in corpus quality, model optimization, and human-machine collaboration. Moreover, LLMs still have limitations in understanding the spatio-temporal dynamics of atmospheric motions and issues with bias and hallucination, which need to be addressed through data cleaning, debiasing, fine-tuning, and retrieval-augmented generation techniques.

Key words: large language model, weather forecasting, retrieval-augmented generation, AI agent

引 言

天气预报作为科学研究与业务实践高度融合的工作,随着气象观测手段与规模的提升、数值天气预报模式的发展(Bauer et al, 2015)以及用户需求的精细化,面临多重挑战。信息过载使预报员从海量数据中准确、高效地提取关键信息变得愈加困难;而大气运动的多尺度特征及预报信息从单一大气层向地球多圈层的扩展,也对预报员的知识经验要求(Stuart et al, 2022)提出了更高的要求。随着预报业务的数字化进程加快,传统手工操作已难以满足自动化需求,亟需先进技术优化流程(Pagano et al, 2022; 唐健等, 2018);此外,预报产品用户日趋多元化,也对针对性和定制化提出了更高要求。这些挑战要求气象部门在技术与方法上持续创新,持续推动天气预报业务向智能化方向发展(代刊等, 2016)。

在科技迅猛发展的背景下,大语言模型(large language model, LLM)迅速崛起,为人工智能在多个垂直领域的应用带来新机遇(Zhang et al, 2024a)。LLM不仅具备文本生成、信息提取和数据分析等多种功能,还能处理表格、图像、视频、元数据及程序代码等多种数据类型(Yin et al, 2024)。统一的模型架构使LLM能高效解决广泛问题,通过生成假设、规划实验和设计复杂研究方案等方式提升研究效率,加速科学发现进程(Zhang et al, 2025)。例如在生物医学领域,LLM分析海量医学文献和电子健康记录,支持疾病预测和治疗方案制定,提升临床决策质量,并通过文献分析识别潜在新靶点(Busch et al., 2025; Singhal et al., 2023)。在化学与材料科学领域,LLM在分子生成、反应预测以及分子性质预测等任务

上的表现不断优化，为新材料设计与合成提供了有力支持（Frey et al, 2023）。在天气气候与环境科学领域，LLM 通过分析气候变化相关文献和数据，帮助理解气候模式，支持可持续发展决策（Li et al, 2024a），为政策制定提供科学依据（Vaghefi et al., 2023）。

Wang et al（2024a）的综述研究将 LLM 在地理空间科学中的关键功能归纳为四个方向：提取关键信息、转换数据形式、创造新内容、类人分析判断。这种对跨领域、多模态数据的强大处理能力，与天气预报整合气象数据、提升预报精度的核心需求高度契合，具有重要借鉴意义。

综上所述，LLM 在科学领域展现出卓越的数据处理、知识整合与创新能力，为解决复杂问题提供了新途径。本文参考 LLM 在多个领域的应用研究，探讨其在提高天气预报精准度和业务智能化中的应用和挑战；首先介绍 LLM 及其相关概念，随后分析其在天气预报中的典型应用场景与潜力，最后讨论推进 LLM 在气象领域应用所需要注意的关键问题，并进行总结。

1 大语言模型概述

自然语言处理领域的迅猛发展催生了大语言模型（LLM）这一革命性技术，其强大的理解与生成能力正深刻改变着人机交互方式。本节将系统介绍大语言模型的基本概念和技术特点，概述提示词工程、模型微调、检索增强生成等关键技术，并探讨智能体的构建原理，为理解 LLM 在气象领域的应用奠定理论基础。

1.1 大语言模型简介

LLM 是一种革命性的自然语言处理（natural language processing, NLP）技术，它通过在大规模语料库上预训练深度神经网络，以掌握广泛语言知识和复杂语义关系，因其参数规模巨大（Brown et al, 2020）和所需语料量庞大而得名。2017 年，Transformer 架构的提出（Vaswani et al, 2017）标志着语言模型进入了“预训练语言模型”阶段。基于 Transformer 的预训练语言模型（如 BERT、GPT 等）通过在大规模无监督语料上进行预训练，掌握了通用的语言知识，再通过微调（Fine-tuning）快速适配特定任务（Zhao et al, 2023）。这一阶段的技术突破为 LLM

的快速发展奠定了基础。随着模型规模的持续扩大，研究表明参数和训练数据量的增加显著提升了 LLM 的语言理解与生成能力，并在超过某一“临界值”后展现出涌现效应（Kaplan et al, 2020; Wei et al, 2022a），包括数学计算、基于上下文的学习、复杂推理和指令跟随等新兴能力。近期，我国 DeepSeek 开源大模型（Seek-AI et al, 2025）在推理能力和计算效率方面取得了显著突破，为气象领域的大模型应用提供了新的可能性。

1.2 大语言模型提示词

提示词用于“指挥”LLM 按照预设的思路去思考问题、输出内容，其作为触发器能激发模型的记忆、推理和生成能力（Sahoo et al, 2024）。提示词工程（prompt engineering）作为一种新兴方法，通过设计和优化提示词来充分挖掘 LLM 的潜力。近年来涌现出的多种高级策略，如少量样本提示（few-shot prompting; Brown et al, 2020）、思维链（chain-of-thought, CoT; Wei et al, 2022b）和验证链（chain-of-verification; Dhuliawala et al, 2023）等，已在提升模型任务表现方面展现出巨大应用前景。例如 Few-shot Prompting 使 GPT-3 模型在英法翻译任务中的 BLEU (bilingual evaluation understudy) 指标评分从 5.2 提升至 25.0，接近专门训练的翻译模型（28.4）（Brown et al, 2020）；CoT 提示词使谷歌的 PaLM-540B 模型在数学和逻辑推理任务中准确率分别提升 40.4% 和 11.2%。上述提示词方法的提出不仅丰富了提示词设计手段，还通过合理设计激发 LLM 的潜在能力，使其无需微调即可完成复杂推理、领域适配和多模态任务，大幅提升了模型的实用性和可靠性。

1.3 大语言模型微调

LLM 在预训练阶段学习了通用语言表达方式和广泛知识，但因语料中的噪声或错误信息，可能导致生成结果不可靠或内容与事实不符的“幻觉”现象（Maynez et al, 2020; Ji et al, 2023）。这种现象在对准确性要求极高的任务中，如医疗诊断、法律顾问以及灾害性天气预报，可能引发严重后果（Shuster et al, 2021; Lawson et al, 2025）。为了减少幻觉现象并提升 LLM 在特定任务中的适配能力，研究者提出了“微调”技术，通过少量领域语料对预训练模型进行微调，注入专

业知识 (Howard and Ruder, 2018)。这种方法有效降低了幻觉现象的发生概率，并显著提升了模型在特定领域的生成质量和可靠性。研究者提出了多种参数高效微调方法 (Han et al, 2024)，如适配层技术 (Houlsby et al, 2019)、低秩适配 (Hu et al, 2021)、提示微调 (Lester et al, 2021) 等。这些方法使 LLM 能够在不显著增加计算成本的情况下快速适配下游任务，提升专业性和准确性。

1.4 检索增强生成

除了模型微调，另一种用于适配专业领域任务的优化方式是基于检索的增强生成 (retrieval-augmented generation, RAG)，通过结合外部知识库和实时检索系统，提升生成内容的准确性和可靠性 (Gao et al, 2023)。RAG 架构包括索引、检索和生成模块，其中索引模块通过嵌入技术将文档向量化，检索模块根据查询向量与文档向量的相似度进行检索，生成模块将检索到的信息与模型内置知识结合生成准确内容 (Izacard and Grave, 2020)。与模型微调相比，RAG 的优势在于其动态更新能力，通过实时检索外部知识库避免因知识过时或不足而产生错误回答。此外，基于知识图谱的检索增强生成受到关注 (Peng et al, 2024)，知识图谱通过结构化的方式帮助智能系统理解实体之间的关联，为复杂检索任务提供准确答案 (Ji et al, 2022)。这种方法在医疗问答系统中表现出显著优势，通过实时检索权威医学数据库，生成更加准确的诊断建议，减少幻觉现象的误导性回答 (Xiong et al, 2025)。

1.5 智能体

智能体 (Agent) 指的是能够感知环境并采取行动以实现目标的实体。LLM 的出现为智能体发展提供了新的设计范式，凭借其在自然语言理解、生成、推理和多任务学习方面的优势，成为智能体核心决策引擎，并通过与外部工具结合，展现出更强的灵活性和通用性，成为当前研究的热点 (Xi et al, 2025; Schick et al, 2023)。基于 LLM 的智能体通常包括三个核心部分：大脑、感知和行动。大脑模块负责推理、决策和内容生成，借助 LLM 的自然语言处理能力、知识整合与推理能力，能够解析用户输入、执行复杂任务。感知模块从多种数据源获取信息并转化为大脑模块能够处理的输入，不仅可以处理文本信息，还能理解语音、图像

和视频等非文本信息。行动模块将大脑模块的决策转化为执行操作，通过工具调用机制从预定义的工具集合中选择合适的工具执行任务，目前比较领先的 LangChain 框架和 ReAct 框架都通过增强的推理与工具调用能力，帮助智能体高效、灵活地完成复杂任务（Yao et al, 2022）。

2 大语言模型的天气应用场景

气象业务面临信息过载、多尺度特征分析与预报产品定制化等多重挑战，LLM 凭借其强大的语言理解、知识整合与生成能力，为天气预报智能化发展提供了新思路。本节将从知识检索、预报基座模型、诊断分析、工具调用和文字生成五个维度，系统探讨 LLM 在提升天气预报业务智能化水平方面的应用场景与技术实践，展现其在气象领域的重要潜能。

2.1 天气预报知识检索

随着天气预报的不断发展，业务范围正在经历深刻的变革。在分析维度上，已从聚焦单一大气层扩展到多圈层的综合分析；在服务目标上，从主要关注灾害天气预警预报，延伸至涵盖天气风险评估的综合性防灾减灾信息服务。这一转变要求预报员具备跨学科知识，以应对复杂气象现象及其对人类活动的广泛影响。然而，传统的预报员培养方式难以满足实际业务需求，亟需新方法和工具提升专业能力。在此背景下，LLM 作为新兴技术，正在革新知识检索模式。它通过深度学习高效压缩海量知识（Del éang et al, 2023），不仅能快速检索信息，还能辅助预报员在复杂气象条件下做出更准确判断，将为气象领域提供强大支持。

在专业领域的知识检索应用中，LLM 在处理复杂数据、生成专业内容和实时更新信息方面展现出巨大潜力。例如 OceanGPT（Bi et al, 2024）作为首个专注海洋科学的 LLM，通过多代理协作生成海洋指令数据集，解决数据获取难题并展示专业知识。ChatClimate（Vaghefi et al, 2023）整合 IPCC 第六次评估报告数据，采用检索增强生成（RAG）技术实时获取最新气象数据和研究成果，确保信息的及时性和准确性。CliMedBERT（Fard et al, 2022）作为预训练语言模型，捕捉气候事件与健康结果的关系，通过广泛训练提升在命名实体识别、文本分类

和事实核查等任务中的表现，助力识别气候变化对人类健康的潜在影响，为政策制定提供科学依据。MapGPT (Fernandez and Dube, 2023) 将 LLM 与空间数据处理相结合，通过自然语言理解与空间分析的融合，提高位置相关查询的响应准确性。

LLM 的气象知识来源多样，除了训练过程中积累的历史气象记录和相关文献之外，还可通过模型微调融入特定领域知识，如天气业务文档、气象数据表格、数值预报及预报员的主观判断等，提升任务表现。RAG 技术进一步增强了 LLM 的信息检索能力，使其在用户提问时能够实时获取最新信息。例如，在极端天气查询中，LLM 不仅基于训练数据生成回答，还能提取最新观测和预报信息，确保信息的可靠性。然而，在实际应用中仍面临数据质量问题，历史气象数据和实况资料可能缺失或不准确，因此需建立完善的数据清洗与处理机制，构建高质量的气象语料库。此外，大气运动的混沌特性使得预报模型在预测未来天气信息时包含不确定性，因此，使用 LLM 进行天气预报信息检索时，需结合专家经验和专业知识进行审核，以提高决策的可靠性。

总之，与传统关键词检索系统相比，LLM 在天气预报知识检索中具有显著优势。传统检索系统依赖精确的关键词匹配和预定义的语义规则，而 LLM 能够理解更加复杂的查询意图并从海量非结构化资料中提取相关信息，利用 RAG 技术，信息的时效性和准确性也将会不断提高。

2.2 天气预报模型基座

业界将专为特定科学任务（如天气预报）训练、并能为下游应用提供基础预报预测的模型，称为科学基座模型。由数据驱动的天气预报 AI 模型，自 2022 年以来利用再分析数据进行训练已取得了显著进展 (Chen et al., 2023c)。以盘古 (Bi et al, 2023)、风乌 (Chen et al, 2023a)、伏羲 (Chen et al, 2023b) 等为例，这些 AI 模型采用先进的深度学习算法，从海量的气象数据中提取复杂的模式和规律，显著提升了气象要素的预测准确性。此外，集成多种数据类型的基座模型，如 AtmoRep (Lessig et al, 2023) 和 Aurora (Bodnar et al, 2025)，也成为天气预报领域的重要发展方向。这些基座模型通过融合再分析数据、遥感数据、实况观测数据等多源信息，增强了模型的鲁棒性和预测能力。通过在不同的时空尺度上

进行优化，基座模型能够在短期、中长期、空气污染等多个下游任务中表现出优异的性能。

除了上述科学基座模型之外，LLM（特别是 GPT-4）逐步具备了通用基座模型的能力，因此也可直接用于作为科学发现或预报预测模型。Wang and Karimi（2024）研究了 LLM 在短期和长期降雨预测中的表现，发现其倾向于生成保守的预测，回归历史平均值。尽管 LLM 在整合专家数据方面存在挑战，但其能够捕捉气象变量间的复杂非线性关系，发现传统模型难以识别的模式。随着技术进步，LLM 有望成为气象预报的重要工具。Li et al（2024b）提出了一个多模态基准数据集 CLLMate。CLLMate 通过整合 2.6 万余篇环境新闻文章与 ERA5 再分析数据，首次实现了数值气象数据与文本事件数据的对齐。对 23 种现有的多模态大型语言模型（MLLM）的系统基准测试结果表明，尽管 MLLM 在某些任务上表现优于启发式基线，但其整体性能仍受限，尤其是在细粒度事件类型预测方面。该研究强调了进一步优化模型架构以提高预测准确性的必要性，并为未来的研究提供了一个有价值的基准。

作为天气预报模型基座，LLM 与由物理规律驱动的数字预报模式、气象数据驱动的科学模型相比具有独特的潜力。LLM 融合了人类的海量科学知识，能够直接模拟预报员进行主观预报的思考，此外其多模态能力能够快速浏览海量的气象数据，因此在理论上具备可行性。目前，LLM 作为基座模型的应用仍处于探索阶段，其物理一致性及在长时间预报中的稳定性有待进一步验证。未来，LLM 与物理模型的混合架构可能成为提升预报精度的重要方向。

2.3 天气预报诊断分析

现代化的精准天气预报，要求预报员具备对各类天气图进行快速诊断分析的核心技能（Stuart et al, 2022）。然而，随着观测数据来源与数值模式迭代日趋多元，传统依赖专业知识和经验的预报方式已难以适应当下数据爆炸的环境，预报员在有限时间内高效筛选海量图形产品并提炼关键信息的难度不断增大。为了高效应对复杂的卫星、雷达及地面站观测数据等，预报员被迫兼具扎实的气象学基础知识，和较好的快速分析与综合判断能力。

2024 年开始，LLM 逐步迈入多模态阶段，跨语言视觉大语言模型（LVLM）

在图表和视频内容的解读领域显示出显著潜力 (Huang et al, 2023)。相关研究表明, LVLM 可高效处理文本、图像与视频等多种输入形式, 从而拓宽科学诊断分析的应用范畴, 并能通过自然语言交互增强多学科分析能力。例如, GeoChat (Kuckreja et al, 2024) 可针对高分辨率遥感影像进行问答, 并以坐标标识对象位置; EarthGPT (Zhang et al, 2024b) 通过统一指令调优整合多传感器遥感任务, 支持场景分类、图像描述与视觉问答等功能。

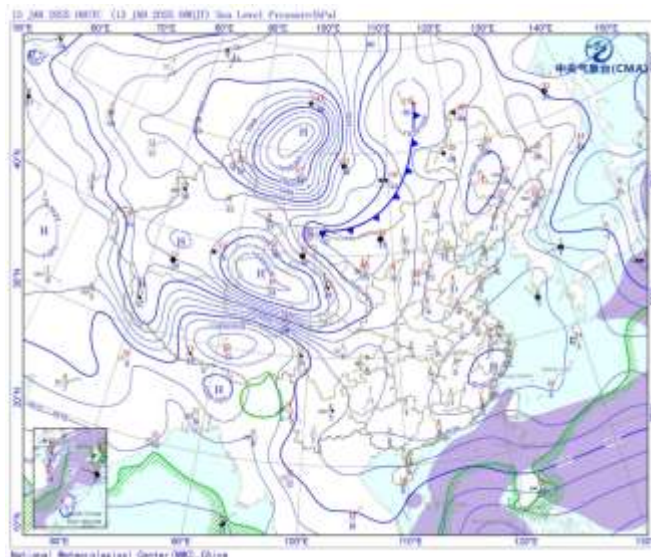
在气象预报应用方面, Lawson et al (2025) 的研究表明, GPT-4V (OpenAI et al, 2023) 在多张气象要素图的综合解读上初显优势。针对对流有效位能 (CAPE) 图, GPT-4V 能够识别密苏里州至伊利诺伊州的不稳定区, 并指明德克萨斯州沿海与墨西哥湾的模式间 CAPE 差异最为显著。在 700Pha 高度上, 能捕捉俄克拉荷马州至伊利诺伊州的高湿度带以及阿肯色州、密苏里州的水汽辐合; 在 300 Pha 高度上, 能识别急流的位置与强度差异; 这展现了其综合分析多层次风场、温度场和湿度场的潜力。在模拟美国强天气预报时, 其所示高风险区域与官方 (美国强风暴预报中心) 结果有较好一致性, 体现出类似资深预报员的综合分析判断能力。

提示词：

你是天气预报员助手工具，可以协助气象学家解读各种气象图表和数据。输入是一张东亚区域的地面天气分析图，采用地面气象站点观测数据制作，包括高/低压系统、气旋锋面和天气现象等。

请输出：识别提供的气象图表类型；识别地面图分析的日期；密切关注等压线，等压线可能紧密或分散。从图表中提取相关的气象信息；解释图表中描绘的预期天气条件；识别并解释重要的天气特征（例如高/低压系统、气旋锋面等）。

图中要素说明如下：气象站点的填图包括：云量、风向风速、天气现象、温度（红色数字）和露点（蓝色数字）；天气系统符号包括：高压中心（H）、低压中心（L）、冷锋（带三角形蓝色线段）、暖锋（带半圆形红色线段）和台风符号；蓝色等值线为海平面气压场等压线（单位 hPa）；紫色填充区为六级以上大风速区；绿色线条为等比湿线，其栅格化一侧为比湿达到 15g/kg 的区域；



模型输出：

气象图表类型：提供的图表为地面天气分析图，基于地面气象观测数据绘制，展示了海平面气压分布、高低压系统、锋面及天气现象等。

地面图分析的日期：图表显示的日期为 2025 年 1 月 13 日 00 时 UTC（北京时间 08 时）。

等压线分布：图中蓝色等压线表示海平面气压分布，单位为 hPa。等压线在北方部分区域（如蒙古东部和中国东北地区）分布较密，显示气压梯度较大，可能伴随较强的风力。东南部沿海地区气压梯度较小，等压线分布较稀疏。

相关气象信息：高压系统（H）：多个高压中心出现在中国西部、中部以及东南沿海地区，显示稳定的大形势，天气可能以晴朗为主。低压系统（L）：东北方向存在一个低压系统，可能带来降雪或雨夹雪天气。冷锋（蓝色三角线）：冷锋从中国北方延伸至东北方向，表示较冷的气团正在向南推进，可能带来降温 and 降水。强风区域（紫色）：在东南沿海和南海部分区域显示六级以上强风，风速较高，需防范海上大风带来的影响。

预期天气条件：中国北方及东北地区：冷锋影响区域，预计会出现降温、降雪或雨夹雪，伴随风力增强。中国中部和西部：受高压控制，天气相对稳定，可能以晴到多云为主，气温较低。东南沿海及南海：大风天气显著，局地可能伴有阵雨或对流性降水。

重要天气特征及解释：高压系统：高压中心位于中国中部和西南部，表明受冷高压影响，天气稳定、干燥，夜间可能出现较强的辐射降温。冷锋：东北地区的冷锋向东南推进，冷空气活动活跃，可能带来降雪或寒潮天气。等比湿区（绿色栅格）：表示比湿达到 15g/kg，东南沿海和南海地区湿度较高，易于形成降水云系，需关注未来的降水变化。

总结：当前天气形势表明，中国北方和东北地区将受冷空气影响，出现降温降雪天气；中部和西部天气相对稳定；而东南沿海和南海将受到大风影响，局部可能出现阵雨。建议关注冷空气和强风的持续发展及其可能带来的影响。

图 1 利用 ChatGPT-4o 诊断分析东亚地面天气分析图的示例

Fig. 1 An example of diagnostic analysis of an East Asian surface synoptic chart using ChatGPT-4o.

图 1 给出使用最新多模态模型 ChatGPT-4o 诊断分析中央气象台发布的地面天气图的示例。模型很好地完成了地面分析图的判识，同时结合自身的知识进

行了详细的推理和总结。由图 1 可见：模型已经具备了基础的气象知识，正确判断了天气图类型和日期；对高/低压中心、锋面等天气系统的分布位置大部分进行了正确的识别，并对系统移动给出了合理推算；对预期天气条件也给出了正确的描述，重要天气特征解释和总结也接近于初级天气预报员水平。同时也可看到，模型对于一些细节的描述存在错误（图中下划线的文字）。例如：认为等压线在东北地区分布较密，与东北地区的低压相矛盾；东南部沿海地区气压梯度较小，不符合有大风的实际情况；高压中心位于中国中部和西南部，对地理位置判断不正确，导致对于西南地区的降水没有描述出来等。当然，实际业务中预报员是通过多种资料综合分析，需要进一步结合实际场景以测试模型在多源资料综合诊断分析方面的能力。

相较于传统依赖人工经验的诊断分析方法，基于 LLM 的气象诊断分析展现出自动化、高效率和综合分析的优势。虽然当前 LLM 在一些细节判断上仍有不足，但其快速提取关键信息并进行综合分析的能力已接近初级预报员水平，为预报流程的智能化提供了可行路径。未来结合更多专业气象知识的微调和 RAG 技术，有望进一步提升 LLM 的诊断准确性。

2.4 天气预报工具调用

在传统天气预报业务中，预报员依赖专业的天气预报系统进行分析和预报产品制作。这些系统作为人机交互平台，既帮助预报员全面、准确地掌握大气三维结构和运动情况（高嵩等, 2017），又支持预报产品制作（贺雅楠等, 2018）和天气个例分析等研究工作（孙卓等, 2020）。随着气象现代化的快速发展，预报员在天气预报中的核心地位依然未变（代刊等, 2016），但传统的交互方式已无法高效应对海量数据和复杂的预报需求。

LLM 的引入为天气预报工具带来了交互方式的革新，相较于传统图形用户界面（GUI），LLM 通过自然语言交互实现智能化工具调用，显著提高了用户体验和工作效率。用户只需输入简单的自然语言指令，便能依托 LLM 的强大意图识别和工具调用能力完成复杂的数据加载、分析和预报产品加工任务，实时反馈机制可根据需求动态调整分析流程，灵活适应不同场景；同时，会话式交互降低了对专业工具和复杂界面的依赖，使预报员能够专注于核心分析工作，避免频繁

切换业务平台和界面；基于 LLM 驱动的“智能体”功能使预报产品生成与发布高度自动化，简化传统流程并提升效率。例如，预报员只需输入“北京过去 24 小时温度变化曲线”，LLM 便自动解析需求并调用相关工具完成数据加载、处理和结果输出，可显著提升数据调用的效率和准确性。如图 2 所示，将大语言模型与后台网络服务 API 接口结合，通过大语言模型的工具调用能力与文档总结能力，可实现指定城市的预报信息查询，并以自然语言显示结果。



图 2 大语言模型调用工具查询指定城市天气

Fig. 2 A large language model querying the weather of a specified city by calling a tool.

部分研究成果已经展示了 LLM 与工具调用带来的优势。Li et al (2024a) 设计了一个由 LLM 驱动的高温热浪气候风险管理系统，用户输入的“展示过去五年中该地区的高温热浪事件及其社会影响”等自然语言指令，会被 LLM 编排为一套自动化分析流程：从自动调用数值模式工具进行分析，到生成关于热浪强度、影响范围及社会反馈的综合分析结果，再到最终的可视化呈现，最终实现对热风险的深度分析。此外，LLM 能够整合多模态数据，扩展工具调用能力。GeoLLM (Manvi et al, 2023) 结合地理信息系统 (GIS)，在自然语言指令驱动下完成复杂的地理空间分析任务，如风暴路径模拟和灾害影响评估；OceanGPT (Bi et al, 2023) 平台则利用多模态模型，自动调用工具获取海洋观测与卫星数据，对海洋生物图像、海底地形、卫星遥感图像进行处理，将自然语言转换为代码或命令，控制水下机器人完成探测、样本采集及设备维护任务。

自 2025 年起，模型上下文协议(model context protocol, MCP)等创新协议的

出现，正推动工具调用能力迈向新的发展阶段。在气象领域，MCP 能够建立预报模型与业务系统间的标准化交互机制。通过该协议，预报员将能以统一、无缝的方式访问各类异构气象数据源，从而大幅提升工作效率。

在 LLM 的基础上，智能体(Agent)技术将工具调用推向了新的高度。近期，以 Manus (<https://manus.org/agent>) 为代表的新一代智能体展现了强大的自主性和决策能力。与传统 AI 工具仅提供信息不同，Manus 采用多代理联邦架构，由规划代理、工具代理和验证代理协作完成端到端的任务闭环。这种架构可以很好地适应气象预报的复杂场景。例如，在强对流天气预报中，智能体能够自动监测多源观测数据，识别潜在的强对流天气信号，主动调取相关模式产品和历史案例，并根据综合分析结果，向预报员提供预警建议，整个过程无需人为手动干预。

气象智能体技术超越了简单的工具调用，实现从被动工具调用到主动决策支持的转变，并能在预报员不在场时持续工作，保证预报服务的连续性。目前在气象领域，“人机协作”模式仍占主导——智能体提供建议，最终决策仍需要预报员介入，但随着技术的不断进步，气象智能体的自主决策能力将不断增强。

2.5 天气预报文字生成

天气预报文字生成将复杂的气象数据和预报结论转化为易于理解的自然语言，以满足决策者和公众的需求。在决策服务中，气象信息是政府防灾减灾和经济社会决策的重要依据（王莉萍等, 2024），但天气预报的不确定性和实时调整使得决策材料难以保持一致，在灾害过程中这一问题尤为突出（邵颖斌等, 2023）。在公众服务方面，随着新媒体和信息发布渠道的扩展，对气象服务的时效性、精准性和可读性要求更高（张振涛等, 2014），而现有平台在交互和文字生成方面的局限性，使得快速生成高质量、个性化的气象文字产品成为一大挑战。

LLM 凭借强大的语言生成和上下文理解能力，为天气预报服务产品的文字生成提供了新机遇。它能够结合天气预报结论与历史相似天气个例，快速生成定制化服务材料，并支持基于对话交互的内容调整，以应对预报的不确定性问题。同时，通过微调历史决策服务材料，LLM 可以自动提取高频表达方式和逻辑框架，减少人工编辑带来的不稳定性，并根据最新预报自动生成服务材料初稿，节

省调整时间。借助 RAG 技术，LLM 还能与交通、健康、农业等专业知识深度融合，确保生成内容的准确性，而在科普宣传中，通过优化表达方式，能够提升气象信息的传播效果，借助提示词工程可进一步扩大公众服务的覆盖面和影响力。

多个领域已开始尝试使用 LLM 进行专业材料编写，并取得显著成效。Liang et al (2024) 阐述了基于 LLM 的可控文本生成 (CTG)，提出通过重新训练、微调、强化学习和提示词工程等方法，确保模型在安全、主题和语言风格上的可控性，并制定了 CTG 的评估方法。在法律领域，Cui et al (2023) 通过微调开源模型，构建了 ChatLaw 法律 LLM，整合法律法规、司法解释和真实案例等知识，创建了知识向量库和测试数据集，显著提升了法律团队的工作效率并减少了错误率。

中央气象台尝试使用 LLM 来进行“灾害性天气快报”的自动化生成。“灾害性天气快报”是综合各种实况、预报及预警信息来源，并结合天气影响的决策服务材料，通常需要预报员收集来自暴雨、强对流、台风、海洋、环境、水文气象、农业气象等不同岗位的大量文字、数据信息。借助 LLM，配合提示词工程，可以快速将上述信息整理成摘要文字，同时根据预报信息，结合灾害天气种类与发生地点，生成关注重点与建议。最终结果如图 3 所示。



图 3 基于大语言模型生成灾害预警服务快报

Fig. 3 A disaster early warning bulletin generated by a large language model.

相较于传统基于模板的天气预报文本生成系统，LLM 在文字生成方面具有显著优势。LLM 能基于对上下文的理解，动态调整生成内容的详细程度、专业

性和风格，适应不同受众群体。然而，当前 LLM 在生成内容的准确性和可靠性方面仍需人工审核把关，特别是在关键预警信息传递方面。

3 应用大语言模型的关键要素

将 LLM 有效应用于气象领域需要克服多种挑战和障碍。本节聚焦应用 LLM 的关键要素，如图 4，包含三个核心层次：数据基础层、模型优化层和应用部署层。数据基础层主要解决气象业务语料库构建和基准测试集开发问题；模型优化层重点解决时空理解增强和偏见幻觉消除两大技术难题；应用部署层则关注数据资源、计算资源和开发模式等支撑环境。这三层相互支撑，共同构成气象 LLM 应用的生态体系。



图 4 大语言模型在天气业务领域应用的关键技术架构

Fig. 4 The key technical architecture for the application of large language models in the weather service domain.

3.1 气象业务基础语料库

LLM 的能力是通过学习海量高质量的文本（即“语料库”）而形成的，这些作为模型训练基础的开源语料库主要可分为两种：通用型和专业型。通用型语料用于模型预训练，旨在提升语言理解和生成能力，主要来源于网站、书籍、学术文献及部分程序代码等公开数据，具有规模大、种类多、通用性强的特点，常见的数据集包括 Common Crawl 和 C4 等(Liu et al, 2024b)。专业型语料则侧重于模型在各垂直领域的迁移应用，通常用于模型微调和 RAG 内容生成。研究表明，

微调阶段的语料质量比数量更为关键(Zhou et al, 2023)。在科研领域, Wang et al(2023)发布了数学语料库 MathPile, 旨在提升大模型在数学推理和问答任务中的能力; 在金融领域, Yang et al(2023)收集整理了金融新闻、公司公告、财务报表、市场趋势和学术论文等数据, 构建了金融语料库, 并通过微调和结合 RAG 技术, 显著提升了 LLM 在量化交易和风险管理等金融任务中的表现。为推动 LLM 在天气预报服务中的应用, 中国气象局自 2024 年起启动了气象服务类 LLM“风和”的研发, 并通过清洗和加工气象部门在预报服务、过程总结、教学培训和科研交流中积累的数据, 构建高质量的气象专业语料库, 以支持 LLM 在气象知识检索、预报文本生成和预报辅助决策等领域的应用。

3.2 气象业务基准测试集

基准测试是一种标准化方法, 用于评估和比较系统性能, 特别是 LLM 在不同任务中的表现, 包括自然语言理解、生成、推理和翻译等, 帮助研究人员改进模型并选择适合特定应用场景的模型 (Chang et al, 2024)。LLM 基准测试的目标还包括比较不同模型在相同任务上的表现, 以及判断模型在具体应用场景中的有效性和可靠性。基准测试集和框架是基准测试的两大核心组成部分, 其中基准测试集通过多样化任务评估模型能力, 分为通用任务、特定下游任务和多模态任务测试集。通用任务测试集评估语言模型在文本分类、机器翻译、阅读理解等自然语言处理任务中的表现, 如 GLUE (Wang et al, 2019a) 和 SuperGLUE (Wang et al, 2019b); 特定下游任务测试集则评估模型在专业领域的能力, 如 GAOKAO-Bench (Zhang et al, 2024c); 多模态任务测试集则评估模型在文字与图像处理中的能力, 如 MMBench (Liu et al, 2024a)。基准测试框架的作用是通过使用数据集对模型进行测试, 记录输出结果, 并利用评估指标对测试结果进行计算与对比, 支持开源和闭源模型, 并能集成多个成熟的数据集。该框架还支持对自定义数据集的扩充, 提供灵活的评估与测试环境。

针对气象领域的 LLM 应用, 气象业务基准测试集应包含多类核心数据要素: 基本气象要素 (温度、湿度、气压、风等)、多时空尺度数据 (从临近到季节预测的时间尺度, 从全球到点位的空间尺度)、多模态数据 (文本类观测报告、图形类天气图、数值类格点数据)、特殊气象场景数据 (台风、暴雨等极端事件和

梅雨、季风等季节性过程)，以及针对 LLM 特定应用场景（知识检索、诊断分析、文本生成、工具调用）的任务数据和验证评价指标（准确性、完整性、可用性）。通过这些全面的数据要素，可系统评估 LLM 在气象业务中的应用能力，为模型优化和应用推广提供科学依据。

3.3 对于大气运动的时空理解

时间和空间是大气运动的重要特征，但当前 LLM 在时空理解上的局限性对其在天气预报辅助和预报服务材料生成中的应用带来了显著挑战。尽管 LLM 通过预训练能够学习到丰富的地理知识，但在处理较为少见的地理信息时，其表现仍不理想，且难以精准判断空间相对位置和距离（Gurnee and Tegmark, 2023）。这种局限性可能导致生成的预报服务材料出现“幻觉”现象，影响内容的质量和可信度。为解决这一问题，提升基座模型的地理推理能力、结合 RAG 扩充地理信息知识、以及结合外部工具如 GIS 进行空间计算和规划是三种可行的改进方法。通过设计专门的地理推理任务并利用提示词工程，可以帮助模型更好地完成复杂的地理推理；结合 RAG 技术可弥补模型对最新地理信息的掌握不足，提升输出的准确性；而结合 GIS 工具则能增强模型在空间计算和规划方面的能力，提供更加精准的文字描述。通过这些优化，模型在天气预报辅助中的表现可以得到显著提升。

3.4 气象领域的偏见与幻觉

“偏见”和“幻觉”是 LLM 在实际应用中面临的两大核心挑战。偏见源于训练数据中的不平衡或社会刻板印象，可能导致模型在学习过程中进一步放大这些倾向（徐磊等, 2024）；而幻觉则是模型生成看似合理但实际上错误的信息，原因包括数据噪声、不一致性、过度依赖参数化知识和推理阶段的暴露偏差（Ji et al, 2023）。在天气预报服务中，偏见问题主要体现在地理信息的准确性和地区公平性上，尤其是模型可能对欠发达地区的信息产生误差，从而影响极端天气预警的精准性。幻觉问题则可能导致虚构的气象事件或不准确的数据，严重影响气象研究、预警系统和公众信息传播，增加核实成本并降低预报服务效率。为了解决这些问题，学术界和工业界提出了数据清洗与增强、去偏见微调技术以及检索增

强生成等方法,在降低偏见和幻觉方面取得了显著进展(Allam, 2024; Lewis et al, 2020)。在天气预报中,可通过构建多样化的数据集、应用去偏见微调技术、结合实时气象知识库和 RAG 技术,进一步提高模型的公平性和准确性,并通过人机协作和持续监测来确保生成内容的可靠性。

3.5 应用部署的资源需求及开发模式

气象智能化业务应用部署平台是实现大语言模型应用于气象领域的关键支撑环境,其构建需要合理规划数据资源、计算资源并采用先进的软件开发模式,以确保平台的可用性和可持续发展。在数据资源方面,在中国气象局已建成气象大数据云平台“天擎”基础上,需进一步构建多源异构数据融合体系,除了气象观测数据、数值预报数据、历史预报案例等之外,需要建立气象专业语料库或知识图谱,确保大语言模型能够获得专业领域信息并支持知识推理。此外,大语言模型对计算能力要求高,传统 CPU 难以满足需求,因此平台应配置 GPU 加速计算集群;同时,平台应构建支持大规模分布式推理的计算架构,采用低延迟推理引擎降低响应时间;利用加速计算技术,结合弹性资源调度,可根据预报业务需求动态调整资源分配,提高利用效率。现代软件开发模式对平台的稳定性和迭代速度至关重要。气象智能化业务平台应采用 DevOps 理念,消除开发和运维壁垒,实现代码提交、测试、构建、部署的自动化。同时,采用微服务架构将平台拆分为多个独立部署、松耦合的服务,便于并行开发和维护;使用容器技术封装应用及其依赖环境,实现标准化交付和弹性扩缩容。

4 结论

本研究系统探讨了 LLM 在天气预报中的应用潜力及其面临的挑战。LLM 在气象知识检索、预报模型基座、诊断分析、工具调用及文字生成等多个场景中展现出显著应用潜力。通过高效处理海量气象数据、整合多源信息、生成定制化预报产品,LLM 为预报员提供了强大的辅助工具,可以提升天气预报的精准度和业务智能化水平。特别是在极端天气事件的预测和复杂气象图表的分析中,LLM 表现出较强的潜力,能够辅助预报员快速提取关键信息,优化决策流程。

然而，LLM 在气象领域的广泛应用仍面临诸多挑战。首先，LLM 在时空理解上的局限性影响了其在天气预报中的应用，尤其是在处理复杂的地理信息和空间关系时，模型的表现尚不理想。其次，LLM 在训练和应用过程中可能产生偏见和幻觉，导致生成内容的不准确或不公平。这些问题需要通过数据清洗、去偏见微调、检索增强生成等技术加以改进，以确保生成内容的准确性和可靠性。

为推进 LLM 在气象领域的应用，构建高质量的气象语料库和优化基准测试框架是关键。高质量的气象语料库能够为 LLM 提供丰富的训练数据，支持其在气象知识检索、预报文本生成等任务中的应用。同时，通过标准化基准测试，可以评估 LLM 在不同气象任务中的表现，帮助选择适合特定应用场景的模型，并推动模型的持续优化。此外，结合外部工具如 MICAPS、天擎数据平台等，能够进一步提升 LLM 的推理和任务执行能力，增强其在气象业务中的自动化能力。

随着 DeepSeek 等开源深度推理大模型的快速发展，气象领域的 LLM 应用迎来新的发展机遇。这些模型在推理能力、跨领域知识整合和语言生成等方面的优势，与气象领域对复杂数据分析和精准预报的需求高度契合。未来，通过领域专家与 AI 研究者的深度合作，构建气象专用大语言模型框架并优化其在业务系统中的部署应用，将显著提升天气预报的智能化水平，为气象现代化发展注入新动力。

参考文献

- 代刊, 曹勇, 钱奇峰, 等, 2016. 中短期数字化天气预报技术现状及趋势[J]. 气象, 42(12): 1445-1455. Dai K, Cao Y, Qian Q F, et al, 2016. Situation and tendency of operational technologies in short- and medium-range weather forecast[J]. Meteor Mon, 42(12): 1445-1455 (in Chinese).
- 高嵩, 毕宝贵, 李月安, 等, 2017. MICAPS4 预报业务系统建设进展与未来发展[J]. 应用气象学报, 28(5): 513-531. Gao S, Bi B G, Li Y A, et al, 2017. Implementation and development plan of MICAPS4[J]. J Appl Meteor Sci, 28(5): 513-531 (in Chinese).
- 贺雅楠, 高嵩, 薛峰, 等, 2018. 基于 MICAPS4 的智能网格预报平台设计与实现[J]. 应用气象学报, 29(1): 13-24. He Y N, Gao S, Xue F, et al, Design and Implementation of Intelligent Grid Forecasting Platform Based on MICAPS4[J]. J Appl Meteor Sci, 29(1): 13-24 (in Chinese).
- 邵颖斌, 江晓南, 毕潇潇, 2023. "两难"困境下决策气象服务的策略探析[J]. 中国防汛抗旱, 33(7): 74-78. Shao Y B, Jiang X N, Bi X X, 2023. Study on strategies of meteorological decision-making service in a paradoxical dilemma[J]. China Flood Drought Manag, 33(7): 74-78 (in Chinese).
- 孙卓, 高嵩, 李江波, 等, 2020. 基于 MICAPS4 的灾害性天气个例检索与训练系统设计与应用[J]. 暴雨灾害, 39(2): 207-212. Sun Z, Gao S, Li J B, et al, 2020. Design and application of disastrous weather case retrieval and training system based on MICAPS4[J]. Torr Rain Dis, 39(2): 207-212 (in Chinese).
- 唐健, 代刊, 宗志平, 等, 2018. 主客观融合定量降水预报方法及平台实现[J]. 气象, 44(8): 1020-1032. Tang J, Dai K, Zong Z P, et al, 2018. Methods and platform realization of the national QPF master blender[J]. Meteor Mon, 44(8): 1020-1032 (in Chinese).
- 王莉萍, 章建成, 梁科, 等, 2024. 国家级决策气象服务满意度指标及应用研究[J]. 武汉理工大学学报(信息与管理工程版), 46(3): 503-509. Wang L P, Zhang J C, Liang K, et al, 2024. Research on satisfaction index and application of national decision-making meteorological service[J]. J Wuhan Univ Technol (Informat Manag Eng), 46(3): 503-509 (in Chinese).
- 徐磊, 胡亚豪, 潘志松, 2024. 针对大语言模型的偏见性研究综述[J]. 计算机应用研究, 41(10): 2881-2892. Xu L, Hu Y H, Pan Z S, 2024. Review of biased research on large language model[J]. Appl Res Comp, 41(10): 2881-2892 (in Chinese).
- 张振涛, 张正文, 陈宇, 等, 2014. 基于天气事件的公共气象服务产品制作系统[J]. 应用气象学报, 25(2): 249-256. Zhang Z T, Zhang Z W, Chen Y, et al, 2014. An event-based public meteorological service product system[J]. J Appl Meteor Sci, 25(2): 24-256 (in Chinese).
- Allam A, 2024. BiasDPO: mitigating bias in language models through direct preference optimization[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: Association for

Computational Linguistics.

- Bauer P, Thorpe A, Brunet G, 2015. The quiet revolution of numerical weather prediction[J]. *Nature*, 525(7567): 47-55.
- Bi K F, Xie L X, Zhang H H, et al, 2023. Accurate medium-range global weather forecasting with 3D neural networks[J]. *Nature*, 619(7970): 533-538.
- Bi Z, Zhang N Y, Xue Y D, et al, 2024. OCEANGPT: a large language model for ocean science tasks [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok: Association for Computational Linguistics.
- Bodnar C, Bruinsma W P, Lucic A, et al, 2025. A Foundation model for the earth system[J]. *Nature*, 641(8065): 1180-1187.
- Brown T B, Mann B, Ryder N, et al, 2020. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc.
- Busch F, Hoffmann L, Rueger C, et al, 2025. Current applications and challenges in large language models for patient care: a systematic review[J]. *Commun Med*, 5(1): 26.
- Chang Y P, Wang X, Wang J D, et al, 2024. A survey on evaluation of large language models[J]. *ACM Trans Intell Syst Technol*, 15(3): 1-45.
- Chen K, Han T, Gong J C, et al, 2023a. FengWu: pushing the skillful global medium-range weather forecast beyond 10 days lead [J]. arXiv preprint arXiv: 2304.02948.
- Chen L, Zhong X H, Zhang F, et al, 2023b. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast[J]. *npj Climate Atmos Sci*, 6(1): 190.
- Chen S C, Long G D, Jiang J, et al, 2023c. Foundation models for weather and climate data understanding: a comprehensive survey[J]. arXiv preprint arXiv: 2312.03014.
- Cui J X, Ning M N, Li Z J, et al, 2023. Chatlaw: multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model[J]. arXiv preprint arXiv: 2306.16092.
- Deléang G, Ruoss A, Duquenne P A, et al, 2023. Language modeling is compression[J]. arXiv preprint arXiv: 2309.10668.
- Dhuliawala S, Komeili M, Xu J, et al, 2023. Chain-of-verification reduces hallucination in large language models[J]. arXiv preprint arXiv: 2309.11495.
- Fard B J, Hasan S A, Bell J E, 2022. CliMedBERT: a pre-trained language model for climate and health-related text[J]. arXiv preprint arXiv: 2212.00689.
- Fernandez A, Dube S, 2023. Core Building Blocks: next gen geo spatial GPT application[J]. arXiv preprint arXiv: 2310.11029.
- Frey N C, Soklaski R, Axelrod S, et al, 2023. Neural scaling of deep chemical models[J]. *Nat Mach Intell*, 5(11): 1297-1305.
- Gao Y F, Xiong Y, Gao X Y, et al, 2023. Retrieval-augmented generation for large language models: a survey[J]. arXiv preprint arXiv: 2312.10997.
- Gurnee W, Tegmark M, 2023. Language models represent space and time[J]. arXiv

- preprint arXiv: 2310.02207.
- Han Z Y, Gao C, Liu J Y, et al, 2024. Parameter-efficient fine-tuning for large models: a comprehensive survey[J]. arXiv preprint arXiv: 2403.14608.
- Houlsby N, Giurgiu A, Jastrzebski S, et al, 2019. Parameter-efficient transfer learning for NLP[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach.
- Howard J, Ruder S, 2018. Universal language model fine-tuning for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics.
- Hu E J, Shen Y L, Wallis P, et al, 2021. LoRA: low-rank adaptation of large language models[J]. arXiv preprint arXiv: 2106.09685.
- Huang K H, Zhou M Y, Chan H P, et al, 2023. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning[C]//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: Association for Computational Linguistics.
- Izcard G, Grave E, 2020. Leveraging passage retrieval with generative models for open domain question answering[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics.
- Ji S X, Pan S R, Cambria E, et al, 2022. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Trans Neural Netw Learn Syst, 33(2): 494-514.
- Ji Z W, Lee N, Frieske R, et al, 2023. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 55(12): 248.
- Kaplan J, McCandlish S, Henighan T, et al, 2020. Scaling laws for neural language models[J]. arXiv preprint arXiv: 2001.08361.
- Kuckreja K, Danish M S, Naseer M, et al, 2024. GeoChat: grounded large vision-language model for remote sensing[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE.
- Lawson J R, Trujillo-Falcón J E, Schultz D M, et al, 2025. Pixels and predictions: potential of GPT-4V in meteorological imagery analysis and forecast communication[J]. Artif Intell Earth Syst, 4(1): 240029.
- Lessig C, Luise I, Gong B, et al, 2023. AtmoRep: a stochastic model of atmosphere dynamics using large scale representation learning[J]. arXiv preprint arXiv: 2308.13280.
- Lester B, Al-Rfou R, Constant N, 2021. The power of scale for parameter-efficient prompt tuning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Lewis P, Perez E, Piktus A, et al, 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the 34th International Conference on Neural Information Processing System. Vancouver: Curran Associates Inc.

-
- Li H B, Kam-Kwai W, Luo Y, et al, 2024a. Save It for the "Hot" day: an LLM-Empowered visual Analytics system for heat risk management[J]. arXiv preprint arXiv: 2406.03317.
- Li H B, Wang Z W, Wang J C, et al, 2024b. CLLMate: a Multimodal Benchmark for weather and climate events forecasting[J]. arXiv preprint arXiv: 2409.19058.
- Liang X, Wang H Y, Wang Y, et al, 2024. Controllable text generation for large language models: a survey[J]. arXiv preprint arXiv:2408.12599.
- Liu Y, Duan H D, Zhang Y H, et al, 2024a. MMBench: is your multi-modal model an all-around player?[C]//Proceedings of the 18th European Conference on Computer Vision. Milan: Springer.
- Liu Y, Cao J H, Liu C Y, et al, 2024b. Datasets for large language models: a comprehensive survey[J]. arXiv preprint arXiv: 2402.18041.
- Manvi R, Khanna S, Mai G C, et al, 2023. GeoLLM: extracting Geospatial knowledge from large language models[J]. arXiv preprint arXiv: 2310.06213.
- Maynez J, Narayan S, Bohnet B, et al, 2020. On Faithfulness and factuality in abstractive summarization[J].arXiv preprint, arXiv: 2005.00661.
- OpenAI, Achiam J, Adler S, et al. 2023. Gpt-4 technical report[J]. arXiv preprint, arXiv: 2303.08774.
- Pagano T C, Casati B, Landman S, et al, 2022. Challenges of operational weather forecast verification and evaluation[J]. Bull Amer Meteor Soc, 105(4): E789-E802.
- Peng B C, Zhu Y, Liu Y C, et al, 2024. Graph retrieval-augmented generation: a survey[J]. arXiv preprint arXiv: 2408.08921.
- Sahoo P, Singh A K, Saha S, et al, 2024. A systematic survey of prompt engineering in large language models: techniques and applications[J]. arXiv preprint arXiv: 2402.07927.
- Schick T, Dwivedi-Yu J, Dessì R, et al, 2023. Toolformer: language models can teach themselves to use tools[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc.
- Seek-AI D, Guo D Y, Yang D J, et al, 2025. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[J]. arXiv preprint arXiv: 2501.12948.
- Shuster K, Poff S, Chen M Y, et al, 2021. Retrieval augmentation reduces hallucination in conversation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Singhal K, Azizi S, Tu T, et al, 2023. Large language models encode clinical knowledge[J]. Nature, 620(7972): 172-180.
- Stuart N A, Hartfield G, Schultz D M, et al, 2022. The evolving role of humans in weather prediction and communication[J]. Bull Amer Meteor Soc, 103(8): E1720-E1746.
- Vaghefi S A, Stambach D, Muccione V, et al, 2023. ChatClimate: Grounding conversational AI in climate science[J]. Commun Earth Environ, 4(1): 480.
- Vaswani A, Shazeer N, Parmar N, et al, 2017. Attention is all you

- need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 5998-6008.
- Wang A, Pruksachatkun Y, Nangia N, et al, 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. , 2019. Red Hook: Curran Associates Inc.
- Wang, A, Singh, A, Michael, J, et al, 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding[C]//2018 EMNLP Workshop BlackboxNLP: analyzing and Interpreting Neural Networks for NLP. Brussels: Association for Computational Linguistics.
- Wang S Q, Hu T, Xiao H, et al, 2024. GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review[J]. *Int J Digit Earth*, 17(1): 2353122.
- Wang Y, Karimi H A, 2024. Exploring large language models for climate forecasting[J]. *arXiv preprint arXiv: 2411.13724*.
- Wang Z Z, Li X F, Xia R, et al, 2023. MathPile: a billion-token-scale pretraining corpus for math[J]. *arXiv preprint arXiv:2312.17120*.
- Wei J, Tay Y, Bommasani R, et al, 2022a. Emergent abilities of large language models[J]. *arXiv preprint arXiv: 2206.07682*.
- Wei J, Wang X Z, Schuurmans D, et al, 2022b. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc..
- Xi Z H, Chen W X, Guo X, et al, 2025. The rise and potential of large language model based agents: a survey[J]. *Science China Information Sciences*, 68(2): 121101.
- Xiong G Z, Jin Q, Wang X, et al, 2025. Improving retrieval-augmented generation in medicine with iterative follow-up questions[J]. *Pac Symp Biocomput*, 30: 199-214.
- Yang H Y, Liu X Y, Wang C, 2023. FinGPT: open-source financial large language models[J]. *arXiv preprint arXiv: 2306.06031*.
- Yao S Y, Zhao J, Yu D, et al, 2022. ReAct: synergizing reasoning and acting in language models[J]. *arXiv preprint arXiv: 2210.03629*.
- Yin S K, Fu C Y, Zhao S R, et al, 2024. A survey on Multimodal large language models[J]. *National Science Review*, 11(12): nwae403.
- Zhang Q, Ding K Y, Lü T W, et al, 2025. Scientific large language models: a survey on biological & chemical domains[J]. *ACM Comput Surv*, 57(6): 1-38.
- Zhang W, Cai M X, Zhang T, et al, 2024b. EarthGPT: a universal multimodal large language model for multi sensor image comprehension in remote sensing domain[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5917820.
- Zhang X T, Li C Y, Zong Y, et al, 2024c. Evaluating the performance of large language models on GAOKAO benchmark[J]. *arXiv preprint arXiv: 2305.12474*.
- Zhang Y, Chen X S, Jin B W, et al, 2024a. A comprehensive survey of scientific large language models and their applications in scientific discovery[C]//Proceedings of

the 2024 Conference on Empirical Methods in Natural Language Processing.
Miami: Association for Computational Linguistics.

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

Zhou C T, Liu P F, Xu P X, et al, 2023. LIMA: less is more for alignment[J].arXiv preprint arXiv:2305.11206.

