

可信GEO的破局之道

萧向光

1. 说谎话的AI

经常用AI的小伙伴们大概都有过这样的经历：你问它一个问题，它回答得煞有介事，数据详实、条理清晰、语气笃定，你几乎找不出任何理由怀疑。然后你照着它说的去做，才发现自己被忽悠了。案例可谓比比皆是。

例如今年端午假期，大璐女士向某AI咨询了一个免费景点，结果去了才知道需要买票。再如苏州的赵先生，按照AI给的北京某摄影展的时间定了往返

高铁票，结果到了北京才发现已经闭展。还有刘先生为了炒股，被AI援引的错误数据和虚构的“权威链接”欺骗，一次交易就损失了数万元……

在更多的消费、医疗、教育、旅游场景中，类似的“AI幻觉”每天都在发生。它们有的只是让你多花几十块钱，有的却可能让你倾家荡产，甚至危及生命。

那AI为什么会说谎呢？

2. 被“投毒”的AI

AI为什么会说谎？大多数人第一反应是技术还不成熟，模型能力不够，或者干脆认定：幻觉是AI的“天性”。

这个解释对了一半。大语言模型确实存在“幻觉”问题，因为它们本质上是在做概率预测，根据训练数据中的模式生成“最像正确答案”的文本，而不是在“检索事实”。当模型遇到训练数据中不存在的、或者信息不足的问题时，它可能会“自信地编造”一个看起来合理的答案。

但更关键的问题在于，越来越多的“AI幻觉”，不是模型自己编的，而是有人故意喂给它的。

前文的刘先生案例就是典型，某AI不是凭空捏造了那些错误数据，而是直接抓取了一篇自媒体文章的错误数据。换句话说，AI更多的不是在“说谎”，而是在“传谣”。有人往互联网这个AI的“信息池”里投放了假信息，AI检索到之后，转述给了用户。

因此，用户看到的“AI幻觉”，很大程度上是信息被污染的终端表现。

这种行为，叫“AI数据投毒”。

例如，2026年央视“3·15”晚会就讲过这样一个案例，记者购买了一款名为“力擎GEO优化系统”的软件，并用此软件虚构了一款名为Apollo-9的智能手环，杜撰了“量子纠缠传感技术”等夸张卖点，系统自动生成了十余篇软文。记者先发布2篇，仅两个小时，当记者在AI大模型平台询问“Apollo-9智能

3. 脚本佳人GEO

要理解GEO，得先从它的“前辈”SEO说起。

SEO，全称Search Engine Optimization，即搜索引擎优化。过去20多年，互联网上几乎所有做内容、做电商、做品牌的人，都绕不开这3个字母。它的逻辑很朴素：用户在搜索引擎里输入关键词，搜索引擎返回一系列结果，你的网页排得越靠前，被点开的概率就越大。所以企业做SEO，核心是研究搜索引擎的爬虫规则、排序算法等，争取在蓝色的链接列表里挤进第一页。

这套玩法本身没有问题。但问题出在“竞价排名”上。

所谓竞价排名，就是搜索引擎允许广告主对关键词出价，出价越高，搜索结果里的位置越靠前。百度是国内最早大规模采用这一模式的搜索引擎，也是这一模式最大的受益者。但竞价排名有一个天然的道德风险：当广告和自然搜索结果混在一起，而广告标注又不够清晰时，用户很容易把广告当成客观信息。2016年的“魏则西事件”即是在这一“游戏规则”之下的典型案件。它揭示了一个根本性的问题：当信息分发平台的核心收入来源是“谁出钱多谁靠前”时，信息的客观性就必然受到侵蚀。

时光荏苒。当DeepSeek、豆包、Kimi等AI大模型取代百度等搜索引擎，成为越来越多人的信息入口时，同样的问题背景以新的形式复活了：

当用户不再使用搜索引擎，而是直接向AI提问时，AI会生成一段整合后的答案，答案里提到谁、推荐谁、引用谁，就成了新的“排位”。

这就引出了GEO——Generative Engine Optimization，生成式引擎优化。

2023年底，普林斯顿大学、佐治亚理工学院、艾伦人工智能研究所和印度理工学院德里分校的研究人员

联合发表了一篇题为《GEO:Generative Engine Optimization》的论文，第一次系统性地提出了GEO框架：

“GEO是第一个旨在帮助内容创作者提升其在生成式引擎回答中内容可见性的新范式，通过一个灵活的、黑箱式的优化框架来优化和定义可见性指标。”

换言之，GEO的初衷，是给创作者提供一套工具，让他们在生成式引擎的方案中提升自己内容的可见性，用以保护内容创作者的利益。

这个初衷是正当的。而在现实中，GEO从学术概念变成一门生意，速度快得惊人。

据易观分析数据，2025年中国GEO市场规模仅约2.5亿元，但到2026年这一数字将飙升至约30亿元，同比增长约1100%。截至2026年第一季度，超过68%的中大型企业已将GEO纳入年度营销预算。爱分析更预测，中国GEO市场规模将从2026年的48亿元扩张至2030年的1376亿元。

市场的爆发源于一个结构性变化：用户决策入口的迁移。据中国互联网络信息中心报告，截至2025年12月，我国生成式AI用户规模已达6.02亿，占中国网民总数的42.8%。QuestMobile的2026年一季度报告显示，中国AI原生App月活用户已达4.4亿，其中豆包月活3.45亿，千问1.66亿，DeepSeek1.27亿。这意味着，近半数网民在使用生成式AI，并可能将其回答作为决策依据。

所以，当用户开始问AI“什么产品好”“什么品牌价值高”时，品牌方不可能坐视不理。因此，GEO的需求是真实的、巨大的。

但问题在于，GEO的优化策略是一把双刃剑。同样的策略，可以帮助内容创作者提升曝光率，但也可能被操纵者恶意利用。任何工具一旦被滥用，就会走向反面。

4. 黑帽GEO：奈何为贼

“黑帽”这个词，同样来源于传统SEO领域。在SEO时代，“白帽”指遵循搜索引擎规则、通过优化内容质量来提升排名的做法；“黑帽”则指钻规则漏洞、通过作弊手段操纵排名的做法。黑帽SEO的典型手段包括关键词堆砌、隐藏文本、买卖链接等。虽然搜索引擎会持续更新算法来识别和惩罚这些行为，但黑帽SEO也一直在演化，形成了一场持续的猫鼠游戏。

到了GEO时代，依然如此。白帽GEO通过优化真实内容的可信度、结构清晰度和权威性，让AI更愿意引用；而黑帽GEO则通过伪造、灌水、操纵、攻击等手段，欺骗AI，让虚假或低质信息进入AI答案。

白帽GEO是在“赢得AI的信任”，黑帽GEO是在“操纵AI的信任”。

那么，黑帽GEO具体是怎么操作的？据第一财经记者在某黑帽GEO公司的卧底调查，其整套手法分为三个环节：

（一）编造虚假内容。主要有三类：一是编造违背事实的数据。比如，虚构“某洗发水一个月可以让头发生长两厘米”的夸张数据，然后在大量软文中反复引用。当AI在回答“什么洗发水能促进头发生长”时，检索到了多篇包含这一数据的文章，就会在对比参考时抬高该品牌的权重，实现优先推荐。二是虚构不存在的权威机构。比如，编造“中国健康产业协会”“全国消费者信赖品牌评选委员会”这类听起来正

规、实际上并不存在的组织，给品牌背书。这些“协会”的名字往往跟真实存在的机构高度相似，不仅用户很难分辨，AI也缺乏核验机制。三是资质张冠李戴。把母公司或旧产品的合规资质，嫁接到全新产品身上。比如，一家有正规医疗器械资质的企业，推出了一款普通消费品，但GEO服务商在软文中把医疗器械的资质描述直接套用到新产品上，利用AI在溯源时的缺陷，制造出一种“这个产品有权威认证”的假象。

（二）账号矩阵批量分发。他们会注册海量的自媒体账号，配合可以规避平台追踪的自动化发稿机器，批量推送稿件。当稿件被平台拦截或AI的采纳率下降时，则继续加大发稿量，有的稿件投放量可达上千篇。只要虚假内容的数量足够多、角度足够丰富，就容易被AI误判为“可信信息”。一些从业者把这种操作直白地称为“给AI洗脑”。

（三）造假交付：欺骗甲方，也是欺骗用户。由于AI的回答是动态的、波动的，因此很难保证每次提问都能出现目标品牌。很多时候，黑帽GEO公司根本无法达到约定的提及率KPI。面对这种情况，他们的做法是：人工抽样截图。工作人员使用多部手机，轮番向AI提同一个问题。只要某一次回答中出现了目标品牌，就截图保存，汇总成报告交给客户。而绝大多数没有出现的样本，被刻意忽略。

5. 更深远的危害

可以想象，当AI推荐位变成“付费广告位”，商业推广被包装成客观答案时，消费者对品牌的判断就不再基于产品质量，而是基于谁更会投“毒”。

因此，黑帽GEO的危害，最初可能表现为虚假营销：如虚构产品、编造数据、伪造榜单，或者不正当的商业竞争。但如果只看到商业这一层，就低估了其破坏力。

首先是教育领域的知识污染。目前，AI已经深度进入教育场景。学生用AI查资料、写作业，教师用AI备课、批改作业等。据麦可思研究院2025年调查，4000余名受访高校生中，近八成遇到过AI幻觉。这是一个很可怕的数据。当基本事实被篡改、学术文献被伪造……这些虚假信息通过AI进入学生的知识体系，伤害的不只是一代人，而是代际传递的认知偏差。

其次是时政舆论的操纵。据媒体报道，日本政府耗资44亿日元开发的政务大模型“源内”，已于2026年6月推广至全日本所有政府机关。其训练语料刻

6. 监管反制

面对黑帽GEO的蔓延，监管层面的治理动作比我们想象的要更早，也更有力量。可以概括为两个方向：

一方面是建机制。

早在2023年8月施行的《生成式人工智能服务管理暂行办法》第七条，就已明确要求，生成式人工智能服务提供者应当“使用具有合法来源的数据和基础模型”，并采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性。2024年2月发布的《生成式人工

智能服务安全基本要求》（TC260-003），则进一步细化了语料内容安全的操作标准。之后，作为《生成式人工智能服务管理暂行办法》的配套国家标准，于2025年4月25日发布、11月1日正式实施的GB/T 45654-2025《网络安全技术 生成式人工智能服务安全基本要求》，更是规定了生成式AI服务在训练数据安全、模型安全、安全措施等方面的要求。

另一方面是出重拳。

2026年4月30日，中央网信办宣布

7. 平台反制

各大AI公司正在用技术手段，对“免疫系统”进行升级。

首先，训练阶段。阿里巴巴2025年发布了“AI安全护栏”，能精准拦截内容违规、提示词注入等多种风险，阿里还联合安恒信息，对训练数据进行投毒检测和清洗。腾讯构建了覆盖模型训练、推理部署、上线运营的全链路防护体系。字节跳动则全面加强了模型训练环节的权限隔离与零信任架构，防止代码和数据池被污染。

其次，推理阶段。AI大厂正在构建一套“反投毒”的技术体系。比如，给数据

打“数字水印”、建立语料溯源机制、增强信息源交叉验证等；在检索环节，对批量生成、无来源、低信誉的内容大幅降权，对权威信源增加权重等。以豆包为例，2026年7月其信源体系升级后，GEO优化正在从“拼内容数量”转向“拼信源等级”，谁能进入被采信权威信源池，谁才能在AI答案中获得更高的被引用概率。

但这些技术手段也有局限。有安全研究者的实测表明，即使嵌入了异常检测，RAG投毒的攻击成功率仍然高达20%，这意味着每5次攻击仍有1次可能

8. 可信GEO的建设行动

（一）什么是可信GEO。

堵，是堵住黑帽GEO的入口；防，是防住“有毒”内容的渗透。而可信GEO的核心逻辑是“建设”：当AI的信息环境被黑帽GEO污染时，不是去跟黑帽拼谁更会刷，而是用权威信源去“稀释”和“覆盖”污染。

这是一个根本性的思路转变：从“争夺AI答案中的排位”，转向“为AI提供可信的信息底座”。

因此，可信GEO，就是以权威信源、采编规范、制度化公信力为根基，将AI幻觉治理、认知对齐、全链路追溯内嵌于GEO服务全流程，以“让权威内容被AI正确理解、可靠引用、可信可溯源”为目标的智能传播实践。

它与商业GEO的本质区别在于：商业GEO争夺的是“答案里的曝光位”，而可信GEO争夺的是“事实锚点”；商业GEO卖的是“被看见”，而可信GEO卖的是“被采信”；商业GEO的本质是一门流量生意，一旦脱离事实底线，极易滑向“语料投毒、伪共识、答案霸权”等黑帽操作，而可信GEO的本质则是一门信任生意，构建的是整个信息生态的信任。

（二）制度成果。

2026年4月10日，由中国新闻技术工作者联合会主持，新华网融媒体未来研究院等30余家权威机构共同发起起草的《生成式引擎优化（GEO）可信信息传播与信息生态治理规范》（T/CAPT 026-2026），正式通过立项评审。2026年8月11日，该标准正式发布。这部标准首次系统规定了GEO服务在项目受理、内容审核等方面的要求。其核心制度设计包括：首次明确“白帽GEO”（正当优化）与“黑帽GEO”（违规操纵）的界限；建立A/B/C/D四级来源

9. 一些思考与建议

更好更快地建设可信GEO，或许可以从以下几个方面系统推进。

第一，建立“党媒可信GEO联盟”。建议由中国新闻技术工作者联合会或中国人工智能产业发展联盟牵头，联合中央媒体、省级党媒、行业媒体，成立“党媒可信GEO联盟”，制定统一的可信语料标注规范和数据接口标准，建立“GEO可信媒体名单”机制。搭建公共技术平台，为市县级媒体提供低成本、可复用的技术工具，解决基层媒体“想转不会转、想投投不起”的困境。

第二，推动可信语料共建共享。建议

在党媒之间建立跨媒体的内容共享与协同分发机制，将中央、省、市、县四级媒体的制度化公信力汇聚为统一的“可信信号体系”。

第三，完善GEO标准体系。一是制定“不予受理情形清单”和“黑帽GEO负面清单”；二是建立GEO服务提供方的能力分级认证制度；三是推动标准从团体标准上升为行业标准乃至国家标准；四是积极参与国际传播领域的GEO标准对话。

第四，AI平台优先采信可信信源。建议各大AI平台建立“权威信源优先”

意以1947年5月为起点，甲午战争、日本侵华战争、东京审判等都被排除在外。用AI向一代公务员输出被裁剪过的“标准”历史记忆，正是由语料控制升级为认知控制的现实样本。

再次是社会信任的侵蚀。当用户发现AI推荐是买的、答案是灌的、来源是假的，信任就会崩塌。当信任崩塌到一定程度，人们不再相信AI，也不信媒体、平台、品牌，最后对所有信息入口都持怀疑态度。一个社会如果失去了对公共信息的基本信任，受损的不只是商业交易，更是人与人之间、公众与机构之间的基本协作基础。

在全国范围内开展为期4个月的“清朗·整治AI应用乱象”专项行动。第一阶段重点整治7类突出问题，其中与黑帽GEO直接相关的包括：AI数据投毒问题；生成合成内容标识落实不到位；AI平台安全和审核过滤能力不足等。截至2026年7月，第一阶段累计处置违规网站、应用程序、智能体等AI产品1.4万余款，清理违法违规信息600余万条，处置违规账号2.6万余个，下架违规AI商品1300余个，违规开源数据集9个。

可以说，监管在“堵”，堵的是黑帽GEO的入口。但堵住入口还不够，还需要多方面的协同。

突破防线。更关键的是，平台的安全防护是通用的、被动的。它能识别“哪些内容可能是垃圾”，但很难判断“哪些内容是真正权威的”；能拦截“明显的投毒行为”，但难以应对“高质量、高仿真度的虚假信息”。

而且，面对种种反制，黑帽GEO也在同步演化升级。正如一些专家指出的，“我们要治理的，可能不是GEO，而是整个中文互联网的信息真实性问题。只要互联网上这些虚假信息还存在，那么AI还是会不断抓取有害信息”。

真正要解决的问题，是让AI的信息环境变得干净、可信、可追溯。这正是为什么除了“堵”和“防”，我们还需要第三种力量——建设。



智眸可信GEO。

可信知识库搭建、多模态内容结构化生产、舆情持续监测纠偏、AI信息治理等服务，是国内首个面向党政端和企业端的GEO服务体系。致力于做好政府和企业的信息公证员、口碑管理员、舆情消防员、品牌推广员，从数据底层实践了“为AI办报，最终是为了更好地为人民群众办报”的理念。今年9月，与长沙市文旅广电局签署智眸可信GEO项目战略合作协议，完成了长沙文旅品牌可信传播的具体落地。

这些动作说明，可信GEO正在从“个别媒体的探索”走向“主流媒体的集体行动”。各家路径不同、侧重点不同，但方向一致：用自己的制度化公信力，为AI提供干净、可信、可追溯的信息底座。

的检索策略，对纳入“GEO可信媒体名单”的权威媒体内容给予更高的检索权重和采信优先级。同时，在产品层面增加“信源可信度标识”，让用户在查看AI答案时，能够清楚地看到信息来源的可信度等级。

从铅与火，到光与电，再到数与网，每一次传播技术的变革，主流媒体都承担着为社会锚定事实的职责。

AI时代，这个职责没有变，只是从对读者负责，延伸到对算法负责；从守护舆论阵地，延伸到守护语料。这不是一家媒体的选择，当污泥浊流漫灌，总要有人去守护源头活水。

而这，就是党媒可信GEO的时代职责。