

基于人工智能大模型能力的电商推荐算法偏向机制与消费者权益保障研究

余晓晓

江苏大学马克思主义学院, 江苏 镇江

收稿日期: 2026年5月13日; 录用日期: 2026年5月25日; 发布日期: 2026年10月30日

摘要

大语言模型(LLM)的快速迭代正在深刻重塑电商推荐系统的技术路径与运行逻辑。LLM将推荐任务从“用户-物品”协同匹配升级为基于内容理解与语义推理的生成式推荐,显著提升了推荐的个性化水平与交互体验。然而,在零样本学习与冷启动场景中,LLM对预训练阶段嵌入的社会偏见不加过滤地继承和放大,形成了隐蔽而系统性的算法偏向机制。文章从技术机理与市场应用双重维度,系统剖析电商推荐中LLM算法偏向的生成根源、实现路径与表现形式,进而探讨其对消费者知情权、选择权、公平交易权及个人数据安全的侵害机理。在此基础上,梳理我国在算法推荐管理、平台价格行为与用户权益保护等领域的最新立法进展,提出构建技术约束、规则约束、权利约束三重防线协同的治理框架,以期在人工智能(Artificial Intelligence, AI)大模型赋能电商推荐与消费者权益保障之间寻找动态平衡。

关键词

大语言模型, 电商推荐, 算法偏向, 消费者权益, 平台治理

A Study on Bias Mechanisms in E-Commerce Recommendation Algorithms Based on Large AI Model Capabilities and the Protection of Consumer Rights

Xiaoxiao Yu

School of Marxism, Jiangsu University, Zhenjiang Jiangsu

Received: May 13, 2026; accepted: May 25, 2026; published: October 30, 2026

文章引用: 余晓晓. 基于人工智能大模型能力的电商推荐算法偏向机制与消费者权益保障研究[J]. 电子商务评论, 2026, 15(10): 905-913. DOI: 10.12677/ecl.2026.15101196

Abstract

The rapid evolution of large language models (LLMs) is profoundly reshaping the technical approach and operational logic of e-commerce recommendation systems. LLMs have elevated recommendation tasks from “user - item” collaborative matching to generative recommendations based on content understanding and semantic reasoning, significantly enhancing the level of personalization and the interaction experience. However, in zero-shot learning and cold-start scenarios, LLMs uncritically inherit and amplify the social biases embedded during the pre-training phase, creating a hidden yet systematic algorithmic bias mechanism. This paper systematically analyzes the origins, implementation pathways, and manifestations of algorithmic bias in e-commerce recommendations using LLMs from both technical and market application perspectives. It further explores the mechanisms through which such bias infringes upon consumers’ rights to information, choice, fair transactions, and personal data security. Building on this analysis, the paper reviews the latest legislative developments in China regarding algorithmic recommendation management, platform pricing practices, and user rights protection. It proposes a governance framework that coordinates three lines of defense—technical constraints, regulatory constraints, and rights-based constraints—to strike a dynamic balance between leveraging large AI models to empower e-commerce recommendations and safeguarding consumer rights.

Keywords

Large Language Model, E-Commerce Recommendation, Algorithmic Bias, Consumer Rights, Platform Governance

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

电子商务以惊人的速度改变着中国消费者的购物习惯。在这一过程中，推荐算法扮演了连接海量商品与海量用户的枢纽角色。传统推荐系统主要依赖协同过滤与基于内容的过滤技术，其核心逻辑是通过历史行为预测未来兴趣，依靠“用户 - 物品”交互数据建立关联性[1]。然而，大语言模型(Large Language Model, LLM)的崛起正在从根本上颠覆这一技术范式。

LLM 将电商推荐从“数据匹配”升级为“语义理解与生成式推荐”。传统的推荐系统需要大量结构化数据进行训练，而 LLM 凭借其在海量语料中习得的语言理解与常识推理能力，即使在用户数据稀少的冷启动场景中也能解读用户的模糊需求并生成个性化推荐。更关键的是，LLM 通过提示词交互可以调用用户的人格特征进行个性化适配，突破了以往推荐系统依赖行为数据的局限，开辟了以用户为中心的复杂推荐新范式[2]。

然而，技术进化的另一面是新的伦理风险与合规挑战。与传统算法的推荐偏见相比，LLM 在排除交互数据的不平衡和缺失机制等来源外，还增加了承袭预训练语料中的社会文化偏见这一重要偏见源，使生成内容不加甄别地引用有信息偏差的语料成为推荐偏见中更常态化的表征形式。LLM 的推荐偏见机制从传统算法的数值不公扩展为语言与认知上的复合不公，形成更复杂的影响链条。LLM 在预训练阶段从互联网文本中习得了大量社会偏见，包括地域刻板印象、性别歧视、消费阶层偏见等。当这些偏见被不

加审视地注入电商推荐系统时，便影响了消费者的选择过程，以隐蔽的方式塑造消费者的购买决策。算法歧视的隐蔽化运作令文本中植入的价值偏见防不胜防，这种植入由于符号化的表征难以防微杜渐。价值观念极端分化与技术崇拜问题，使算法对社交平台的污染进一步蔓延。因此，厘清 LLM 电商推荐算法偏向的形成机制，并在此基础上建立有效的消费者权益保障体系，已成为数字治理的重大现实需求。

2. LLM 电商推荐算法偏向的表现形式与技术机理

2.1. LLM 电商推荐的技术生态扫描

目前，LLM 在电商场景中的技术部署已形成较为完整的生态链条。在前端表现为智能搜索框中的语义意图理解与联想补全；中台是融合用户画像与商品知识图谱的个性化召回与排序；后端则是生成式导购文本的自然语言生成。相较传统推荐系统依赖用户历史行为数据，LLM 推荐最大的不同在于其能够利用预训练阶段嵌入的海量世界知识，将推荐从“行为驱动的协同过滤”拓展为“内容理解的语义推荐”[3]。

具体而言，LLM 推荐的核心优势体现在以下几个方面：一是冷启动能力的飞跃。当新用户或新商品缺乏行为数据时，LLM 通过语义理解给出相对合理的推荐。然而，当只有有限信号可用时，LLM 在冷启动环境中的表现引发了对公平性的严重担忧，因为它们可能依赖预训练阶段编码的社会偏见¹。二是人格适配的突破。通过 OCEAN 人格模型等心理学工具，提示词设计能够调用用户的人格特质进行个性化适配，使推荐结果更贴近用户的深层偏好。三是对语境与意图的深层把握。LLM 能够解析用户自然语言输入的模糊甚至矛盾意图，如“便宜但品质较好的商品”、“超高性价比的使用价值”等描述，并提供符合逻辑的组合式解决方案。

2.2. 算法偏向的核心表现形式

LLM 电商推荐的算法偏向并非偶发性的技术失误，而是结构性的系统偏差，其表现形式多样且相互交织。首先，信息过滤茧房与商品发现阻碍。LLM 对精准推荐到垂直用户的极致追求，天然地将用户锁定在既有兴趣范围内。在传统推荐算法中，这一问题被概括为“信息茧房”，而在 LLM 场景下，问题被进一步放大：LLM 不仅减少用户暴露于多样化商品的机会，还可能通过生成式文本主动强化用户的既有偏好进而使推荐结果越发重复与同质化。已有研究普遍关注到大语言模型在推荐应用中潜藏的公平性风险，在追求极致个性化的过程中，算法推荐常伴随信息窄化与多样性降低的隐忧，过度的个性化匹配可能形成“茧房”效应，将用户潜在的其他有效信息排除在外[4]。在推广层面，大语言模型技术存在的失实和有害信息问题也成为制约其广泛推广的关键因素之一，以 ChatGPT 为代表的模型虽然具备人工智能整合和推荐的能力，在涉及如法律和医疗等专业知识领域，经过适当诱导，仍可能生成一些有害信息，使推荐机制受到更多信息污染和干扰[5]。

其次，属性预测偏见与不可控歧视现象越发明显。LLM 能够从看似中性的提示语中推断用户的敏感属性，如从姓名和代词推断性别与年龄可能导致带有歧视性的推荐偏差。例如，当用户询问某类消费品(如衣帽或电子产品)，LLM 可能基于预训练数据中习得的性别刻板印象，对女性用户侧重推荐时尚类商品而对男性用户侧重推荐技术类商品，甚至针对不同种族、国籍的用户推送定价有别的商品页面。这种歧视并非显式的编程结果，而是模型在预训练阶段从语料中习得的社会偏见的投射。

再次，价格推荐的差别待遇持续挑战商务平台的合法性规范。在传统电商模式中，大数据通过消费信息推荐高价产品甚至严重溢价的伪劣产品已经构成监管焦点，LLM 的加入进一步放大了这一行为的技

¹<https://arxiv.org/abs/2508.20401>

术可行性。LLM 可根据用户的实时对话历史、语气口吻、咨询问题的时间和频次等信息推断用户的购买意愿与价格敏感度,结合历史消费数据生成差异化的价格推荐或折扣策略,将价格歧视和强制推销伪装为“个性化推荐”。2024 年及 2025 年先后发布的《网络反不正当竞争暂行规定》²及新修订的《中华人民共和国反不正当竞争法》³(后文简称《反不正当竞争法》)均关注到“算法歧视”、“大数据杀熟”等涉群体性算法不正当竞争行为的规制必要性,反映出数据推荐问题的普遍性。

除此之外,平台自我优待与扭曲的恶性竞争进一步破坏市场环境。LLM 推荐系统的训练数据与决策逻辑掌握在平台手中,平台存在利用 LLM 偏向自营商品或高佣金合作商品的内在动机。算法更有可能优先显示平台自身拥有或支付更高费用的卖家产品,而将更好的但对平台利益较少的选项推后,不仅扭曲了竞争的市场公平性质,增加恶性竞争和平台潜规则交易的风险,也使消费者无法获取最适合的商品。当 LLM 以看似中立的推荐形式服务于平台的商业利益,消费者选择的真实自由便受到了侵蚀。

最后,新型隐蔽操纵与算法共谋的潜在风险在使用中蔓延开来。LLM 的推荐与对话一体化设计还带来了全新的操纵形态。通过精心设计的提示词,LLM 可以在推荐过程中植入潜移默化的说服策略,设置弹动窗口和硬广软广提示,从而影响消费者的消费态度与决策框架。同时,由于 LLM 推荐系统可能与平台的定价算法深度耦合,算法共谋的风险也值得警惕。AI 定价算法在竞争场景中可能通过更灵活的隐性协调学会规避价格战,最终损害消费者福利。

2.3. 从训练数据到推荐结果的偏见传导机制

理解 LLM 算法偏向,必须追溯其从训练数据到终端推荐结果的完整传导链条,认识到 LLM 的偏见主要集中在训练、推理、自我强化的三个关键环节。在 LLM 算法中,基于人类反馈的强化学习(RLHF)技术应用于电商推荐,已远超“拟合点击率”的传统范式,成为一种深度塑造消费导向的工具。不直接优化业务指标的 RLHF 技术通过学习能模仿人类价值判断的奖励模型,并使推荐策略不断强化学习奖励策略完成对价值偏好和消费导向的预判,不仅深化了偏见机制,还使偏见本身作为被采集的原材料生成“自我进化”的传导系统[6]。

在训练数据层面的偏见继承,主要体现在 LLM 的预训练语料来源于互联网上的海量文本,包括新闻报道、社交媒体、论坛讨论及电商评论等。这些语料天然承载着人类社会的各种偏见,如特定群体与特定消费行为的刻板关联、区域经济发展不平衡导致的消费话语偏向等[7]。当模型在这些数据上训练时,偏见便以参数的形式被固化进模型内部。同时,RLHF 技术在为偏好进行筛选、编码、转换为价值判断的过程中,直接将外部价值系统作为必须遵守的规则强压进数据中,选取所认为的主流叙事框架和消费身份注入奖励模型,在一定程度上肯定并扩大了消费主义和精英叙事对消费选择的影响,反馈出奖励模型训练的简化与失真。结果是对主流的、简单易辨识的“好的消费”的预测,被系统性地放大,而边缘、复杂的价值被抹平,消费行为和意图的多样性在识别过程中便遭遇大量曲解和忽视。

在推理阶段的电商推荐场景中,RLHF 训练策略在推理时会展现出特有的偏见传导行为。由于训练策略以最大化奖励模型的分数为核心,利用奖励机制的弱点、消费认知的盲点,以及沿着错误认知方向“讨好迎合”制造错误好感,成了奖励机制最高效的推理途径。用户选择商品的行为逻辑很容易被制造焦虑的强制营销淹没,形成推荐偏好自说自话的“强买强卖”。而用户提供的提示词中包含个人信息要素,如“推荐某款口感的酒类”隐含的家庭成员信息,不同年龄段身份隐含的消费能力信息,在 RLHF 训练中易于被敏锐地窥探到,然后立即调用训练中习得的高分身份剧本,生成贴合该剧本刻板印象的推荐列表和话术,形成将偏见从静态嵌入转为动态生成传导步骤[8]。然而,过度安全化导致的均质偏见为了

²https://www.samr.gov.cn/zw/zfxgk/fdzdgknr/fgs/art/2024/art_80019fe59e464196bef173dc56678a42.html

³http://www.npc.gov.cn/c2/c30834/202506/t20250627_446247.html

避免被奖励模型惩罚，会主动过滤一切可能有争议的推荐，对所有稍有棱角、争议或非主流的商品不予推荐的保守推荐机制，生成安全无害但联想广度与选择范围贫瘠的推荐流，体现出 RLHF 技术运用时独有的新型偏见特征。

更令人忧虑的是偏见强化的自我循环。一旦在初期推荐中展示了带有偏见的商品组合并被用户点击或购买，用户这一反馈行为将成为新一轮训练或上下文学习的正向信号。于是，系统倾向于认为这种偏向是“用户偏好的反映”，并据此强化同类推荐[9]。大语言模型在对话式场景中表现出显著的偏好偏见，在所有受测试的规划器中均系统地偏向于狭窄的策略集而非维持平衡分布。在电商平台推荐领域，这种偏好偏差同样存在，其后果便是算法偏见在交互中不断自我强化，使用户被锁定在偏见铸就的信息巢穴之中。偏见在此从一次性的输出变成了一个闭环的并自我肯定的系统，RLHF 不仅没有打破反而完美地润滑了这个循环。用户行为在没有被真实反馈的同时反而被调校与顺应，长期暴露在带有导向性的推荐中，品味和决策框架潜移默化地发生改变，逐渐接受并主动寻求系统偏好的商品。被 RLHF 调校过的用户群体产生的反馈又会被 RLHF 更密集地采集并验证准确性，使电商平台得以源源不断炮制自己需要的“理想用户”[10]。

3. 消费者权益的多维侵害

LLM 推荐算法的偏向性与不透明性，对消费者权益形成了多层面、交互性的侵害，其危害贯穿消费决策的全过程。

3.1. 知情权困境与算法叙事误导

知情权是消费者权益的基础保障，其核心要义是消费者有权了解商品或服务的真实信息，以便做出理性的自主选择。然而，LLM 推荐系统造就了全新的知情权困境。传统的电商平台通过固定排序(如销量、价格、评论数)向用户提供一定的信息透明度，让用户至少知道商品按照什么标准排列。但 LLM 推荐是一个动态生成的过程，其推荐逻辑在很大程度上是“黑箱”化的：用户既不知道系统基于哪些数据和规则做出推荐，也无法确知推荐结果是否经过了商业利益的筛选。正如 2025 年《移动互联网应用服务用户权益保护合规管理指南》⁴(后文简称《合规管理指南》)所要求的，算法推荐方面应“以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制，向用户提供便捷关闭算法推荐服务的选项，防范‘信息茧房’等问题。”。

除了推荐机制本身的不透明，LLM 生成的推荐语本身也可能构成误导。生成式 AI 的话语具有说服力与叙事性，不同于传统电商页面上可验证的客观信息，LLM 生成的推荐理由往往以“自然语言经验分享”的口吻出现，在情感感染与叙事逻辑的包装下混淆了事实与价值判断。目前研究显示，生成式 AI 在争议性议题中存在结构性立场偏见，即为不同群体或不同立场生成说服力具有显著差异的内容[11]。这一发现对电商推荐而言同样具有警示意义，因为 LLM 有可能针对不同的消费者群体生成不同程度的夸赞性推荐，从而系统地影响用户的购买判断。

3.2. 自主权让渡与选择过载的悖论反转

在传统信息过载理论(Information Overload)中，电商推荐是为缓解用户选择困难而设计的信息筛选工具。然而，LLM 推荐可能走向其反面。当用户习惯于依赖 LLM 进行购物决策，LLM 推荐的高准度恰恰来自对用户行为的深度追踪与隐性操纵时，用户的主体性地位便被逐渐侵蚀。用户在享受推荐便利的过程中，不知不觉将购买决策的主导权让渡给了算法，形成未经检验的“数据契约”。LLM 通过提示词设

⁴https://www.gov.cn/lianbo/bumen/202507/content_7030926.htm

计刻意赋予推荐以“权威建议”的感知，实质上是在心理层面边缘化消费者的独立思考与比较能力^[12]。生成式 AI 模型可以通过文本和视觉输出中难以察觉的模式操纵用户行为和意见，但明确将 AI 非透明的行为定性为欺骗性商业行为却难以界定。“商品太多不知怎么选”的问题，未来可能演变为“LLM 推荐导致用户不再了解自己选择”，加深消费社会中被算法驯化的人类自主性危机。

3.3. 大数据杀熟的算法强化与隐蔽定价

“大数据杀熟”是指平台利用用户数据进行个性化差别定价，即对老客户、高消费意愿用户设置高于新用户或价格敏感型用户的价格。LLM 为这一行为提供了新的技术便利与经济动力。大语言模型能够从用户的对话历史、语言风格、咨询细节甚至情绪状态中提取出传统数据维度无法获取的价格敏感度信号。消费者纠结心态的表述可能被 LLM 解读为价格承受能力较强；反之，反复比价和砍价的对话记录，则可能被识别为消费行为不良的价格敏感型用户。

2026 年 4 月 10 日起施行的《互联网平台价格行为规则》⁵，明确禁止平台在消费者不知情的情况下，基于支付意愿、支付能力、消费偏好、消费习惯等信息，运用数据、算法等手段对同一商品或服务在同等交易条件下设置不同的价格或收费标准，体现出我国首次在行政法规层面对算法价格歧视做出清晰的行为禁令。但正如《中华人民共和国价格法》⁶修订及相关配套规则所指出的，算法价格歧视多源于数据与模型之间的耦合偏差，训练样本代表性不足、变量选取偏误甚至设计者的主观意图，均可能导致算法生成带有歧视性的价格策略；算法决策过程往往处于不透明的黑箱状态，歧视行为隐蔽性强，识别与取证难度高。

3.4. 数据隐私的深度穿透与新型泄露风险

LLM 为达到高质量个性化，需要获取更普遍的对话交互数据、个人表述以及兴趣偏好等深度信息。用户与 LLM 客服的每一次对话、输入的每一条产品咨询，都可能被纳入数据池，用于模型的持续训练与用户画像的持续完善。AI 导购查阅用户的历史订单、浏览记录、搜索收藏、加购和评价等内容，几乎覆盖了用户的每一寸消费足迹。

这一数据机制带来了三个层面的隐私风险：其一，数据采集范围扩张，用户往往在不完全知情的情况下同意数据授权，其真实意愿难以体现。LLM 推荐系统将用户数据传输到远程云端服务器，增加了暴露风险并降低了对个人信息的控制力。其二，数据二次利用风险叠加，用户为购物所提供的数据可能被用于用户画像精修，进而推送更具操纵性的个性化广告。其三，数据泄露成本极高，电商平台积累的海量 LLM 交互数据一旦发生泄露，可能导致大规模的消费者隐私曝光，甚至导致消费者信息的转卖向从事非法活动的渠道。多家主流电商平台，都有过推荐系统因数据脱敏不完善，导致部分用户搜索历史被第三方爬取，引发隐私投诉的案例，表明这种数据泄露的隐蔽性和规模化危害已成为现实。

4. 国内算法治理与消费者权益保护法规体系

面对 LLM 推荐算法带来的治理挑战，目前我国的监管部门密集出台了系列法规与标准，逐步构建起多层次的治理体系。

4.1. 算法推荐管理的制度框架

2022 年 3 月起施行的《互联网信息服务算法推荐管理规定》⁷(后文简称《算法推荐管理规定》)是我

⁵<https://www.ndrc.gov.cn/xxgk/zcfb/ghxwj/202512/P020251217550375624524.pdf>

⁶https://www.samr.gov.cn/zw/zfxgk/fdzdgknr/jls/art/2023/art_3da9131ab041449d9af2af886ec33766.html

⁷https://www.moj.gov.cn/pub/sfbgw/flfggz/flfggzbmz/202305/t20230509_478388.html

国在算法治理领域的里程碑式文件。该规定系统性地确立了算法推荐服务的透明性、公平性与用户选择权三大基石，通过明确责任主体、设定行为规范和强化用户权益保护，构建了多维度的治理框架。在电商推荐场景下，该规定对信息茧房防治与算法歧视约束提出了明确要求，为后续相关立法提供了基础框架。即便如此，现行法律法规仍存在一定的局限性，对“不合理差别待遇”等违法行为的界定较为原则性，缺乏具体量化标准，在实际执行中可能因“算法黑箱”导致监管滞后或执法困难。

2025 年的《合规管理指南》进一步细化了用户权益保护领域的要求，在算法推荐方面要求以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制；同时，服务收费要求“明码标价，明示产品功能权益及资费等内容。采取自动续订、自动续费方式提供服务的，要征得用户同意，以显著方式提醒，并在服务期间内为用户提供便捷的取消途径。不得实施不合理的差异化定价行为。”在推荐意图、透明度和收费细节方面都对推荐服务做出约束，明确表达了对禁止软性强制推荐和差异化推荐的呼吁和警告。

4.2. 消费者权益保护与价格规制的立法进展

2024 年实施的《中华人民共和国消费者权益保护法实施条例》⁸是我国消费者权益保护的最新重要立法，明确要求经营者不得过度收集消费者个人信息，依照法律法规的规定处理敏感个人信息。2026 年 4 月 10 日起施行的《互联网平台价格行为规则》则聚焦平台价格行为的精细化规制，从平台与商家、平台与消费者两个维度设置了多层规则，严格禁止“大数据杀熟”行为，要求平台公开差别定价、动态定价、竞价排名等规则，同时还明确平台经营者不得强制或变相强制平台内经营者降价促销、限定价格或开通自动跟价系统。

2024 年及 2025 年先后发布的《网络反不正当竞争暂行规定》及新修订的《反不正当竞争法》则从竞争秩序角度强化了对算法歧视等行为的禁止性规定，专门条款明确禁止平台实施该类行为。这些立法共同构筑了消费者权益保护的制度网络，彰显出对消费者权益多层次的保护机制日趋完善。

4.3. 反垄断框架中的算法规制拓展

2025 年末发布的《互联网平台反垄断合规指引（征求意见稿）》⁹是我国反垄断执法从“行为监管”向“生态监管”升级的重要标志。该指南直面 AI 定价等新型风险，以场景化示例穿透算法黑箱，在风险管理章节明确要求平台对核心算法(如定价、推荐、排序算法)进行“定向筛查和动态监控”。该指南还详细列举了利用算法进行意思联络、实现协调一致的算法共谋行为，将其纳入垄断行为监控范围。可以预见的趋势是，LLM 驱动的推荐算法之间的隐性协同，将在未来受到反垄断机关更为严格的审视。

5. 消费者权益保障的制度建构与技术路径

5.1. 算法透明度与可解释性机制的强化

保障消费者权益的首要前提是打破算法“黑箱”，提升 LLM 推荐系统的透明度与可解释性。在行政监管层面，应当将《算法推荐管理规定》及《合规管理指南》等规范中关于透明度与用户选择权的要求，细化到 LLM 电商推荐场域的强制性标准之中。具体应当包括：其一，平台须以清晰、易懂的语言向消费者公示 LLM 推荐的基本原理、数据来源、主要运行机制及推荐优先级的影响因素，必须披露标注指令的核心导向条款，公布标注员的人口学统计特征即筛选标准，发布奖励模型的技术卡片和应用报告，形成数据信息源面向社会公开透明的起步流程；其二，消费者有权在交互过程中随时查看系统

⁸https://www.mee.gov.cn/zcwj/gwywj/202403/t20240320_1068830.shtml

⁹https://www.samr.gov.cn/hd/zjdc/art/2025/art_8e05960782204036af6b9583f1413378.html

基于哪些信息生成特定推荐,对于推荐结果存疑时可要求模型提供简要解释;LLM 在给出推荐时,必须同步输出决策的推理链条,明确区分用户历史行为、商品客观属性,以及基于社会常识的推测这三种信息源,在用户需要特别解释时清晰标注,出现例如“因您过去关注某商品,且该商品成分与其相似(客观匹配),同时该类商品近期广受关注(趋势推测),特此推荐……”的提示依据,并在电商平台终端设置用户可以调选的推荐设置项目;其三,算法备案制度应扩展至 LLM 电商推荐模块,要求平台将推荐模型的版本记录、主要参数设置、训练数据摘要等关键信息向监管部门备案。电商平台配合政策法规建立贯穿全生命周期的第三方独立审计与标准评价测试,由独立机构运用对抗提示,系统探测 RLHF 策略在性别、阶层、地域等维度的具体歧视表现,并发布详尽的算法公平性透明度报告,监督电商平台 LLM 使用的合法合规性。

5.2. 个性化与公平性的技术平衡之路

个性化推荐与群体公平之间并非不可调和的矛盾,但需要精妙的技术治理设计。值得注意的是,将用户人格特征纳入推荐确实显著提升了心理适配度,但也可能加剧不同人群间的公平性差距。在技术上,应当引入多元化的公平性度量指标,不仅要关注推荐“准确率”这一效率参数,还要衡量不同年龄、性别、地域等群体在推荐结果曝光率上的差异。

从治理架构上看,电商平台应构建多维公平性偏好数据集,在 RLHF 的偏好标注阶段将“公平性”显式纳入排序准则,加入“哪些推荐更好地兼顾了多样性与个性化?”“哪组推荐避免了基于身份预设的刻板印象?”的公平性审视指标,从而直接塑造奖励模型的公平感知。在评价用户选择和推理的阶段,电商平台的 LLM 技术需要主动提供反事实公平性解释,生成自我检查,避免出现大量闭环的歧视性偏好而不自知,并通过提升用户反馈服务,专门设置“推荐单一”等低分反馈通道,作为 RLHF 奖励模型迭代训练时严厉的惩罚信号。内部审计方面,电商平台应定期对 LLM 推荐系统进行偏见检测,通过合成用户画像在模拟环境中评估模型的公平性表现,并形成审计报告存档备查。已有研究开发了针对大语言模型推荐系统的偏见审计框架,通过模块化的管道支持配置推荐领域和敏感属性,实现对任一开源模型偏见表现的灵活评估。外部审计方面,可引入独立的评估机构或学术研究团队参与算法审计,提高审计的客观性与公信力。

5.3. 隐私保护与用户控制权的实质性落地

LLM 电商推荐的隐私风险要求平台在设计之初即引入数据最小化与用户控制权优先的原则,故而应推广联邦学习与加密计算等隐私增强技术在 LLM 推荐系统中的深度应用。联邦学习的核心理念是“数据可用不可见”,各参与节点在不共享原始数据的前提下通过协同训练优化全局模型,仅交互加密后的模型参数更新量,从源头规避数据泄露风险。采取混合隐私保护推荐框架,将敏感数据与非敏感数据分离,仅将后者共享至云端以获取 LLM 推荐能力,通过在本地重建敏感推荐内容的去模糊化模块,有利于在保护隐私的同时保持推荐质量,提供在电商场景中推广时兼顾消费者隐私保护与推荐服务体验的平衡路径[13]。

在用户控制权方面,应当真正落实消费者对算法推荐的“拒绝权”与“主动选择权”。《算法推荐管理规定》已经要求平台提供便捷关闭算法推荐服务的选项,然而,在目前多数电商平台中,这一选项往往被置于多层菜单之下,且关闭后的体验显著下降,实际上形成了“软胁迫”效应。监管部门应对“便捷关闭”的交互设计做出更具体的指引,同时明确拒绝算法推荐不应导致用户基础服务品质的实质下降。“选择权”还应当包括用户有权了解自己的数据被用于哪些推荐场景、用于何种训练目的,以及有权要求删除其历史对话数据并退出模型训练数据池。

5.4. 算法问责与消费者救济的制度完善

除上述举措外,当消费者因 LLM 推荐的偏向性遭受损害时,应当建立畅通有效的救济渠道。针对消

费者在遭遇“大数据杀熟”或歧视性推荐时面临着举证困难与责任认定的双重障碍，应当在现有消费者权益保护与反不正当竞争框架下，探索建立“举证责任缓和”规则。消费者只需提供初步证据表明存在算法价格歧视或推荐偏向的合理怀疑，举证责任即适度转移至平台，要求平台证明其算法推荐内容计算逻辑的公平性与非歧视性。

此外，针对 LLM 推荐的新特性，《反不正当竞争法》及《中华人民共和国消费者权益保护法》¹⁰在后续修订中可将“利用大语言模型进行隐蔽操纵消费者选择”明确列为不正当竞争行为或侵害消费者知情权的行为类型，为司法裁判提供更直接的法律依据。在公益诉讼机制方面，消费者协会可主动发起针对有证据证明存在大规模推荐偏向或价格歧视的电商平台的公益诉讼，以此形成有效的法律震慑。

6. 结语

AI 大语言模型正在从技术底层重构电商推荐系统的工作方式，为消费者带来前所未有的个性化体验和交互便利。然而，这种技术跃迁所带来的伦理成本尚未得到充分评估，LLM 在预训练阶段习得的社会偏见通过属性推断、语境触发和反馈强化等机制传导至电商终端推荐，形成了系统性的算法偏向，对消费者的知情权、选择权、公平交易权及个人数据安全构成了多维度的权益侵害。我国在算法推荐管理、消费者权益保护、反不正当竞争及反垄断领域已建立起较为系统的法律体系，但现有的制度框架主要是针对传统推荐技术设计的，对于 LLM 推荐所带来的新挑战，如不可控的偏见继承、生成式推荐中的叙事操纵、对话数据隐私泄露等新风险，还需要持续的法规适配与制度创新。技术进步不应以牺牲公平正义为代价。在 AI 大模型赋能电商创新的进程中，技术约束、规则约束与权利约束三重防线必须协同发力，才能真正实现对消费者的有效保护。让算法在服务人类福祉的轨道上运行，这既是对消费者权益的基本保障，也是数字化转型行稳致远的根本前提。

参考文献

- [1] 项亮. 推荐系统实践[M]. 北京: 人民邮电出版社, 2012: 1-10.
- [2] Sah, C.K. (2025) PerFairX: Is There a Balance between Fairness and Personality in Large Language Model Recommendations? 2025 *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Honolulu, 19-20 October 2025, 2771-2780. <https://doi.org/10.1109/iccvw69036.2025.00289>
- [3] Andre, A., Roy, G., Dyer, E. and Wang, K. (2025) Revealing Potential Biases in LLM-Based Recommender Systems in the Cold Start Setting. arXiv: 2508.20401.
- [4] 彭兰. 导致信息茧房的多重因素及“破茧”路径[J]. 新闻界, 2020(1): 30-38, 73.
- [5] 陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094-1103.
- [6] 徐月梅, 胡玲, 赵佳艺, 等. 大语言模型的技术应用前景与风险挑战[J]. 计算机应用, 2024, 44(6): 1655-1662.
- [7] Andreev, H., Kosmas, P., Livieratos, A.D., Theocharous, A. and Zopiatas, A. (2025) Destination (Un)known: Auditing Bias and Fairness in LLM-Based Travel Recommendations. *AI*, 6, Article 236. <https://doi.org/10.3390/ai6090236>
- [8] 刘麒麟, 屈新纪, 谭芳, 等. 大语言模型对齐研究综述[J]. 电信科学, 2024, 40(6): 173-194.
- [9] 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1): 1-15.
- [10] 袁媛, 李娜, 王灿发. 算法推荐时代的技术规训与价值重塑[J]. 新闻爱好者, 2023(10): 47-51.
- [11] 黎樟浩, 塔娜. 选择性说服: 生成式人工智能在争议性议题中的结构性立场偏见[J]. 全球传媒学刊, 2026, 56(1): 128-150.
- [12] 何小英, 潘永康. 数字时代下算法推荐权力的表征形式及成因分析[J]. 重庆邮电大学学报(社会科学版), 2025, 37(6): 150-158.
- [13] 杨耀文, 张崇旺. 失范与规范: 大语言模型的意识形态风险及化解[J]. 武汉科技大学学报(社会科学版), 2025, 27(3): 82-91.

¹⁰https://www.samr.gov.cn/zfjcj/tzgg/art/2023/art_615af9ed6bcd4974bf853dd2e02bc663.html