E-Commerce Letters 电子商务评论, 2025, 14(12), 6614-6620
Published Online December 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14124653

Hans汉斯

电商平台AI客服“幻觉”致损的综合治理研究

刘梦哲

上海理工大学出版学院，上海

收稿日期：2025年12月5日；录用日期：2025年12月24日；发布日期：2025年12月31日

摘　要

随着人工智能技术在电子商务领域的广泛应用，AI客服已成为电商平台和商家提升服务效率的重要工具。然而，AI客服的“幻觉”现象却可能给用户权益带来损害。本文立足于语言模型幻觉机理的分析，结合电商领域AI客服致损的现实案例，系统分析了不同情形下商家、平台及AI服务提供商之间的法律责任分配问题。研究发现，当前治理工作面临技术瓶颈、法律定性困境及举证困难等多重挑战。为此，本文从技术改进、法律规制与政策引导三个维度提出构建三位一体的综合治理体系，以期为电商平台AI客服的规范发展提供理论参考。

关键词

AI幻觉，法律责任，电商，综合治理

Comprehensive Governance Research on Losses Caused by AI Customer Service “Hallucination” on E-Commerce Platforms

Mengzhe Liu

College of Publishing, University of Shanghai for Science and Technology, Shanghai

Received: December 5, 2025; accepted: December 24, 2025; published: December 31, 2025

Abstract

With the widespread application of artificial intelligence technology in the field of e-commerce, AI customer service has become an important tool for e-commerce platforms and merchants to enhance service efficiency. However, the “hallucination” phenomenon of AI customer service may potentially harm user rights. Based on an analysis of the mechanisms underlying language model hallucinations and incorporating real-world cases of damages caused by AI customer service in e-commerce, this

文章引用: 刘梦哲. 电商平台 AI 客服“幻觉”致损的综合治理研究[J]. 电子商务评论, 2025, 14(12): 6614-6620.
DOI: 10.12677/ecl.2025.14124653

paper systematically analyzes the allocation of legal responsibilities among merchants, platforms, and AI service providers under different scenarios. The study finds that current governance efforts face multiple challenges, including technical bottlenecks, difficulties in legal characterization, and evidentiary hurdles. Accordingly, this paper proposes the construction of a tripartite comprehensive governance system from three dimensions: technological improvement, legal regulation, and policy guidance, with the aim of providing a theoretical reference for the standardized development of AI customer service on e-commerce platforms.

Keywords

AI Hallucination, Legal Responsibility, E-Commerce, Comprehensive Governance

1. 引言

近年来，人工智能技术驱动的客服系统在电商领域得到大规模应用，从简单的自动回复到复杂的售前咨询、售中支持及售后服务，AI 客服正逐步替代传统人工客服，成为电商平台与用户交互的首要接口。据行业统计，主流电商平台日均 AI 客服交互量已占客服总交互量的 70%以上[1]，这一趋势在“双十一”等大促期间尤为明显。AI 客服的广泛应用在降低人力成本、提升服务效率、实现 24 小时不间断服务等方面展现出显著优势，但同时也带来了新型技术风险——其中最为典型的便是 AI“幻觉”问题。

所谓 AI“幻觉”，是指大语言模型生成的并不符合其训练数据的自信反应，与事实不符或无法从源出处中得到验证的内容[2]。在电商语境下主要指 AI 客服自信地生成与商品特性、服务承诺、促销政策等事实不符的虚假信息，并以高度可信的方式呈现给用户，导致用户基于这些错误信息作出购买决策或采取相应行动，进而遭受经济损失或其他权益损害。OpenAI 在 2025 年 9 月发布的专题研究报告《Why Language Models Hallucinate》中明确指出，幻觉现象并非大型语言模型的可修复缺陷，而是其基于统计概率的技术架构下不可避免的产物。这一发现为理解电商平台 AI 客服的幻觉问题提供了科学依据。

电商平台 AI 客服幻觉已从理论风险转化为现实损害。2025 年 8 月，新华网报道了多起消费者利用 AI 工具伪造商品瑕疵图片骗取“仅退款”的案例，从侧面反映了 AI 生成内容滥用对电商交易秩序的冲击。与此同时，商家使用 AI 生成虚假商品展示、“种草笔记”及根本不存在的“权威机构认证证书”等行为，也进一步加剧了电商交易中的信息不对称问题。这些现象表明，AI 幻觉及其相关滥用行为正在侵蚀电商交易的信任基础。

在此背景下，厘清 AI 客服幻觉致损情形下的法律责任分配，不仅关系到消费者权益的救济效率，也直接影响电商平台的健康生态建设以及 AI 技术的负责任创新。然而，现有研究多集中于 AI 幻觉的技术机理分析，缺乏从电商实践角度探讨其法律责任的分配以及综合治理问题。本文试图从这一角度展开研究，通过案例分类与法理分析，构建一个适应电商平台 AI 客服特点的法律责任分配范式，并为综合治理 AI 幻觉问题提供路径建议。

2. 电商平台 AI 客服幻觉的概念性分析

2.1. AI 幻觉的定义与类型

在人工智能领域，“幻觉”特指模型生成看似语法正确、逻辑严谨，但实际上存在事实错误或无法

验证事实的现象。幻觉即“模型自信地生成不真实答案的情况”，在电商平台 AI 客服的应用场景中，研究将幻觉现象区分为“事实性幻觉”和“忠实性幻觉”[3]两种核心类型：

事实性幻觉是指 AI 生成的内容与客观事实不一致或存在事实捏造。在电商平台中，具体体现在 AI 客服错误地描述商品的材质成分、功能特性、产地信息等事实要素。在促销场景中，AI 客服虚构不存在的优惠活动或错误解释活动规则。直接导致“货不对板”或虚假宣传，构成对消费者知情权的实质侵害。

忠实性幻觉则指 AI 生成的内容与用户指令不一致、上下文不一致或存在逻辑矛盾。在电商平台 AI 客服的具体应用中，当用户询问特定尺寸商品的库存情况时，AI 客服提供错误库存信息，或在同一对话流程中前后回复矛盾，由于 AI 客服在语义理解与上下文跟踪方面的技术局限，导致服务体验的下降与决策误导。

2.2. AI 幻觉的产生原因

AI 幻觉的产生并非源于单一因素，而是语言模型的技术架构、训练数据与评估机制共同作用的结果。研究揭示了幻觉产生的双重根源：训练过程中的统计必然性和评估体系的结构性缺陷。为理解电商平台 AI 客服的幻觉现象提供了理论基础。

从技术架构角度看，基于转换器结构的大语言模型主要采用自回归生成方式，依据历史标记的概率分布逐词预测输出。这种机制缺乏对上下文语义一致性的把控能力，容易优先选择概率更高但与事实不一致的词汇组合，导致“语法正确但内容失真”的幻觉。从数字文化产业视角看，“算法黑箱带来 AI 幻觉现象与技术霸权，AIGC 生成式泡沫将增生版权风险与主体性危机”[4]。在电商客服这一专业化的场景中，AI 模型需要处理大量商品参数、促销规则等结构化信息，但其生成机制本质上仍依赖于统计模式而非事实核查，因而容易出现与用户真实需要相悖的幻觉。

训练数据的局限性是幻觉产生的另一关键因素。大模型基于互联网海量数据进行深度学习，但电商领域的专业知识更新迅速，训练数据往往滞后于实际商品信息与规则变化。通过 LDA 主题建模分析发现，“训练数据的源头性污染是导致幻觉生成的重要诱因。主流大语言模型往往以海量互联网文本作为训练基础，虚假信息、陈旧数据、歧视性内容广泛存在于网络生态中。模型在缺乏真实性过滤机制的条件下进行无差别学习，往往将错误信息编码为语言模式”[5]，特别是对于新产品、新活动等长尾信息，模型更倾向于依赖预训练阶段的通用知识而非当前上下文中的特定信息，从而产生事实性幻觉。OpenAI 通过数学证明表明，对于仅在训练数据中出现一次的事实，AI 的幻觉率至少等于这类事实在训练数据中的比例。更为明显地体现在电商场景中，易引发幻觉。

评估机制的缺陷则进一步加剧了幻觉问题。研究指出，当前主流的 AI 评估基准普遍采用正确计 1 分、错误或不知道计 0 分的二元评价体系，这种机制实质上激励模型在不确定时仍倾向于猜测而非拒绝生成。在电商平台的客服质量评估中，猜测偏好更为明显，AI 客服被训练为尽可能提供确定答案，即使其置信度较低，因为“我不知道”之类的回复通常会被系统视为服务质量低下。由于此类评估文化使得 AI 客服倾向于过度自信，从而提高了幻觉产生的概率。

ICLR 2025 会议上发表的研究进一步揭示了多模态大模型中的幻觉机理：当模型在最终输出层生成图像中不存在的虚假物体时，其早期中间层实际上能够正确判断该物体的存在性[6]。这证明语言模态的强知识先验在解码过程中逐渐压制了视觉证据，最终导致语义漂移。

2.3. AI 幻觉对电商交易的潜在影响

电商平台 AI 客服的幻觉现象不仅是一个技术问题，更是一个关乎交易公平与消费者权益的实际问题。其潜在影响主要体现在三个层面：

对于消费者而言，AI 幻觉可能导致其基于错误信息作出购买决策，产生直接经济损失。AI 客服错误地承诺某商品具备特定功能，或虚构优惠条件，诱导消费者下单。此外，幻觉还可能延误问题解决时机，如提供错误的退货地址或退款流程，导致消费者维权成本增加。

对于商家而言，AI 幻觉会带来商誉损害与法律风险。由于 AI 客服错误描述商品特性或作出无法兑现的承诺，商家会面临消费者投诉、行政处罚甚至民事诉讼。近期媒体报道了多起消费者利用 AI 工具伪造商品瑕疵图片以骗取仅退款的案例。有水果卖家遭遇买家使用带有不自然黑斑的 AI 图片指控芒果腐坏，玩具卖家收到显示玩偶开裂的伪造图片，甚至还催生了售卖退款教程的灰色产业，显著增加了商家的运营成本和平台的治理难度。这些通过“造假式索赔”牟取不当利益的行为不仅造成商家的直接经济损失，还增加了其审核成本与运营负担。

对于电商平台而言，AI 幻觉会侵蚀其信誉基础与生态健康。平台内置的 AI 客服若频繁产生幻觉，将削弱用户对平台整体可靠性的信任。在 AI 幻觉致损的责任认定中，平台可能因为监管不到位或应急处置不及时而承担相应法律责任。更重要的是，系统性幻觉可能扰乱平台正常交易秩序，形成诚实商家因严格审核而成本上升，不诚信用户利用 AI 造假技术获利的逆向选择之风，最终损害平台的可持续发展能力。

3. AI 客服幻觉对归责逻辑的解构与可验证性范式构建

随着人工智能与数字经济的深度融合，电商平台 AI 客服的幻觉问题，已经不能被传统的技术缺陷框架所完全涵盖。这本质是一种“认知层面的僭越”[7]，在这种机理下，AI 系统以其高度拟真但事实空缺的输出，动摇了法律归责赖以建立的过错认定基础与因果关系链条。因此，法律责任的分配必须超越对“过错”的简单追溯，转而构建一个核心在于追查“生成过程可验证性”的有限责任的新型范式。

3.1. AI 幻觉动摇归责的逻辑前提

生成式 AI 输出的是语法完备、格式规范但事实空缺、逻辑断裂的伪知识。也就是说，AI 客服生成的回复，往往在句式和语气上高度模仿人类专业客服，构建出一种符号权威。这种符号权威的底层逻辑是通过概率运算预测下一个词，而非对事实的指涉。它能以极其肯定的语气引用根本不存在的国家标准，杜撰出细节丰富的用户好评来吸引用户购买。当消费者基于伪知识作出购买决策时，损害便发生了。然而，在诉讼中，商家或平台完全可以以“一个理性的消费者不应完全信赖 AI 的单方面陈述”作为抗辩，试图切断法律上的因果关系链条。

传统法律中，判断行为人是否有过错要看其是否未尽到“合理注意义务”。但在 AI 幻觉场景下，注意义务的边界变得极其模糊。对于平台和商家而言，他们面临的困境是：即便投入了行业通行的技术成本，也无法从根本上杜绝幻觉。AI 的“算法黑箱”特性使得错误发生的原因是非线性、不可预测的，这就导致技术上的不可控性，正在消解法律上过错责任的前提，当损害发生时，责任很容易从技术的使用者向基于对人工智能输出内容的信任而作出决策的普通消费者转变。

3.2. “管理幻觉”的治理范式转向

鉴于完全消除幻觉在技术上不现实，治理思路应当从消除幻觉转向管理幻觉，在损害发生后实现公平的责任分配。明确在电商领域部署 AI 客服，核心注意义务在于确保其关键陈述具备可验证性。参考学界提出的“平台数据权利的双层结构”思想[8]，对 AI 客服致损的责任进行分层认定，在数据层表现为，对于材质、标准等事实性陈述，系统应通过权威检测报告提供可追溯的数据来源。对于优惠、送达时间等承诺性陈述，系统应生成不可篡改的电子承诺凭证，如果平台或商家的 AI 系统无法对生成的信息提供可验证的锚点，则一旦信息被证伪，无论幻觉原因为何，均应直接推定平台及商家存在过错；在映射层

面表现为，关注输出内容造成的经济损害，只要消费者能证明其基于 AI 客服的承诺产生了信赖并遭受了财产损失，且该承诺与实际情况不符，就应支持其合同撤销或损害赔偿的请求。

4. AI 幻觉治理在现行模式下的系统性困境

尽管前述框架在一定程度上指明了 AI 幻觉产生的原因及治理范式转向，但在落地实践中，电商平台 AI 幻觉的治理仍面临着一系列源于技术本质与制度适配的深层困境，构成了一个系统性的治理僵局。

4.1. 技术特性带来的治理约束

治理面临的首要困境，是 AI 幻觉的技术固有性与法律对可控性要求之间的矛盾。研究表明，幻觉源于概率生成范式下的表征偏移与训练数据偏差，是大型语言模型固有的技术特性。这种技术局限性导致法律规制面临过于严格的责任分配会抑制技术创新与应用的争议性障碍。

与此同时，电商领域 AI 客服的评估标准与质量规范尚付阙如。现有主流的 AI 评估基准普遍采用二元评价体系，这种机制实质上激励模型在不确定时仍倾向于猜测而非拒绝生成。在电商客服场景中，缺乏针对性的幻觉评估指标与测试基准，导致不同系统的幻觉率无法客观比较，监管部门也难以建立合理的准入门槛与最低性能要求。

另一个关键技术困境在于标识技术的局限性。《人工智能生成合成内容标识办法》要求 AI 生成内容需添加显式标识，但在电商客服的实时交互场景中，标识技术面临响应速度与准确性的平衡难题。若标识过于频繁，会造成用户体验下降；若标识不足，则无法有效提示风险。

4.2. 司法实践的不确定性

在诉讼程序中，AI 客服幻觉致损案件面临举证与验证的双重难题。《中华人民共和国电子商务法》第三十八条规定了平台在“知道或应当知道”侵权行为时应负有的连带责任。然而，AI 幻觉的不可预测性，为平台提供了抗辩理由，因为 AI 系统是在数以亿计的交互中随机产生幻觉，在技术层面无法实现针对特定风险的“知道”或“应当知道”。这使得平台责任被巧妙地悬置。

在举证环节，AI 幻觉纠纷呈现出证据不对称性。AI 对话记录是动态电子证据，极易被修改且不留痕迹。消费者手中的截图，在技术上无法与平台后台记录的包含完整元数据的日志相抗衡。在这种信息不对称格局下，无法严格遵循“谁主张，谁举证”的原则。虽然《最高人民法院关于民事诉讼证据的若干规定》第九十五条规定了证据妨碍规则，但该规则在 AI 幻觉案件中的具体适用标准尚不明确。

算法的黑箱特性使得 AI 决策逻辑不透明，导致归责困难。当就是否存在 AI 幻觉产生争议时，往往需要专业机构对系统进行测试验证。然而，电商平台 AI 客服通常处理海量并发请求，其输出可能因负载、上下文、模型随机性等因素而存在差异，使得事后复现特定幻觉场景面临挑战。此外，对于未开源的商业模型，其内部决策逻辑属于核心商业秘密，司法鉴定难以触及。即便获得鉴定许可，海量的非线性计算也使得鉴定结果无法被准确作出因果解释。法官最终也只能依赖专家的倾向性意见，这使得裁判结果充满了不确定性。

5. 电商平台 AI 幻觉的多维治理路径

5.1. 技术改进：从模型优化到全链路监控

治理电商平台 AI 客服幻觉，将技术改进作为第一道防线。根据 OpenAI 的研究启示，治理幻觉应从评测机制入手优化模型确定性。具体而言，即在电商 AI 客服评测中引入不确定性奖励机制，对恰当表达不确定性的行为给予加分，而非简单奖励回答行为。同时，电商平台可借鉴“信心阈值”[9]方法，在客

服决策流程中设置置信度门槛，只有当 AI 对答案的把握度高于特定阈值时才直接回答，否则就引导用户转向人工客服或表达不确定。借鉴新闻业的技术治理经验，应“构建 AI 幻觉的全链条防控体系，提升内容的准确性；建立 AI 生成内容审核机制，将人工核查嵌入 AI 生产全流程；开发专业幻觉检测工具，强化内容生产的技术防御能力”[10]。

数据质量的提升同样至关重要。针对电商数据动态变化的特点，建立知识库同步机制，确保 AI 客服系统能够及时获取最新的商品信息、规则变更与政策调整。引入检索增强生成技术，使 AI 客服在回答事实性问题时优先从权威商品数据库获取信息，而不是仅仅依赖模型内部的参数化知识。

在系统层面，构建全链路监控体系。电商平台对 AI 客服对话进行实时采样分析，建立幻觉风险预警机制。具体包括：设置关键词，若触发关键词，即引入人工客服审核；对高频率投诉对话模式进行追踪分析，筛选出幻觉率高的 AI 客服，对其进行调整升级；定期对 AI 客服进行幻觉率评估，并纳入服务质量考核体系。

5.2. 法律规制：完善责任体系与适用规则

针对 AI 客服幻觉致损的法律规制，需从立法完善与司法适用两个维度同步推进。在立法层面，可考虑在《消费者权益保护法》修订或电子商务专门立法中增设 AI 服务提供者的责任条款，明确不同主体在 AI 幻觉致损情形下的责任分配原则。特别是要明确平台在提供 AI 客服服务时的注意义务边界，以及免责或减责事由的认定标准。制定人工智能专门立法，明确 AI 系统的法律属性及缺陷认定标准。可借鉴欧盟《人工智能法案》的风险分级思路[11]，对电商客服这一特定应用场景设定差异化监管要求。

在司法层面，通过发布典型案例或制定司法解释的方式，为 AI 幻觉案件审理提供指导。明确 AI 客服幻觉致损案件的举证责任分配规则，在特定条件下可适当减轻消费者举证负担；界定《消费者权益保护法》第五十五条“欺诈”行为在 AI 客服幻觉场景中的认定，厘清平台责任与商家责任的区分标准。

此外，应强化行政监管与行业自律的协同作用。市场监管部门应针对电商 AI 客服发布行业指导规范，明确风险提示、应急处置等管理要求。鼓励电商平台建立 AI 客服专门投诉渠道与快速赔付机制，降低消费者维权成本。

5.3. 构建多元共治生态

治理电商平台 AI 客服幻觉，需要构建企业、政府、消费者共同参与的多元共治生态。在政策引导方面，将 AI 客服可靠性纳入电子商务示范企业评价体系，对幻觉率低，投诉处理及时的优质服务提供者给予政策倾斜。设立专项资金或税收优惠，鼓励企业研发幻觉抑制技术与解决方案。加快制定电商 AI 客服技术标准与服务规范，包括数据质量要求、测试评估方法等。推动建立跨平台共享的欺诈案例数据和模型，形成联合防控机制。

对消费者加强 AI 素养教育。电商平台在 AI 客服界面设置明确提示，告知用户对话对象为 AI 系统及其潜在的局限性，引导消费者对关键信息通过多种渠道核实。消费者协会发布 AI 客服使用指南，帮助消费者了解典型幻觉模式与维权渠道。从长远看，积极探索 AI 责任保险机制。鼓励保险公司开发针对 AI 系统风险的保险产品，为商家及平台提供责任风险保障，通过保费杠杆激励企业加强风险管理。

6. 结论与展望

电商平台 AI 客服幻觉致损的综合治理，涉及技术、法律与商业多个维度。本文分析表明，AI 幻觉源于概率生成范式下的技术与评估体系的结构性缺陷，在电商客服这一高交互场景中尤为突出。针对幻觉致损的责任分配，根据 AI 客服的致损原因及各方控制能力等因素进行综合判断，形成责权利相一致的分配方案。

当前，治理电商平台 AI 幻觉面临技术瓶颈、法律定性难题与举证障碍等多重困境，亟需构建“技术法律政策”三位一体的综合治理体系。在技术层面，通过信息阈值设置、动态校正解码等前沿技术降低幻觉频率；在法律层面，完善责任体系与适用规则，明确各方义务边界；在政策层面，加强标准建设与生态培育，形成多元共治格局。

从更宏观的发展视角看，治理应遵循自生、共生、再生的发展逻辑，建立多中心主导的高质量发展体系。随着《人工智能生成合成内容标识办法》等规章的实施，我国已初步建立 AI 治理框架。但电商平台 AI 客服的专门规制仍处于起步阶段。未来研究在以下方向仍有较大研究空间：一是探索基于区块链的 AI 对话存证机制，解决举证难题；二是研究适应电商场景的幻觉检测标准，为行业提供统一基准；三是设计合理的责任保险产品，保证新兴技术风险分担。

AI 技术的快速发展将持续重塑电商生态，只有通过技术创新、法律规制与行业自律的协同推进，才能在享受 AI 客服效率红利的同时，有效管控其幻觉风险，最终构建安全、可信、健康的电子商务环境。

参考文献

[1] 艾媒咨询. 2025-2026 年中国智能客服行业研究及消费者洞察报告[EB/OL]. https://www.iimedia.cn/c400/106439.html, 2025-06-06.

[2] 张铮, 张慧敏. AI 幻觉在文创中的应用可能与风险防范[J]. 南京社会科学, 2025(9): 137-147.

[3] 梁昭. AI“幻觉”: 认知困境、术语反思与范式嬗变[J]. 民族学刊, 2025, 16(8): 82-87+161.

[4] 解学芳. 数智时代数字文化产业高质量发展范式——基于“技术-文化-制度”模型[J]. 华中师范大学学报(人文社会科学版), 2025, 64(5): 166-176.

[5] 张梦晗, 沈文乾. 错位的升级与结构性失序: AI 幻觉主导的信息迷雾风险与分类治理可能[J]. 传媒观察, 2025(10): 53-63.

[6] Wang, C.X., Chen, X., Zhang, N.Y., *et al*. (2025) MLLM Can See? Dynamic Correction Decoding for Hallucination Mitigation. https://arxiv.org/html/2410.11779v1

[7] 曹咏萍. 技术与文化: 生成式人工智能幻觉的认知僭越及其治理转型[J]. 新经济, 2025(11): 56-69.

[8] 房慧颖. 人工智能时代平台数据权益的刑法介入机制[J/OL]. 理论与改革: 1-11. https://link.cnki.net/urlid/51.1036.D.20251127.1234.002, 2025-11-30.

[9] 李斌, 薛希萌, 张琪睿, 等. 人工智能对企业创新韧性的影响——基于技术能力适应性视角[J/OL]. 研究与发展管理: 1-13. https://doi.org/10.13581/j.cnki.rdm.20241839, 2025-11-23.

[10] 张迪, 张晔. AI 幻觉冲击下新闻业的守正路径[J]. 视听, 2025(21): 118-120.

[11] European Union (2024) https://eur-lex.europa.eu/eli/reg/2024/1689/oj