第 51 卷第 1 期
2024 年 1 月
Vol. 51, No. 1
January, 2024

四川师范大学学报(社会科学版)
Journal of Sichuan Normal University
(Social Sciences Edition)

# 论大语言模型材料的证据属性
## ——以 ChatGPT 和文心一言为例

徐继敏　严若冰

**摘要**：以 ChatGPT 和文心一言为代表的大语言模型产生海量大语言模型材料，此类材料进入社会生活并产生广泛影响，讨论大语言模型材料证据属性具有重要意义。大语言模型材料可以成为证据，但是否具有证据资格需要在具体个案中进行判断。从诉讼效率出发，大语言模型证据在不同诉讼中的呈现形式可以有所区别：一般案件可以仅举示人机交流材料和使用者本地环境信息，重大案件则应完整举示。大语言模型材料区别于大数据证据和一般的人工智能证据，具有直观性强、可解释性弱、偏在于少数技术公司、可识别性弱等特点。

**关键词**：大语言模型；大语言模型材料；大语言模型证据；AIGC

**DOI**：10.13734/j.cnki.1000-5315.2023.0316

**收稿日期**：2023-05-24

**基金项目**：本文系国家社会科学基金项目“在线行政行为证据规则研究”(21XFX004)的阶段性成果。

**作者简介**：徐继敏，男，四川内江人，四川大学法学院教授，E-mail：xujimin168@163.com；
严若冰，女，浙江台州人，四川大学法学院博士研究生。

### 一　新型大语言模型材料必将广泛运用于法治实践

2022 年 11 月，美国公司 OpenAI 推出的预训练生成式通用大语言模型(Large Language Model，LLM)[①]ChatGPT 获得巨大反响。用户通过浏览器访问的 ChatGPT 聊天机器人是基于 ChatGPT 模型形成的产品，但目前较少对两者进行区分[②]。ChatGPT 是单一模态大语言模型，理解和生成信息都限于文本(含代码)形式，而 OpenAI 在 2023 年 3 月发布的 GPT-4 为多模态大语言模型，在理解图像、处理复杂任务等方面有所提升。两者都缺乏对数据截止时间后相关信息的了解，且给出的回答不一定正确，大语言模型的这些错误被称为“幻觉”(hallucination)[③]。3 月 23 日，ChatGPT 允许用户通过添加插件进行联网，实现了信息实时检索等功能，能力得到进一步提升[④]。在国内，百度于 3 月 16 日发布大语言模型“文心一言”[⑤]，阿里于 4 月 7 日开放“通义千问”大语言模型体验申请，商汤科技于 4 月 10 日发布大语言模型产品“商量”，我国的大语言模型也在不断推进。关于大语言模型对文字行业的影响，我们可以从 AI 绘画对绘画行业的影响中窥

①“大语言模型”也称“大型语言模型”或“语言大模型”，均指英语中的 LLM(Large Language Model)，可以作完全相同理解。

②ChatGPT 既是 OpenAI 发布的聊天机器人的名字，也是产生该聊天机器人的大语言模型的名字，即大语言模型与大语言模型产品同名。报道和研究常称产生 ChatGPT 聊天机器人的大语言模型为 GPT-3.5，但根据 OpenAI 官网上的 ChatGPT 常见问题解答(ChatGPT General FAQ)，更准确地说，ChatGPT 模型是以 GPT-3.5 为基础进行微调得到的。2023 年 3 月 14 日，OpenAI 发布了 GPT-4 模型，升级后以 GPT-4 模型为基础的聊天机器人叫作 ChatGPT Plus。

③如 GPT-4 的数据截止时间为 2021 年 9 月，参见：“GPT-4 is OpenAI's most advanced system, producing safer and more useful responses,” OpenAI, updated March 15, 2023, accessed May 22, 2023, https://openai.com/product/gpt-4。

④“ChatGPT plugins,” OpenAI, updated March 23, 2023, accessed May 22, 2023, https://openai.com/blog/chatgpt-plugins.

⑤文心一言可以生成文本、语音、图像和视频等内容，具备多模态能力。根据笔者自 3 月 17 日参加内测的情况，文心一言与 ChatGPT 尚有明显差距，但是可以参与文稿初步生成、文字校对、文本翻译等日常工作。

见一些线索。AI 绘画工具能够基于文字提示生成图像，成本低、效率高，自推出时就受到绘画行业的抵制[①]。笔者通过对国内游戏从业者的访谈了解到，目前 AI 绘画对游戏美术业务的冲击已经显现，游戏公司的美术外包业务甚至内部美术团队都面临着被裁撤的风险，因为绘画行业的抵制行动难以对抗企业降低成本的意愿。

可以预见，大语言模型将广泛深入地参与到社会生活中。首先，大语言模型已受到各科技企业和科研团队的高度重视，国外有谷歌的 Bard、Anthropic 的 Claude、Meta 的 BlenderBot3 等，国内有百度的文心一言、阿里的通义千问、商汤科技的商量、华为的盘古、腾讯的混元、中国科学院自动化研究所的紫东太初等。其次，大语言模型有较成熟的技术和已经落地且取得商业成功的产品，其热度性质与暂时缺少技术铺垫的元宇宙概念不同。再次，大语言模型作为基石性模型（Foundational Model）具有较强拓展潜力，可以和搜索引擎、内容平台等应用相结合，也可以与各种公共服务场景相结合[②]。最后，大语言模型与个人生活工作结合紧密，有潜力成为新的枢纽平台。ChatGPT 允许联网后，用户可以通过它进行订餐、订机票和订酒店。ChatGPT 独特的产品能力和工作性能或可吸引一批用户改变原本的互联网消费习惯，转而以 ChatGPT 作为新的信息处理枢纽，甚至通过路径依赖形成深度绑定。随着大语言模型的铺展，大语言模型材料将大量出现并得到普遍运用。

目前大语言模型产品的主要形式是聊天机器人和搜索引擎，虽然两者都非新兴事物，但是大语言模型相比以往技术的区别存在诸多特殊性，至少包括以下四项：第一，大语言模型材料以生成性的文本（含代码）为主，多模态大语言模型材料还可能包括图片、语音和视频；第二，大语言模型是基石性模型，未来许多产品和功能可以在其基础上搭建，大语言模型材料的形态和运用将非常丰富[③]；第三，大语言模型能通过自然语言与使用者进行交流，甚至能让使用者察觉不到自己在与机器对话，难以意识到大语言模型材料的存在；第四，大语言模型有很强的黑箱效应，而且 ChatGPT、GPT-4 和文心一言等主要大语言模型的训练数据和模型均未公开，使大语言模型及其形成材料的可解释性受到更深质疑。

由上可知，大语言模型材料的广泛出现将带来新的法律问题，而证据问题在其中尤为突出。比如在民事领域，大语言模型如果有不当广告行为，使用者和监管机构如何进行证据提取和审查；在刑事领域，大语言模型可能被利用来进行犯罪活动，公检法机关应如何调查取证和运用证据；在行政领域，大语言模型如果被用于政府建设[④]，可以在多大程度上影响甚至取代行政机关工作人员的行政行为，是否会出现智能行政行为，行政相对人如何在复议、诉讼中运用大语言模型材料等。虽然尚无案例出现，但随着大语言模型研究和运用的深化，大语言模型材料进入证明活动或许是司法实践和学术研究必然要面对的挑战。目前大语言模型在产品应用上还相对保守，主要以聊天机器人和搜索引擎的形式呈现，但大语言模型产品的未来远不止这些：ChatGPT、GPT-4 已经开放联网，使用者可以通过下载插件实现实时检索等多项需求；微软宣布将 GPT-4 集成到 Copilot，今后 Office 系列软件的使用方式将发生巨大变化；超三百家公司宣布与文心一言合作，涵盖媒体、文娱、金融等行业领域。当大语言模型作为一项基础工具与我们日常生活的方方面面结合，它产生的材料将远比本文能够看到的更加复杂。正如电子数据随着互联网和计算机技术进入法律实践，或许大语言模型证据也将占有类似的重要地位。

## 二　大语言模型材料的定义、技术、内容和类型化

### （一）大语言模型材料的定义

大语言模型材料是指与大语言模型使用行为相关的、在证明活动中可能作为证据使用的材料。广义的大语言模型材料包括人机交流材料、大语言模型本体相关材料和大语言模型运行环境信息三个部分；狭义的

①陶凤、李想《“离谱”AI 绘画赚钱不离谱》，《北京商报》2023 年 2 月 2 日，第 4 版。

②於兴中、郑戈、丁晓东《生成式人工智能与法律的六大议题：以 ChatGPT 为例》，《中国法律评论》2023 年第 2 期，第 2 页。

③於兴中、郑戈、丁晓东《生成式人工智能与法律的六大议题：以 ChatGPT 为例》，《中国法律评论》2023 年第 2 期，第 2 页。

④如张效羽认为，GPT 技术的长处与法治政府建设的基本需求具有技术亲和性，未来法治政府建设要为人工智能嵌入行政执法程序做好充分准备。参见：张效羽《ChatGPT 等人工智能内容生成技术对法治政府建设的影响及应对》，《电子政务》2023 年第 4 期，第 12－14 页。

大语言模型材料则仅指由大语言模型生成的材料①。从概念关系上看,大语言模型属于人工智能技术,如果承认大语言模型材料可以成为证据,那么,“大语言模型证据”也属于人工智能证据的范畴。人工智能证据已经受到司法实践和法学研究的关注,可以成为研究大语言模型材料的重要参考。目前实践中还缺少将大语言模型材料作为证据的情形,而且大语言模型材料的具体内容和证据资格问题暂无学术共识或规范依据。为求稳妥,本文主要称“大语言模型材料”而非“大语言模型证据”,不过本文认为大语言模型材料可以成为证据,因此也将适当使用“大语言模型证据”的提法。

(二)大语言模型技术的概况及相关法律规范

自然语言处理(Nature Language Processing,NLP)技术被广泛应用于搜索引擎、智能客服、机器翻译、舆情监测、自动摘要等领域,大语言模型是自然语言处理领域的前沿成果②。ChatGPT、GPT-4、文心一言、通义千问等都属于预训练生成式通用大语言模型,具有通用性、超多参数、生成式等特点。通用性是指模型不局限于某一行业或细分领域,旨在理解和回应常识性、通用性问题。ChatGPT、通义千问是单一模态大语言模型,GPT-4 和文心一言则具备多模态能力。基于通用、多模态的模型特点,大语言模型材料的内容也具有通用性,可能包含文字、图片、语音和视频等多种模态。

模型指通过学习算法(Learing Algorithm)“从数据中学得的结果”③,大模型和小模型以参数量(Parameter Size)为划分标准,ChatGPT 的参数量达到千亿级,GPT-4 的参数量据说更达到兆级④。大模型和知识图谱(Knowledge Graph,KG)是人工智能发展的不同路径⑤,由于 ChatGPT 的成功,许多人认为大语言模型更可能是未来发展的方向,但知识图谱仍然可以作为大语言模型的训练数据或以外接方式发挥作用⑥。知识图谱的成本相对较低,且能够通过直观可视的关系网络呈现,可解释性较强;大语言模型以“大算力+强算法”为重要特征,内部极为复杂,可解释性较差。大语言模型的可解释性弱决定了大语言模型材料容易受到质疑,但这不意味着大语言模型材料缺乏证据法上的可靠性。

大语言模型能够生成新的内容,其生成材料属于人工智能生成内容 AIGC(AI Generated Content)。AIGC 包括 AI 绘画、AI 翻唱、大语言模型写作等,虽然在著作权上存在争议⑦,但是已经得到企业和个人的广泛运用。4 月 11 日,国家网信办就《生成式人工智能服务管理办法(征求意见稿)》公开征求意见,涉及对大语言模型生成材料进行规范。从生成式特点出发,大语言模型材料属于深度合成内容。我国在深度合成治理上走在世界前列,2022 年 11 月出台的《互联网信息服务深度合成管理规定》对深度合成进行了定义和规范。但大语言模型的深度合成能力超越以往算法,给立法提出了新的挑战,要求深度合成立法从算法治理走向人工智能治理⑧。

(三)大语言模型材料的具体内容及其证据价值

---

①由于缺少既有研究,“大语言模型材料”和“大语言模型证据”均为本文提出的概念。

②传统人机交互由人类以机器语言、编程语言实现和机器的交流,而自然语言处理旨在让机器理解和生成人类语言,通过机器对人类语言的适应实现人机交互。自然语言理解(NLU)支持机器理解人类自然语言文本的内容,自然语言生成(NLG)支持机器以自然语言文本的形式输出信息。因此总体来说,自然语言理解和自然语言生成都属于自然语言处理。除自然语言处理外,人工智能领域还有计算机视觉(Computer Vision,CV)、计算机听觉(Computer Audition,CA)等研究方向,与人类的自然感知系统形成对应关系。

③周志华《机器学习》,清华大学出版社 2016 年版,第 1 页。

④目前 ChatGPT 和文心一言参数量均未公布,此处 ChatGPT 和 GPT-4 参数量数据采用美国网站 Semafor 援引 8 名知情人士消息。参见:Reed Albergotti, “The secret history of Elon Musk, Sam Altman, and OpenAI,” Semafor, updated March 25, 2023, accessed May 22, 2023, https://www.semafor.com/article/03/24/2023/the-secret-history-of-elon-musk-sam-altman-and-openai.

⑤大语言模型是给一个有超多参数的模型网络喂海量文本数据进行训练,再对其进行微调;知识图谱则是显式地抽取出实体、关系等信息,构成由节点(Node,表示实体)和边(Edge,表示关系)组成的关系网络。

⑥根据 2023 年 3 月 16 日百度公司文心一言发布会内容,文心一言以 5500 亿事实的知识图谱为训练数据。知识图谱的“事实”指由节点 a,关系 r,节点 b 共同构成的一个三元组。

⑦如“AI 孙燕姿”、“AI 周杰伦”等翻唱作品引发著作权相关讨论。参见:顾敏、陈月飞等《技术迅猛发展,AI 如何“向善”》,《新华日报》2023 年 5 月 16 日,第 5 版;李欣璐《专家:“AI 歌手”或涉嫌多项侵权》,《四川法治报》2023 年 5 月 17 日,第 5 版;刘凡《解码 AI 歌手习艺之道》,《海南日报》2023 年 5 月 22 日,第 B09 版。

⑧张凌寒《深度合成治理的逻辑更新与体系迭代——ChatGPT 等生成型人工智能治理的中国路径》,《法律科学(西北政法大学学报)》2023 年第 3 期,第 38-51 页。

大语言模型材料是复合型材料，可以被分解为三个主要部分：人机交流材料、大语言模型本体相关材料和大语言模型运行环境信息。由于大语言模型本体材料和云计算服务平台环境信息提取收集的难度极高，且对证明一般案件的案件事实作用不大，大语言模型材料在实践中或将主要以人机交流材料和使用者本地环境信息的形式呈现。

1.使用者与大语言模型交流形成的材料

使用者与大语言模型交流的材料（简称为“人机交流材料”）以文本（含代码）为主，可能包含图片、语音和视频，一般载于计算机或者移动设备的网页，呈现为人类使用者与大语言模型一问一答组成的连续性对话。以同一使用者为范围，涉及的人机交流材料可以被分为三个层次。第一，最低层次材料是使用者与大语言模型一问一答形成的“问答”，均由使用者先输入信息或提出问题，再由大语言模型进行回答。第二，中间层次材料是由连续问题组成的“对话”，这是人机交流材料中最重要的单位，适宜成为一份人机交流材料的基础范围。“对话”并不以话题内容和时间间隔为识别标准，而是要考虑大语言模型的“记忆”范围。本文所称“记忆”指大语言模型联系上下文的能力，在“记忆”范围内，如果使用者能够通过恰当的“提示”（Prompt）引导对话，大语言模型将给出更加精确的回答。这种与大模型进行交流、得到更优结果的方法被称为提示工程（Prompt Engineering），已经受到技术和产业领域的肯定和重视①。目前影响大语言模型记忆范围的因素主要有：问答是否属于大语言模型产品中的同一对话框，以及技术公司设置的大语言模型产品记忆上限，如微软的必应就有记忆问答的上限。第三，最高层次材料是同一使用者账号下的所有人机交流材料，它包括同一使用者与大语言模型的所有对话。需要注意的是，账号所有者与大语言模型使用者未必是同一人，需要结合使用者本地环境信息、相关言词证据等进行综合判断。

使用者输入的材料是大语言模型理解的对象。根据 GPT-4 和文心一言的多模态能力，使用者输入大语言模型的材料可以是文本（含代码）和图像，且以文本为主要形式。从证据角度看，使用者输入大语言模型的材料至少具有以下价值。第一，能够反映使用者的主观心态，比如体现使用者对某类信息的需求和认知。刑事案件中犯罪嫌疑人在浏览器留下的相关搜索记录能证明其主观上对犯罪行为的认识和心态，今后或许会有嫌疑人就类似问题向大语言模型产品提问，那么证据也相应地从电子数据转为大语言模型材料。第二，能够反映大语言模型生成的材料是否合理。大语言模型生成的材料应与使用者输入材料有一定对应关系，两者过于不匹配或说明大语言模型生成材料的可靠性较弱。第三，能够反映使用者对大语言模型生成材料的认识。在人机连续对话环境中，使用者输入的材料除了独立表达意思，还是对大语言模型生成材料的回复，因此能反映大语言模型生成材料对使用者的影响，比如使用者是否受到了不正当广告行为的诱导。

大语言模型生成的材料可以被看作是最狭义的大语言模型材料。目前大语言模型可以生成的材料类型包括文本（含代码）、图像、语音甚至视频。本文认为大语言模型生成材料包含了“机器意见”和“人类意见”两种元素，“机器意见”指大语言模型的创新性元素，“人类意见”指使用者的指令、干预、引导元素。“机器意见”的可靠性不强，即使是目前公认性能最优秀的 GPT-4 模型也存在“幻觉”。大语言模型还不能像一些文章中提到的人工智能证据那样——比如人脸识别系统判断照片中人是特定某人——能以机器自身的“意见”发挥证据作用②。大语言模型生成材料可以反映大语言模型服务提供者和使用者的不当甚至违法行为，比如服务提供者通过大语言模型推送不恰当广告信息，使用者利用大语言模型进行诈骗、开展“网络水军”活动等。

2.大语言模型本体材料

与大语言模型本体相关的材料有：第一，用于形成大语言模型且影响大语言模型生成内容的材料，主要包括训练数据和深度学习算法；第二，大语言模型算法本身，如 ChatGPT 模型、GPT-4 模型和文心一言模型；第三，大语言模型产品，指在大语言模型基础上形成的具体产品，如 ChatGPT 聊天机器人、接入 GPT-4 的必应搜索引擎。

---

①百度创始人李彦宏预测称，十年以后全世界或有 50％的工作会是提示词工程（Prompt Engineering）。参见：《李彦宏独家回应 36 氪：如何看待 AI 代替人类工作》，36 氪，2023 年 3 月 22 日发布，2023 年 5 月 23 日访问，https://36kr.com/newsflashes/2182652773859072。

②马国洋《论刑事诉讼中人工智能证据的审查》，《中国刑事法杂志》2021 年第 5 期，第 158 页。

大语言模型生成的具体内容由训练数据和深度学习算法决定:训练数据是指用于大语言模型训练的海量数据,深度学习算法可以理解为大语言模型学习的方法。训练数据和深度学习算法涉及到大语言模型在技术公正和算法黑箱方面的核心问题,具有相当的证据意义。但它们的可解释性非常弱,难以被人类的自身能力感知,因此在证据运用上存在困难。其一,训练数据对模型可靠性有重要影响,如样本数据少容易"过拟合"①,且"在不可信数据上训练的模型的性能将会大幅下降,甚至在模型中留有严重后门"②。评价训练数据的因素包括数据质量、规模、多样性以及是否经过预处理等。《生成式人工智能服务管理办法(征求意见稿)》提出,训练数据应符合法律法规要求,不得侵犯知识产权、个人信息权,应当保证真实性、准确性、客观性和多样性。从证据角度看,训练数据规模巨大、内容复杂,人类无法通过自身感知能力对其进行有效审查,这与大数据证据有一定相似性。本文认为,可以参照刘品新对大数据证据的观点,让训练数据通过司法鉴定,作为鉴定意见进入证明活动③。其二,深度学习算法是一类超多层神经网络学习算法,其复杂性是大模型强黑箱效应的重要原因④。从证据角度看,深度学习算法的内在逻辑难以为一般人理解,比如 ChatGPT 的深度学习算法结合了 Transformer 架构、多头注意力机制、自监督学习和语言模型预训练等技术。鉴于深度学习算法对证据运用的要求超出一般人能力水平,本文认为其也较适合作为鉴定意见进入证明活动。

大语言模型是在海量数据上训练得到的参数规模巨大的深度学习模型,其参数量一般在百亿级以上,代表模型有 Open AI 的 GPT-4、百度的文心一言和阿里的通义千问等。从证据角度看,大语言模型至少有以下值得关注的特点:其一,大语言模型是生成新内容而非简单检索,属于深度合成技术⑤;其二,程序员编写的代码是大语言模型的骨架,但决定大语言模型预测结果的还是机器学习算法学习到的参数,换言之,直接决定黑箱输出结果的大模型的核心是参数而非代码,因此,即使程序员也未必能理解大语言模型的"黑箱";其三,大语言模型的黑箱效应极为显著,但蕴含着大量人为因素,比如训练数据的选取和深度学习算法的设计。因此,大语言模型由大量代码构成且可解释性差,与训练数据和深度学习算法存在一定相似性,也较适合以鉴定意见的形式进入证明活动。

大语言模型是一种基础性工具,软件开发者可以将其集成到自己的应用中,形成功能丰富的产品⑥。目前大语言模型产品主要包括 ChatGPT、文心一言等聊天机器人,必应等搜索引擎,以及 Copilot 等办公应用。大语言模型产品直接影响大语言模型材料的呈现,比如聊天机器人形成的大语言模型材料常以对话文本形式呈现,联网大语言模型形成的材料则包含较丰富的网络链接。现阶段,不同种类大语言模型产品形成的材料还没有脱离"一问一答"的基础文本形态,但未来大语言模型材料可能形态多样甚至难以辨认。相应地,大语言模型材料的提取收集、固定保全、审查判断都会面临新的挑战。

3.大语言模型运行环境信息

大语言模型需要强大算力的支持,比如微软 Azure 云计算平台是 OpenAI 运行和管理 ChatGPT 的重要基础。但是大语言模型生成的内容与所部署的云计算平台无关,云计算平台主要影响到大语言模型产品运行的稳定性,比如云计算平台的状况和故障可能导致大语言模型的响应时间较长,或者无法正常运行。本文将云计算环境信息纳入大语言模型材料是出于完整性考虑,但云计算环境对证据法视角下的大语言模型材料影响极为有限,因此云计算环境信息的证据价值不高。

使用者本地环境信息是指反映使用者操作大语言模型产品时的计算机或移动设备环境的信息,主要包括使用的日期、时间和地区,所用大语言模型的产品版本,计算机或移动设备的型号、操作系统和浏览器,互联网协议地址(IP 地址)等。使用者本地环境信息在证明活动中的作用主要有二:一是保障大语言模型材料

①周志华《机器学习》,第 13 页。

②何灿《机器学习模型训练数据的安全性研究》,南京航空航天大学 2021 年硕士学位论文,第 1 页。

③刘品新《论大数据证据》,《环球法律评论》2019 年第 1 期,第 28 页。

④张博伦《超越算法的黑箱想象》,《清华社会学评论》第 18 辑,社会科学文献出版社 2022 年版,第 152－153 页。

⑤《互联网信息服务深度合成管理规定》,国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部令第 12 号,2022 年 11 月 25 日公布,中国网信网,2022 年 12 月 11 日发布,2023 年 5 月 23 日访问,http://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm。

⑥於兴中、郑戈、丁晓东《生成式人工智能与法律的六大议题:以 ChatGPT 为例》,《中国法律评论》2023 年第 2 期,第 2 页。

的真实性，尽量避免人机对话材料被伪造或篡改；二是确定使用者的身份，大语言模型使用者和账号所有者未必是同一人，因此需要结合使用者本地环境信息进行身份同一性判断。可见，使用者本地环境信息具有一定证据价值，且可以参照电子数据环境信息的相关程序规范进行收集和举示，证据运用成本不高。

（四）大语言模型材料的类型化及其证据属性

1.以反映“人类－机器意见”的程度为标准判断其证据属性

（1）“机器意见型”大语言模型材料

大语言模型可以对已经学习到的事物、事件，或使用者输入的复杂内容进行分析，提出生成性的观点和判断。“机器意见型”大语言模型材料中体现了较多大语言模型的机器判断，而使用者的人类意见较少得到体现，至少包括：其一，大语言模型对客观事件、事物作判断形成的材料，如使用者要求大语言模型对某家公司、某所高校、某项产品、某个历史事件、社会事件等进行的判断；其二，对使用者输入的弱主观性内容分析形成的、主要体现大语言模型意见的材料，如使用者要求大语言模型对其输入学术文章所作的分析评价；其三，对输入的复杂内容进行鉴定形成的材料，如使用者要求大语言模型审查书证可靠性形成的分析意见；其四，对输入的复杂内容进行推理形成的材料，如案件侦办人员输入已经较确定的案件情况，大语言模型据此作出的案情推理。那么，“机器意见型”大语言模型材料能否成为证据呢？比如某公司在广告中宣称其产品全国知名，依据是大语言模型在对话中肯定该产品全国知名，那么相关大语言模型材料能否成为支撑其广告行为合法性的依据？大语言模型基于海量训练数据和深度学习算法产生，其“机器意见”有一定客观性，与待证事实之间存在关联性，因此“机器意见型”大语言模型材料可以在合法前提下作为证据使用。但是，大语言模型本身存在“幻觉”现象，而且使用者可以通过提示对大语言模型输出的内容进行误导。比如使用者可以先告诉大语言模型该产品全国知名，再进行提问，就能得到想要的回答。因此本文认为，“机器意见”型大语言模型材料可以成为证据，但对其客观性和关联性的审查需要格外谨慎。对“机器意见型”大语言模型材料的审查尤其要注意人机对话的上下文，排除使用者提示对机器意见的诱导。

（2）“人类－机器意见平衡型”大语言模型材料

此类材料指人类意见和机器意见对大语言模型生成材料发挥作用较为平衡的类型。在目前使用场景下，“人类－机器意见平衡型”大语言模型材料至少包括以下情形：其一，人机合作创造性工作形成的材料，以法律职业为例，GPT-4 能够通过美国模拟律师考试，并且分数位于应试者前 10％左右①，文心一言的法律能力有较大进步空间②，两者都无法完全取代律师在处理复杂案情和证据、调查取证等方面的作用，需要通过人机深度合作形成可用的工作成果；其二，对使用者输入的强主观性内容进行分析形成的材料，如案件侦办人员将数份言词证据输入大语言模型材料，要求其梳理前后是否有矛盾之处，对涉及人员言论的可信度进行评估；其三，经使用者重要“提示”（Prompt）形成的材料，比如使用者在对话上文给出“某公司为知名企业、有良好商誉”的信息，能够在大语言模型记忆范围内影响其回答；其四，在强人机交互环境中形成的材料，大语言模型已经被计划用于智能客服领域，由其形成的客户服务记录有较强的人机意见交换性，一般属于“人类－机器意见平衡型”大语言模型材料。“人类－机器意见平衡型”大语言模型材料可以通过反映机器意见和人类意见发挥证据作用，这里的“平衡”不要求人机意见占比持平，而是一种基于人机交互复杂性的折中描述。以案件侦办人员通过大语言模型分析言词证据为例：言词证据形成的过程和侦办人员选取言词证据的过程都含有较强主观因素，体现的是人类意见；大语言模型分析言词证据得出结论，体现的是机器意见。因此，审查“人类－机器意见平衡型”大语言模型材料时需要对机器意见和人类意见进行一定区分，根据具体案件需要排除人类意见或机器意见的干扰，抑或对两种意见分别进行审查判断。

---

①“GPT-4 is OpenAI's most advanced system, producing safer and more useful responses,” OpenAI, updated March 15, 2023, accessed May 22, 2023, https://openai.com/product/gpt-4.

②根据笔者在 2023 年 3 月 18 日的测试，文心一言可以定位到《中华人民共和国民法典》部分具体条款，但会编造法条的条数和内容；它对《中华人民共和国刑法》了解得非常笼统，只到“章”；它对商法、公司法的掌握也比较笼统。但在适当和充分的提示下，文心一言对《中华人民共和国土地管理法》修改的情况作出了较为完善的描述和评价，甚至能对其中的土地征收程序修改情况进行描述和评价。总体而言，文心一言在法条检索和法律咨询上的表现不尽如人意，但充分恰当的提示可以提高它的回答质量。

(3)“人类意见型”大语言模型材料

一些大语言模型生成材料几乎完全是对人类意见的反映,至少包括以下情形:其一,基于使用者提供的内容经简单加工形成的材料,如不含艺术性的语言翻译、文字语法校对、文章润色等;其二,完全按照使用者要求生成的、基本不含机器意见的材料,如“网络水军”评论文本。“人类意见型”大语言模型材料在一定情况下可以成为证据,比如使用者通过大语言模型翻译违法文章用于不当宣传,通过大语言模型大量生成垃圾信息用于“网络水军”活动等。在这种情况下,大语言模型材料通过反映人类意见来证明案件事实,可以用来证明使用者行为的主观方面。

2.以证明活动中的作用为标准判断其证据属性

(1)用于证明案件事实的大语言模型材料

“案件事实”是证据定义、证明对象等问题的核心概念之一,既往研究对“案件事实”的理解存在争议,本文支持“案件事实就是实体法事实”的观点,所称“案件事实”即指对解决案件实体问题具有法律意义的事实①。大语言模型材料可以在多种情况下对案件事实起证明作用:在著作权案件中,大语言模型材料可以证明通过大语言模型进行的改写、抄袭等事实;在商业案件中,它可以证明大语言模型服务提供者在大语言模型对话中违规植入广告、进行不良诱导等不当商业行为;在刑事案件中,它可以证明向大语言模型咨询犯罪法律问题的犯罪嫌疑人主观心态。在这些情况下,大语言模型材料能够证明案件事实,具有作为证据的不可替代性,可以被称为“大语言模型证据”。

(2)用于证据审查的大语言模型材料

大语言模型可以用来审查已经收集到的证据,尤其是书证、言词证据等以文本内容发挥证明作用的证据。比如办案人员可以将大量言词证据输入大语言模型,要求大语言模型梳理陈述中的前后矛盾。正如本文对“人工智能证据审查方法”和“人工智能证据”的区分,本文认为这种发挥证据审查作用的大语言模型材料也不宜称为“大语言模型证据”(详后)。对证明对象范围最广的理解是,证明对象包括实体法事实、程序法事实和证据事实②:实体法事实指对解决案件实体问题具有法律意义的事实③;程序法事实指引起诉讼法律关系发生、变更和消灭的事实,包括诉讼行为和诉讼事件④;证据事实指证据提供的内容⑤。20 世纪 90 年代中期之后,我国诉讼法通说观点基本认同证明对象范围包括实体法事实和程序法事实,并大多否定诉讼证明对象中包括证据事实⑥。证据事实不属于证明对象范围的通说印证了“审查证据的方法不是证据”的观点。本文认为,可以将通过大语言模型审查证据的方式称作“大语言模型证据审查方法”,与“大语言模型证据”相区分。

(3)用于辅助案件调查的大语言模型材料

除了证明案件事实和证据事实,大语言模型材料还可以用于辅助案件调查。比如在刑事案件侦办中,办案人员可以将案件背景和收集到的证据情况输入大语言模型,要求其推理、还原出可能的案件情况,甚至尝试推理具备作案嫌疑的人。辅助案件侦查形成的大语言模型材料无法证明案件事实,只是拓宽使用者认识案件的思路,因此不属于证据。

## 三 大语言模型材料的证据资格和运用

(一)大语言模型材料和证据资格

大语言模型材料将深度广泛地进入证明活动,那它能否成为证据?诉讼法学研究对证据的定义存在分

①陈光中、周国钧《论刑事诉讼中的证明对象》,《中国政法大学学报》1983 年第 3 期,第 58 页。

②也有观点主张不采用传统证明对象范围理论(区分实体法事实、程序法事实和证据事实),认为证明对象的范围是诉辩双方的诉讼主张。该理论与本文讨论内容有一定距离,因此未作展开。参见:鲁杰、曹福来《论证明对象的范围是诉辩双方的诉讼主张》,《政治与法律》2009 年第 1 期,第 128-132 页。

③陈光中、周国钧《论刑事诉讼中的证明对象》,《中国政法大学学报》1983 年第 3 期,第 58 页。

④卞建林编《证据法学》,中国政法大学出版社 2000 年版,第 279 页。

⑤陈光中、周国钧《论刑事诉讼中的证明对象》,《中国政法大学学报》1983 年第 3 期,第 62 页。

⑥闵春雷、刘铭《证明对象研究走向评析》,《吉林大学社会科学学报》2009 年第 2 期,第 48 页。

歧，其中对我国立法影响较大的观点主要有三种，分别是“事实说”、“根据说”和“材料说”①。事实说认为证据是“证明案件真实情况的一切事实”，曾在研究中占主导地位②，我国 1979 年《刑事诉讼法》采纳这一观点③。根据说认为“证据是查明和确定案件真实情况的根据”，代表学者有陈一云、龙宗智、何家弘、刘品新等④，《最高人民法院关于贯彻执行〈民事诉讼法（试行）〉若干问题的意见》采纳这一观点⑤。材料说认为证据是“可以用于证明案件事实的材料”⑥，以 2012 年《刑事诉讼法》修改的采纳为标志，材料说成为我国证据定义的主流观点⑦。本文从我国现行立法出发，认为证据是可用于证明案件事实的材料。大语言模型材料证明案件事实的情形至少包括：证明通过大语言模型产品进行的不当行为（如不当广告行为、“网络水军”行为），证明咨询犯罪问题的犯罪嫌疑人的主观心态，等等。因此，大语言模型材料可以成为证据。

大语言模型证据能否在证明活动中被采纳？这是证据资格的问题，研究中常见的“证据能力”⑧、“证人能力”、“证据的采纳标准”等描述的都是证据资格⑨。大陆法系常采用证据资格（Competency of Evidence）、证据能力概念，英美法系中则表述为证据的可采性（Admissibility of Evidence）⑩。证据资格的内容在不同证明活动中、面对不同的证据形式时有所不同，基本内容包括客观性、关联性和合法性。首先，客观性是指证据应当具有客观存在性，包括证据在内容上是对客观事物的反映，在形式上是一种客观存在，能够被人通过某种方式感知⑪。在内容上，大语言模型材料能够反映以使用者行为为代表的多种客观事物；在形式上，大语言模型材料中的人机交流材料能被人直观感知，本体材料和使用环境信息也能通过鉴定和技术公司公开为人感知。其次，关联性是指证据必须与待证事实存在联系。大语言模型材料深入社会生活，能够在民事、刑事、行政等多种场景下与案件事实相联系，因此具备关联性。最后，合法性是指证据的调查主体、形式、收集程序或提取方法应符合法律规定。证据是否需要具有合法性在研究中有较大争议，何家弘认为该争议的存在是由于证据概念与证据资格发生混淆：合法性是证据资格的考量因素，经非法主体、形式、程序得到的材料依然可以是证据，只是不一定能在证明活动中被采纳⑫。本文认同这一观点，尽管大语言模型材料的取证主体、证据形式和取证程序尚无法律依据，但这并不影响其成为证据，只影响其在证明活动中能否被采纳。

综上，大语言模型材料可以成为证据出现在证明活动中。由于大语言模型证据在一些情况下能够证明案件事实，具有真实性，本文认为通过完善法律，它也具有证据资格。

（二）大语言模型证据和相关类型证据比较

1.大数据证据和大语言模型证据

大语言模型是人工智能领域中自然语言处理的前沿成果，与大数据技术密切相关。人工智能的发展基于大量数据，而大数据技术的分布式存储和分布式计算为人工智能提供了强大的存储和计算能力⑬。大数

①何家弘、刘品新《证据法学》，法律出版社 2022 年版，第 118－120 页。

②何家弘、刘品新《证据法学》，第 118－119 页。

③《中华人民共和国刑事诉讼法》（1979 年）第三十一条：“证明案件真实情况的一切事实，都是证据。”《中华人民共和国行政诉讼法》（1989 年）和《中华人民共和国民事诉讼法》（1991 年）未对“证据”作定义；《行政诉讼法》（1989 年）第三十一条对证据种类进行列举，规定“以上证据经法庭审查属实，才能作为定案的根据”；《民事诉讼法》（1991 年）第六十三条对证据种类进行列举，规定“以上证据必须查证属实，才能作为认定事实的根据”。

④陈一云、王新清、严端编《证据学》，中国人民大学出版社 2013 年版，第 3 页；龙宗智《诉讼证据论》，法律出版社 2021 年版，第 8 页；何家弘、刘品新《证据法学》，第 119、121 页。

⑤《最高人民法院关于贯彻执行〈民事诉讼法（试行）〉若干问题的意见》（已废止），[1984]法办字第 112 号。其中第四节“证据问题”规定：“证据是查明和确定案件真实情况的根据。”

⑥龙宗智的观点部分体现了证据的材料说：“具体的证据，是指承载证据信息（事实与意见），而以特定形式表现出来的证明材料。”参见：龙宗智《诉讼证据论》，第 8 页。

⑦《中华人民共和国刑事诉讼法》（2012 年修正）第四十八条：“可以用于证明案件事实的材料，都是证据。”

⑧“证据能力，是指能够成为证据的资格。”参见：田口守一《刑事诉讼法》，张凌、于秀峰译，法律出版社 2019 年版，第 437 页。

⑨林志毅《论刑事证据资格之多重性》，《中国法学》2022 年第 1 期，第 263 页。

⑩参见：田口守一《刑事诉讼法》，第 438 页；何家弘、刘品新《证据法学》，第 124－125 页。

⑪何家弘、刘品新《证据法学》，第 124－128 页。

⑫何家弘、刘品新《证据法学》，第 128－132 页。

⑬林子雨编著《大数据导论——数据思维、数据能力和数据伦理》，高等教育出版社 2020 年版，第 54－55 页。

据证据和大语言模型证据都随前沿技术发展产生,面临着相似的黑箱质疑、证据资格问题和证据种类问题,两者也存在区别。第一,在技术基础上,大数据技术的重心在于对海量数据的处理和对相关性关系的发掘,是一种"寻找结果"的传统计算;而大语言模型属于人工智能技术,是一种"允许机器执行认知功能"的计算方法,目的在于辅助或者替代人类完成某些任务,进行某些决定①。第二,在具体内容上,大数据证据由海量基础数据、大数据分析技术和大数据分析结果组成②;最完整的大语言模型证据由人机交流材料、大语言模型本体材料和运行环境信息组成。第三,在运用难度上,大数据证据在证明活动中一般以大数据分析报告、说明报告或鉴定意见的形式呈现,有一定专业门槛;大语言模型证据或多以人机交流材料(如对话文本)形式呈现,运用难度相对较小。

大数据证据已经在司法裁判中得到运用,法律实务和学术研究均认可其证据资格,但在证据种类问题上存在分歧。在司法实践中有将大数据证据归为鉴定意见、电子数据、书证、证人证言,甚至是将其作为"侦破经过"或"情况说明"的做法③;学术上对大数据证据的种类有纳入鉴定意见④、独立类型⑤等不同看法。本文认为大数据证据有别于传统证据种类,但是不宜作为"大数据证据"进入立法。证据分类应当实现识别、适用和交往的基本功能⑥,而"大数据证据"这一分类未必具有交往性(即普遍性)。随着技术发展不能被归入法定证据种类的新技术证据只会越来越多,比如我们正在讨论的大语言模型证据⑦。有学者认为,鉴于法定证据种类在面对新技术证据时存在的困难,应该放弃将证据种类作为证据门槛的做法⑧。本文支持这一观点,证据资格才是"证据门槛",证据种类是我们认识证据的工具。可以看到,大数据证据在证据种类上的混乱并未影响它在司法实践中被广泛运用,大语言模型证据或许也将走上类似的道路。

2.人工智能证据和大语言模型证据

大语言模型证据属于人工智能证据,但是人工智能技术有多种研究方向,比如知识图谱和大语言模型是两种完全不同的方案。因此,人工智能证据研究成果难以套用到大语言模型证据上,却可以成为理论来源和重要参考。人工智能证据研究在刑事诉讼领域和民事诉讼领域都已展开,但现有研究存在将"人工智能证据审查方法"和"人工智能证据"混用的情况,这与谢登科等指出的"电子数据区块链存证"与"区块链证据"混用的情况具有一定相似性⑨。有文章举例的"人工智能证据"是人脸识别系统分析结论,该分析结论在诉讼中被用来证明特定照片上的人是特定某人⑩。本文认为该例子不一定妥当,人脸识别系统分析结论在诉讼中起到的是补强书证(即本案中照片)的作用,是作为辅助证据(或称补助证据)用来证明证据事实的⑪。用人工智能方法对其他证据进行审查判断形成的材料或不宜称为"人工智能证据",可以将这种方法称为"人工智能证据审查方法"。

可能会有这样的反对意见:验证其他证据形成的人工智能材料也与案件事实相关,所以是"人工智能证据"。本文认为该观点有一定道理,而且符合司法实践和通常认识,但尚有可商榷之处。在区块链证据领域,"区块链证据"和"电子数据区块链存证"的混用已经较为普遍,有文章指出了既往研究中存在的混用情况及

①林子雨编著《大数据导论——数据思维、数据能力和数据伦理》,第55页。

②严若冰《以定义为中心的大数据证据独立种类研究》,《山东警察学院学报》2020年第5期,第87-89页。

③严若冰《以定义为中心的大数据证据独立种类研究》,《山东警察学院学报》2020年第5期,第80-91页。

④刘品新《论大数据证据》,《环球法律评论》2019年第1期,第28页。

⑤徐惠、李晓东《大数据证据之证据属性证成研究》,《中国人民公安大学学报(社会科学版)》2020年第1期,第47-57页。

⑥识别性指分类能将某类证据与其他证据进行有效区分,适用性是指证据分类有助于适用证据规则,交往性即普遍性,是指证据分类获得普遍认可,因此便利交流与交往。参见:龙宗智《诉讼证据论》,第44页。

⑦严若冰《以定义为中心的大数据证据独立种类研究》,《山东警察学院学报》2020年第5期,第80-91页。

⑧郑飞、马国洋《大数据证据适用的三重困境及出路》,《重庆大学学报(社会科学版)》2022年第3期,第207-218页。

⑨谢登科、张赫《电子数据区块链存证的理论反思》,《重庆大学学报(社会科学版)》2022年12月20日网络首发,第1-14页,http://kns.cnki.net/kcms/detail/50.1023.c.20221219.1201.001.html。

⑩马国洋《论刑事诉讼中人工智能证据的审查》,《中国刑事法杂志》2021年第5期,第158页。

⑪陈光中、周国钧《论刑事诉讼中的证明对象》,《中国政法大学学报》1983年第3期,第58-64页;田口守一《刑事诉讼法》,第438-439页。

其给研究带来的困难①。人工智能证据研究尚处初期，厘清概念有助于今后研究的顺利开展，因此本文更倾向于区分“人工智能证据”和“人工智能证据审查方法”。比较符合这一“人工智能证据”定义的有金融领域的智能投顾材料②，由 AI 绘画工具生成的 AI 绘画作品，由 AI 语音工具生成的 AI 翻唱作品，以及大语言模型证据等。

（三）大语言模型材料的运用场景

1.民事法律证明场景中的大语言模型材料

民事领域或将是各法律部门中最早出现大语言模型材料的。大语言模型通过广告营利的商业模式几乎是板上钉钉，其中蕴含着法律风险。大语言模型以一问一答的形式向使用者提供意见，使用者省去了在搜索引擎中筛选信息的过程，但这种“不必选择”也意味着“难以选择”和“易被误导”。如果大语言模型在对话过程中推荐商业广告，用户甚至可能意识不到广告存在，这种广告在涉及医疗、法律服务等敏感行业时会更具危险性③。

据路透社报道，微软已经在尝试向搭载 GPT-4 的必应搜索引擎中加入广告，比如在机器回复中提供付费链接④。又以文心一言为例，大模型的实现和维持依赖强算法和大算力，这意味着文心一言在开发阶段就消耗了巨量资源，且后续业务开展需要以大量资金投入为保障。广告业务是百度公司的重要收入来源，百度 2022 年第四季度的在线营销收入（Online Marketing Revenue）为人民币 181 亿元，占该季度营收（331 亿元）的54.68％⑤。文心一言作为国内推出的第一款大语言模型炙手可热，承接广告业务的经济效益相当可观。大语言模型的技术复杂性使其较难受到外界有效监督，且法律本身存在滞后性，但法律人对大语言模型的民商事合规风险应有一定预见和警惕。

2.刑事法律证明场景中的大语言模型材料

一项新技术出现后，社会群体内接受新技术的速度和能力不同，由此带来的信息差将让犯罪分子有机可乘。大语言模型以假乱真的对话能力可能被用于违法犯罪活动中，比如“网络水军”活动和电信诈骗犯罪。以“网络水军”为例，目前“水军”在互联网上的发言较为生硬，辨识难度不高。但大语言模型可以高效编写大量自然流畅的虚假文案，提高“水军”活动的效率，增强了违法犯罪的隐蔽性和危害性。又以诈骗案件为例，在以婚恋为诱饵的“杀猪盘”骗局中，犯罪嫌疑人或可用大语言模型聊天机器人和受害者进行对话“培养感情”，降低犯罪成本。在这些情况下，使用者与大语言模型交流形成的材料将成为证明案件事实的证据。

除了直接证明案件事实，大语言模型材料在刑事活动中还可以作为破案线索，或是审查其他证据的辅助证据。比如在案件侦破阶段，警方可以将已经搜集到的案件信息和经过确认的部分证据输入大语言模型，要求其推理出案件最有可能的几种情况，以此拓宽办案思路。对于待初步审查的书证、言词证据，警方可以将证据文本内容和搜集该证据的相关情况输入大语言模型，要求其梳理案件中的人物关系和主要情节，进行内容、程序上的审查。作为破案线索和辅助证据的大语言模型材料虽然可靠性不一定高，但是在保证算法公正的前提下具有较高公正性，有助于提高办案效率。

3.行政法律证明场景中的大语言模型材料

根据数字政府建设和 2023 年国务院机构改革体现的发展方向，我国在政府领域引入大语言模型或许只是时间问题。一方面，数字化智能化是我国政府发展的重要方向，“十四五”规划中明确要求“全面推进政府

①谢登科、张赫《电子数据区块链存证的理论反思》，《重庆大学学报（社会科学版）》2022 年 12 月 20 日网络首发，第 1－14 页，http://kns.cnki.net/kcms/detail/50.1023.c.20221219.1201.001.html。

②徐凤《人工智能算法黑箱的法律规制——以智能投顾为例展开》，《东方法学》2019 年第 6 期，第 83－86 页。

③如曾经发生过莆田系医院通过商业竞价在百度搜索结果中投放广告，患者通过广告被引导到莆田系医院就医导致治疗延误的事件。参见：张燕《揭“莆田系”医院盈利秘密》，《中国经济周刊》2016 年第 19 期，第 24－26 页。

④Sheila Dang, “Exclusive: Microsoft’s Bing plans AI ads in early pitch to advertisers,” Reuters News, updated February 18, 2023, accessed May 22, 2023, https://www.reuters.com/technology/microsofts-bing-plans-ai-ads-early-pitch-advertisers-2023-02-17/.

⑤“Baidu Announces Fourth Quarter and Fiscal Year 2022 Results,” Baidu IR, updated February 22, 2023, accessed May 22, 2023, https://ir.baidu.com/investor-overview/.

运行方式、业务流程和服务模式数字化智能化”①,2022年,《国务院关于加强数字政府建设的指导意见》提出“构建数字化、智能化的政府运行新形态”。另一方面,2023年国务院机构改革方案要求“中央国家机关各部门人员编制将统一按照5%的比例进行精减”②,在安全可靠的前提下将大语言模型引入政府工作将是精简编制、集中编制资源攻克重点问题的合理方案。极为强调安全性的国内银行业已经开始“拥抱”大语言模型,文心一言将在银行的客服、风控、投研、营销等领域开展应用③,如果文心一言能够实现令人较为满意的对话和文本生成能力,这些银行的职位需求将相应减少。

文心一言已经与一些政府部门、国有企业和事业单位达成合作,如工信部新闻宣传中心④、邮储银行⑤,大语言模型进入行政领域或不遥远。从大语言模型目前的应用来看,它对外可以受理业务投诉、为群众提供咨询服务、参与网络行政执法,甚至进行自动化的行政许可形式审批⑥;对内可以成为每一位行政机关工作人员的“私人助手”,处理重复性和日常性较高、非核心机要的文书工作,提供政策和决定咨询。我国基层公务员的工作负担中有相当一部分是重复繁琐的文书工作,如果能在保证意思准确、不影响工作质量的前提下引入大语言模型,将有助于解放基层劳动力。当政务活动中开始应用大语言模型技术,大语言模型材料也将迈入行政程序活动、行政复议和行政诉讼领域。

(四)司法实践中大语言模型证据的运用

证据在实践中的表现形式可能与法律规范要求的并不相同,非常典型的例子是民事诉讼中的电子证据,尤其是在网络交易型证明活动中。有学者指出,网络交易型诉讼的证明活动高度依赖电子证据,但在实践中原告举示的电子证据常常以截图、打印稿的形式呈现⑦。这在一定程度上是因为此类案件中的电子证据偏在于互联网平台,但足以反映出证据实践表现形式与法律规定之间的落差。结合电子数据和大数据证据在实践中的举证状况,本文对大语言模型证据在诉讼证明活动的运用进行如下猜测:一方面,大语言模型证据举示方出于成本效率的考量,或将以截图、打印稿的形式对人机交流材料进行举证;另一方面,质证方将从大语言模型的算法公正性(黑箱效应),大语言模型材料的完整性,账号所有者与人机交流者身份的同一性等角度质疑人机交流材料;同时,被质证的一方可以通过大语言模型黑箱属性的固有性和极高昂成本对抗黑箱质疑,通过充分举示人机交流材料、大语言模型本体材料和大语言模型运行环境信息对抗完整性质疑,通过举示使用者运行环境信息对抗身份同一性质疑。

从证明活动效率考虑,本文认为,一般案件可以仅举示人机交流材料和使用者本地环境信息,重大案件才需要对人机交流材料、大语言模型本体材料和大语言模型运行环境信息作完整举示。一方面,从成本上看,大语言模型黑箱效应突出,本体材料和云计算环境信息的提取和审查有较高门槛,对相关人员专业能力要求极高;另一方面,从与待证事实的关联性来看,大语言模型训练和运行的成本极高,为实施普通违法犯罪行为故意调整大语言模型的可能性较小,云计算环境一般不影响大语言模型生成的内容,因此大语言模型本体和云计算环境与一般案件事实的关联性不强。而人机交流材料和使用者本地环境信息与案件事实的联系紧密,且运用难度较小,因此本文支持在一般案件中将人机交流材料和使用者本地环境信息认定为完整的大语言模型证据。

---

①《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》,中国政府网,2021年3月13日发布,2023年5月29日访问,https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm。

②肖捷《关于国务院机构改革方案的说明——2023年3月7日在第十四届全国人民代表大会第一次会议上》,中国政府网,2023年3月8日发布,2023年5月23日访问,http://www.gov.cn/guowuyuan/2023-03/08/content_5745356.htm。

③李海颜《牵手百度 多家银行寻求中国版ChatGPT新应用》,《北京商报》2023年2月28日,第7版。

④赵乐瑄《工信部新闻宣传中心(人民邮电报社)宣布接入百度文心一言 树立行业媒体智能化新标杆》,中国工信产业网,2023年2月14日发布,2023年5月23日访问,https://www.cnii.com.cn/rmydb/202302/t20230214_446697.html。

⑤《邮储银行宣布接入百度“文心一言”提供更智能更有温度的金融服务》,中国邮政集团有限公司网站,2023年2月18日发布,2023年5月23日访问,http://www.cptu.org.cn/xhtml1/report/23021/7167-1.htm。

⑥GPT-4已经能够对输入的文本和图像进行分析,如果未来该类技术更加成熟且能保证安全性,或可用于自动化行政许可形式审查,甚至能实现一定程度的实质审查。

⑦比如在一起食品网络交易诉讼中,原告提供了4项电子证据,分别是网页商品快照打印件、网页订单详情截图打印件、快递单打印件、电子支付账单详情截图打印件。参见:周翔《论电子证据的偏在性及其克服》,《大连理工大学学报(社会科学版)》2020年第1期,第92页。

## 四　大语言模型材料作为证据的特点

### （一）直观性强：人机交流材料能够被人类直观感知

人机交流材料呈现为一问一答的连续人机对话，这种直观性使大语言模型证据与需要鉴定的科学证据形成区别。人机交流材料的形成过程处于极强黑箱效应中，但我们并非一定要通过司法鉴定打开这个黑箱。一方面，对人机交流材料形成过程的探究需要对大语言模型本身进行分析，鉴定成本较难得到有效控制；另一方面，大语言模型的形成和运行需要巨额资金投入，黑箱内部与普通个案的关联性不强。因此，本文认为，强行要求对人机交流材料进行鉴定将极大提高诉讼成本，缺少必要性。人机交流材料可以凭借其直观易理解的优势，参照互联网聊天记录、网页信息进行举证质证。从成本效益上看，人机交流材料最有可能成为未来诉讼活动中大语言模型证据的表现形式。

在收集提取人机交流材料时，当事人应尽可能保证材料的可链接性和真实性。人机交流材料目前多以浏览器网页为载体，且缺少规范便捷的证据提取收集方法，存在伪造、篡改的可能性。比如在 Chrome 浏览器中打开网页、按 F12 进入开发者工具，可以通过修改网页元素（Elements）来改变网页呈现内容。在司法实践中运用大语言模型材料可注意以下两点：第一，保留原始材料，使审判人员和其他诉讼参与人能够通过网络链接查阅到人机交流材料原件；第二，在提取过程中进行屏幕录像，并对提取到的网页文件计算哈希值[①]，或者通过司法区块链工具将相关网页信息以链上数据的形式进行保全。

### （二）可解释性弱：不等于大语言模型证据可靠性弱

可解释性的定义存在争议，涉及到认知科学、计算机科学、心理学和哲学等领域[②]。本文中人工智能的可解释性指人类对人工智能自动决策的理解，包括人工智能自动决策的原因、方法和内容等。大语言模型属于深度学习模型，可解释性弱是其目前最为人诟病的特征之一。深度学习模型的黑箱程度高于社会主流观点对算法黑箱的认识。有社会学者指出，算法黑箱问题常被归结到“专业知识”和“透明度”上，许多观点认为只要人类具有关于算法的专业知识而且能够接触到相应代码，就能够“探查到算法的社会影响，消除存在于其中的可能的偏见”。但深度学习模型的黑箱与传统“算法想象”对黑箱的理解不同，黑箱是深度学习模型的固有特征，它不仅对用户来说是一个黑箱，甚至对开发它的程序员和公司来说也是如此[③]。

大语言模型的可解释性弱不仅表现在与主流黑箱认识的对比上，还表现在与知识图谱和小模型等其他人工智能技术的对比上。知识图谱由一系列包含实体和关系的事实组成，直观可见，因此具有强可解释性。与小模型相比，大模型不仅有更庞大的参数量，还具有涌现能力（Emergent Abilities）。涌现性（Emergence）是指系统中数量性的变化引起了行为上性质的变化，可以理解为量变引起质变。大语言模型的涌现能力则是指当模型的训练量到达一定程度，就会有新的推理结构在神经网络中自发涌现，使其精准度得到大幅提升。这种涌现能力基于大量数据和强大计算能力，是较小模型所不具备的[④]。涌现能力意味着，大语言模型在程序员设计框架之外拥有非人为设计的能力，机器决策不能被完全预测，故可解释性难以得到保障。

大语言模型的可解释性弱决定了大语言模型材料的可解释性弱。尽管人机交流材料的内容直观可见，但我们难以认识大语言模型理解用户输入材料、生成输出材料的过程。此外，部分大语言模型材料由于客观原因并不在我国境内存储，我国在大语言模型技术方面与国际最先进水平仍存在客观差距。一些国内用户使用 ChatGPT 等国外大语言模型产品辅助工作，提高效率。对于这部分在国内使用、但由国外大语言模型生成且存储在国外的材料，如何进行收集、保存和审查，或将成为我们在技术和国际关系上需要面临的挑战。

大语言模型材料的可解释性弱，并不意味着大语言模型证据的可靠性弱。一方面，人机交流材料与电子

---

①孙百昌《网页取证 网页电子数据证据获取固定步骤与方法（2022）》，中国工商出版社 2022 年版，第 156－175 页。

②Roberto Confalonieri, Ludovik Coba et al., “A Historical Perspective of Explainable Artificial Intelligence,” *Wires Data Mining and Knowledge Discovery* 11, no.1 (January/February 2021) : 2-4.

③“算法想象”是张博伦提出的概念，指社会对算法的一般认识。参见：张博伦《超越算法的黑箱想象》，《清华社会学评论》第 18 辑，第 152－153 页。

④Roberto Confalonieri, Ludovik Coba et al., “A Historical Perspective of Explainable Artificial Intelligence,” *Wires Data Mining and Knowledge Discovery* 11, no.1 (January/February 2021) : 2-4.

数据相似,它可以通过可链接性来保障真实性,并且适宜通过司法区块链和公证的方法进行存证,较适应现行电子数据保全框架。另一方面,大语言模型本体材料和大语言模型云计算环境信息被"封装"在黑箱中,一般不影响大语言模型材料对具体案件事实的反映。对于确有必要进行举示的大语言模型本体材料和大语言模型云计算环境信息,也可以通过司法鉴定、以鉴定意见的形式进行举示。因此,大语言模型证据有能力反映一定案件事实,在诉讼证明活动中具有可靠性。

(三)偏在性:部分材料仅由少数技术公司掌握

证据偏在现象是指负有证明责任的一方无法掌握相应证据,因而难以履行证明责任,面临败诉风险。证据偏在问题产生于20世纪初的现代型诉讼,如医疗案件里医院和医生掌握患者病历。现代型诉讼的证据偏在问题未脱离诉讼双方,但随着互联网兴起和平台经济发展,电子证据常由互联网平台掌握,即电子证据常偏在于控辩双方之外的互联网平台①。大语言模型证据的偏在与互联网平台案件中电子证据的偏在有一定相似性,部分大语言模型证据仅由少数技术公司掌握。在大语言模型技术存在国家和地区间差距的情况下,这种证据偏在的状况还可能涉及到国际关系问题。OpenAI会收集用户使用ChatGPT服务时的各种信息,且OpenAI未在中国大陆正式开展服务,我国使用者作为海外用户被收集的各类信息均存储在美国②,这意味着我国司法机关获取ChatGPT生成材料的难度极大。

大语言模型证据偏在和互联网时代的电子证据偏在有一定相似性,因此也可以参考各国应对互联网时代电子证据偏在的方案。欧陆模式以证明责任减轻理论为核心,在法官主导证据调查的传统下展开诉讼证明活动;英美模式采取证据开示,坚持由当事人收集证据。我国立法与欧陆模式较为一致,但有观点指出这一方案正越发难以回应互联网平台垄断电子数据的问题,认为我国可以适当借鉴英美法系,适时提出网络平台的信息公开义务③。

本文更支持借鉴英美的证据开示模式,如果继续按照欧陆的证明责任减轻模式,我国法官将主导对大语言模型材料,尤其是本体材料和云计算服务信息的调查。但一般法官并不具备相应技术能力,加之法官群体工作量普遍较大,这样的制度设计难以发挥作用。而另一方面,大语言模型材料和相关专业知识均由技术公司掌握,根据百度公司的《文心一言(测试版)个人信息保护规则》和OpenAI公司的个人隐私政策,这些主要技术公司掌握着包括人机交流材料、大语言模型本体材料和大语言模型运行环境信息在内,所有可能被作为证据运用的大语言模型材料。因此本文认为,可以借鉴英美法系的电子数据证据开示制度,明确科技公司作为社会信息垄断者的证据开示义务。

(四)可识别性弱:大语言模型与深度合成治理

本文提出的大语言模型材料"可识别性"指人类能否识别一份材料是由大语言模型生成的,主要在于人机交流材料的可识别性。ChatGPT在对话时相当流畅自然,以至于能够让使用者感觉像与一名真正的人在对话。互联网上常有关于ChatGPT能否通过"图灵测试"的讨论④,虽然该问题尚无定论,但应该能够达成共识的是,当人类在不知情状态下与类ChatGPT水平的大语言模型对话,他有相当概率无法正确判断与其对话的是人类还是机器。也就是说,人类在缺少明确信息的情况下,未必能识别一份文本材料是否属于大语言模型的人机交流材料。该问题在刑事侦查阶段会影响案件调查的方向,影响案件性质和涉案主体的确定,在诉讼阶段也是庭审举证质证中难以回避的问题。它在民事领域也有一定影响,比如消费者要求与商家的真人客服进行沟通,能否有效判断对方提供的是大语言模型聊天机器人还是人类客服。

大语言模型属于深度合成技术,从理论上看,大语言模型材料的可识别性问题可以通过深度合成治理得到缓解。我国的深度合成治理立法走在世界前列,2023年1月开始实施的《互联网信息服务深度合成管理规定》要求深度合成服务提供者应当在"生成或者编辑的信息内容的合理位置、区域"进行显著的深度合成标

①周翔《论电子证据的偏在性及其克服》,《大连理工大学学报(社会科学版)》2020年第1期,第94-96页。

②"Privacy Policy," OpenAI, updated April 27, 2023, accessed May 23, 2023, https://openai.com/policies/privacy-policy.

③周翔《论电子证据的偏在性及其克服》,《大连理工大学学报(社会科学版)》2020年第1期,第91-102页。

④Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59, Issue 236 (October 1950): 433-460.

识，避免公众混淆或者误认[①]。目前 ChatGPT 和文心一言在对话中都会强调自己作为大语言模型的身份，这在某种程度上符合我国立法关于深度合成标识的要求。

但在实践层面，深度合成标识相关规定未得到充分落实：深度合成服务提供者未充分遵守立法关于深度合成标识的规定，且深度合成标识难以约束深度合成服务使用者的不当利用。一方面，从深度合成服务提供者的角度来看，笔者通过百度文心一格（AI 绘图工具）生成了四张图片，成品图片上并无人类可以感知的深度合成标识，这是当前深度合成服务的普遍状况。另一方面，从深度合成服务使用者的角度来看，运用和传播无深度合成标识的 AIGC 内容也相当普遍，且管理部门对此缺乏有效识别和规范手段，如目前互联网内容平台上充斥着由 AI 配音但未加标注的视频，以及由 AI 绘图生成的图片（甚至包括为数众多能够以假乱真的“虚拟人类”图片）。

国家网信办在 2023 年 4 月 11 日发布的《生成式人工智能服务管理办法（征求意见稿）》体现了国家对大语言模型运用的态度：生成式人工智能服务提供者应当指导用户合理利用相关服务，对利用过程中违反法律法规、商业道德或社会公德的用户暂停或终止服务。基于深度合成的立法现状和生成式人工智能的立法方向，本文认为可以考虑增设以下规定：第一，强调大语言模型的深度合成属性，使大语言模型运用与我国现行的深度合成治理规范相衔接，明确大语言模型和大语言模型产品适用于有关深度合成的法律法规；第二，要求应用大语言模型技术的产品至少在交互界面和生成文本中充分、明确地强调其大语言模型身份，并提醒用户合理合法地使用大语言模型产品；第三，使用大语言模型产品代替其进行对外交往的机构或个人应当表明其正在使用大语言模型产品，否则将承担不利法律后果。

**致谢**：本文在撰写过程中得到许多专业人士和同学的帮助，王钰薇女士对游戏行业 AIGC 使用情况给予了指导，李琳婕女士就“大语言模型材料的运用场景”部分与笔者进行讨论，柏林洪堡大学（Humboldt-Universitaet zu Berlin）苏泓宇同学对金融行业相关情况给予了指导，北京航空航天大学张雪峰同学，西湖大学高文炀同学，清华大学李思磐同学、邱浩先生、李岚皓先生在大语言模型、云计算、知识图谱等方面进行了技术指导。在此向他们表示衷心的感谢！

［责任编辑：苏雪梅］

①参见：《互联网信息服务深度合成管理规定》，国家互联网信息办公室、中华人民共和国工业和信息化部、中华人民共和国公安部令第 12 号，2022 年 11 月 25 日公布，中国网信网，2022 年 12 月 11 日发布，2023 年 5 月 23 日访问，http://www.cac.gov.cn/2022-12/11/c.1672221949354811.htm；张凌寒《深度合成治理的逻辑更新与体系迭代——ChatGPT 等生成型人工智能治理的中国路径》，《法律科学（西北政法大学学报）》2023 年第 3 期，第 39 页。