国家科技图书文献中心
NATIONAL SCIENCE AND TECHNOLOGY LIBRARY

中国科学院文献情报中心
NATIONAL SCIENCE LIBRARY , CHINESE ACADEMY OF SCIENCES

ChatGPT
对文献情报工作的影响（简版）

报告得到如下项目的支持：

- 国家重点研发计划项目：科技文献内容深度挖掘及智能分析关键技术和软件
- 国家社会科学基金重大项目：大数据驱动的科技文献语义评价体系研究

2023年2月21日

ChatGPT

对文献情报工作的影响（简版）

报告撰写者：张智雄、钱力、谢靖、常志军、刘熠、于改红、胡懋地、汪璐、李雪思、赵旸、王宇飞、王猛、林歆、张梦婷、黎洋、张琴、王雅娇、管铮懿、孟旭、吴欣雨、曹晓丽、谢子纯、李西雨、时慧敏、王倩、许钦亚、杜悦、范萌

ChatGPT 对文献情报工作的影响
（简版）

1 ChatGPT 是什么？

ChatGPT（Chat Generative Pre-training Transformer，生成型预训练转换程序）发布于 2022 年 11 月 30 日，是由 OpenAI 公司研发的人工智能对话系统。由于其能在诸多知识领域中给出清晰、详尽的答案，甚至写出接近真人撰写的文章，自推出后便迅速获得关注。

ChatGPT 是什么，可以从以下五个方面来把握。

（1）**ChatGPT 的对外表现是一个聊天机器人**。它能够通过学习和理解人类语言来与人进行对话，具有依据对话的上下文环境来回答问题的能力，就像人一样来与人类进行聊天交流。

（2）**ChatGPT 的实际本质是人工智能生成技术**。它是人工智能内容生成（Artificial Intelligence Generate Content，AIGC）技术的具体应用。它在学习人类语言和相关领域知识的基础之上，具有了智能化的内容创作能力，能够自动生成特定的内容。

（3）**ChatGPT 的关键基础是生成式大规模语言模型**。即基于生成式预训练的变换器（Generative Pre-trained Transformer，GPT），它以生成式的自监督学习为基础，从 TB 级训练数据中学习隐含的语言规律和模式，训练出的千亿级别参数量的大规模语言模型。

（4）**ChatGPT 的核心技术是 InstructGPT**。它采用了基于人类反馈的强化学习(Reinforcement Learning with Human Feedback，RLHF)，让人工智能模型的产出和人类的常识、认知、需求、价值观保持一致。

（5）**ChatGPT 的主要特点是与前期类似产品相比，编造事实大幅下降，生成的毒内容更少**。它在一定程度上解决了传统语言模型在复杂多领域的知识利用、演绎推理、欺骗性反应等方面的缺陷，使回答更具有用性和真实性，具有编造事实大幅下降，生成的模仿性谎言(imitative falsehoods)、毒内容（toxic output）更少的重要特征。

ChatGPT 的核心技术体系如图 1 所示。

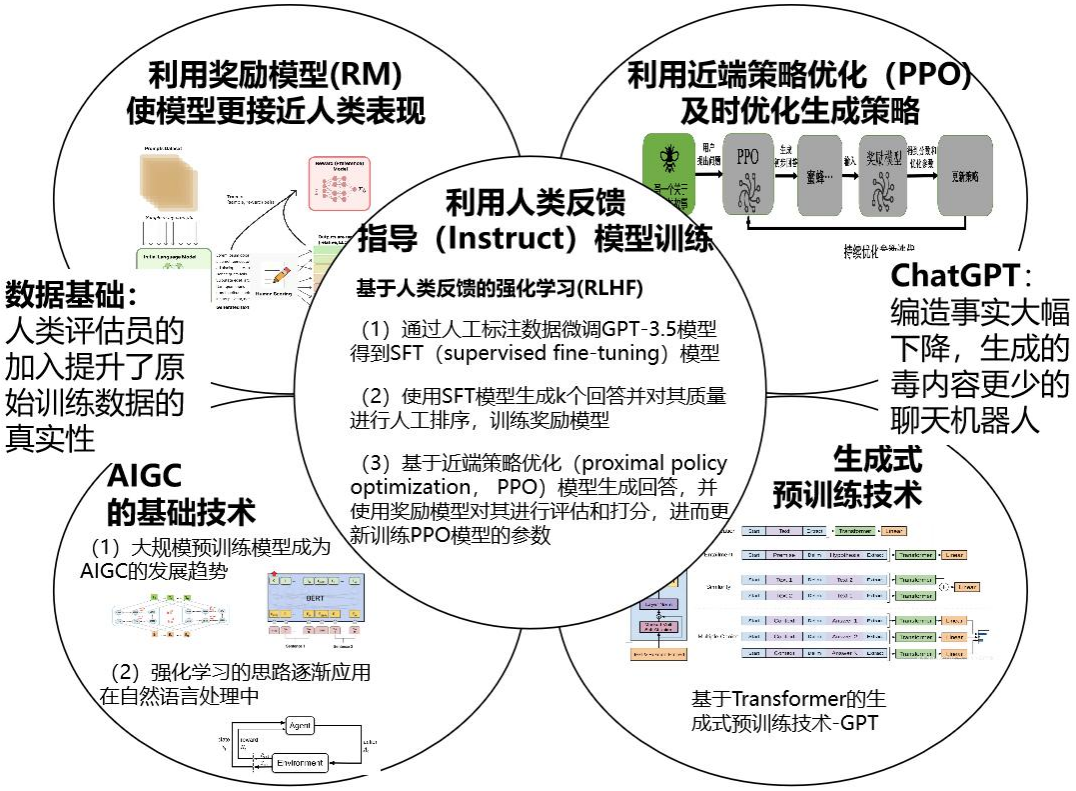

图 1 ChatGPT 的核心技术体系

2 人工智能技术迅速发展给文献情报工作的启示

以 ChatGPT 为代表的人工智能技术近年取得飞速突破，相关成果广泛应用于各个领域，对社会各行业都产生了巨大冲击。

总结近十余年来人工智能的主要发展历程，我们认为人工智能取得突飞猛进的根本原因在于计算机学习知识、开发利用知识的模式已经改变。能够从各类数据资源（语料）中快速高效地学习到隐藏于这些数据资源中的知识是 AI 飞速突破的本质所在。

以 ChatGPT 为代表的 AI 技术的迅速发展源于知识学习能力的大幅提升。它带给文献情报机构提升知识学习有以下启示：

（1）**计算机解决问题模式已改变，机器学习成为获取解决问题所需知识的重要手段**。机器学习改变了计算机解决问题的模式。原来是人输入知识让机器解决问题；而现在，是让机器从相关语料中学习知识，再让机器利用学习到的知识去解决相关的问题。在这一过程中，隐藏着重要人类知识的大样本训练语料至关

重要。这些以语料为表征的人类知识才是机器学习取得飞速突破的关键。

（2）**深度学习的性能提升，除模型突破之外，更要归功于语料和算力**。各类深度学习模型是基础，大量可计算数据资源（训练语料）是前提，大规模计算能力是催化剂。拥有大样本训练语料和大规模计算能力，使得基于人工神经网络深度学习的知识学习性能大幅提升。

（3）**自然语言处理的技术已经重写，无监督的预训练对于知识学习有重要价值**。基于预训练（Pre-Training）和微调（Fine-Tuning）的两阶段学习方法，改写了自然语言处理（NLP）方式，无监督的预训练具有重要价值。利用大规模非标注语料的无监督的预训练，能够使模型从语料中学习到语言表达模式、文字前后逻辑、知识元间关系等知识内容，提高了模型的泛化能力和鲁棒性。在此基础上只用少量标注语料进行微调，即可在特定下游任务中取得较好的效果。

（4）**ChatGPT 并不是无来由的横空出世，而是学习能力从量变到质变的重大突破**。回顾历史来看，从最初的 1.17 亿参数、5GB 语料、12 层 Transformer 的 GPT-1 模型，到目前的 1750 亿参数、45TB 语料、96 层 Transformer、采用人类反馈强化学习的 ChatGPT 模型。ChatGPT 是语料、模型、算法，通过迭代训练不断积累而成的。人工智能知识学习能力上，每一个小小的进步都是有价值的，久久为功，不断进步，最终实现了从量变到质变的转换。

（5）**ChatGPT 是集成创新的成果，学习能力的提升得益于软件、硬件、技术、语料的有效集成**。为了适应模型参数量的激增，OpenAI 收集、标注了更多的原始训练语料；为了实现更贴近人类的对话效果，研发了基于人类反馈的强化学习方法；为了加速模型训练，部署了 28 万个 CPU 内核、1 万个 GPU 的超级计算机。通过软件、硬件、技术、语料有效的集成，才使得 ChatGPT 的知识学习能力获得质的飞跃，造就了当前 ChatGPT 出色的表现。

3 ChatGPT 对文献情报工作的影响

（1）**改变文献情报数据组织方式，从表面信息组织到语义内容组织**。科技文献情报原始的组织方式往往以题目、摘要、关键词、机构、期刊等表面信息组织为主，较少深入到文献内容中。随着 AI 技术发展，从科技文献中精确挖掘细粒度知识对象的能力得到提升，以科技文献中研究问题、研究方法、实验步骤、数据资料等深入到文献内容的语义内容组织成为可能。

（2）**改变文献情报知识服务的模式，从信息检索到知识问答**。目前，文献情报知识获取服务主要基于文献元数据，通过元数据索引实现对海量科技文献数据的检索与获取。ChatGPT 等技术可以在语义层面理解论文内容，识别结构化细粒度知识元，形成大规模知识网络。ChatGPT 推动了从索引式信息检索方式向问答式知识应答方式的转变。在未来，或许能够实现一种新型的知识问答服务，即用户向智能知识服务平台提问后，平台能够直接生成该问题的答案，并给出答案的相关证据链。

（3）**改变文献情报分析方法，从手工作坊到大规模智能分析。**文献情报分析过程包括数据准备、统计分析、观点提炼以及报告撰写等一系列复杂工作，往往由人类手工完成。类 ChatGPT 人工智能技术已具有观点提炼、内容综述、场景问答、语言翻译、语义分析、智能推荐、辅助决策的潜在能力，可以为情报分析人员提供智能化工具，辅助文献情报分析工作。

（4）**带来文献情报服务安全问题，须建立风险管控机制**。泛知识化大模型不能保证回答质量，而文献情报领域对数据可信度具有更高的要求，基于伪数据、伪造事实生成的情报报告必然是不可信的。掌握智能服务的数据控制权是做好应用的重中之重，同时建立完善的数据循证体系，附加数据证据链、数据来源详情，实现对风险的有效管控和溯源。

（5）**对用户阅读习惯的影响，引导人机协同阅读新模式**。类 ChatGPT 技术可能对用户阅读文献资源的方式带来颠覆性的影响。用户输入待读文献资源，智能技术自动实现知识抽取、关系揭示，通过可视化方式进行展现，支持多维度的语义分析，并以交互式的方式应答用户的问题和设定，形成用户与人工智能协同阅读的新模式。

（6）**对传统图书情报工作形成挑战，需要统筹谋划图书馆的队伍能力与岗位体系**。从基础的书目录入、客服解答、代码撰写到资讯编辑、热点论文推荐、动态感知、情报分析等都会在不同程度上受到人工智能技术的影响，一部分“重复性、技术含量低”的工作将被人工智能优化或替代。同时，人工智能也带来了新的工作机会，更多智能服务的工作需要设置新的岗位，形成新的业务方向，扩展图书情报工作的业务范围。

4 对文献情报领域的建议

ChatGPT 重在内容生成，而文献情报工作重在循证。ChatGPT 主要解决自然语言处理中内容生成的问题，但文献情报工作的重点并不在此，我们的机会在于如何循证，挖掘支撑可信情报的证据及证据链。文献情报工作在 AI 时代要找到自己的不同价值取向,有关建议如下:

（1）**文献情报领域要把从科技文献内容中挖掘和利用知识的能力作为核心能力来建设**。科技文献蕴含人类知识、表达科学机理、揭示科研成果，是国家科技创新的核心战略资源。ChatGPT 利用智能技术从海量文本数据中对知识挖掘与利用的巨大成功告诉我们，鉴于科技文献的重要价值，文献情报领域要将从科技文献内容中实现知识挖掘和利用作为核心能力建设。

（2）**充分认识到文献情报机构在 AI 时代的优势和价值**。语料是人工智能获取知识的源泉，高价值语料工作是一切人工智能的基础。科技文献蕴含大量知识，文献情报机构应充分认识自己在新时代的使命和定位：AI 语料提供者，做好“语料”基础工作。

（3）**充分加强人工智能新技术方法的研究和应用**。BERT、ChatGPT 等人工智能新技术方法突破，表明一代又一代的 AI 技术还在突飞猛进。文献情报领域不能浅尝辄止。

（4）**文献情报领域需积极参与“专业和垂直”知识系统建设**。ChatGPT 开启了一个新模式，带来了强大的综合性问答系统，而针对科学领域，开展更加深入的专业化知识内容获取与分析的技术方法研究，还存在很多可以开拓的空间。我们需要利用自身专业领域的文献情报优势，积极参与“专业和垂直”知识系统建设。

（5）**文献情报领域要努力创新知识服务模式**。ChatGPT 让我们看到检索和问答已经相互交融，文献情报不能还仅仅停留在检索之上，要充分利用新思路、新技术、新模式、新方法支持知识服务应用。例如，面向知识获取场景的问答式知识检索，面向阅读辅助场景的科技文献集的自动综述等。

（6）**应用 ChatGPT 在情报研究工作上启发创意**。ChatGPT 能够通过简单提示进行具有创意的创作。在情报研究工作上，用之来启发创意可能是一个不错的选择，但需要专家来指导。

（7）**情报的溯源和真实可靠性检测将变得更加重要**。当很多“情报”由 ChatGPT 生成之后，情报的溯源和真实可靠性检测将变得更加重要。避免“滥用

ChatGPT”带来错误虚假信息传播、信息泄露、抄袭等一系列问题。

（8）**要进行数据资源、基础设施、智能技术一体化的能力建设**。ChatGPT这样真正实现应用的AI产品，是软硬件以及各种技术方法有效集成的结果。文献情报工作能力的提升，要统筹数据资源的积累、基础设施的升级、智能技术的研究等，实现各方面从量的积累到质的飞跃，最后进行一体化的有效集成，开发出真正好用、耐用、用户愿意用的文献情报产品。

文献情报需要自我革新，拥抱新技术与新机会。ChatGPT作为一种工具，它本身不会打败人。但是它肯定会带来：会使用这种工具的人打败那些不会使用这种工具的人。传统的文献情报工作依然有价值，但新技术带来改变已是大势所趋。在此背景之下，文献情报领域需要守正创新，图书情报研究必须把握机遇，既要守正继承传统科学研究范式，也要拓展以ChatGPT等新技术助力科学研究。

由于撰稿人能力水平有限，如有不当之处，欢迎大家批评指正！若需要详细报告，请联系我们。联系人：刘熠，liuyi@mail.las.ac.cn

国家科技图书文献中心
NATIONAL SCIENCE AND TECHNOLOGY LIBRARY

中国科学院文献情报中心
NATIONAL SCIENCE LIBRARY , CHINESE ACADEMY OF SCIENCES