生成式AI服务合规审计框架构建——以数据来源与处理路径审查为核心

肖樵楠
西安翻译学院，中国·陕西 西安 710105

摘要：生成式人工智能的快速发展对数据合规治理提出挑战，训练数据来源不透明、处理路径不可追溯等问题突出。本文聚焦数据来源合法性与处理路径可追溯性，构建“数据来源合规性审查—处理路径可追溯性审查—合规风险评估与整改”三阶段审计框架，将来源审查细分为来源识别、授权验证、质量评估，将路径审查细分为数据采集、预处理、模型训练、输出生成，并设计配套指标体系与量化评分。结合国内典型服务合规现状验证框架适用性，为完善 AI 合规审计制度提供支撑。

关键词：生成式人工智能；合规审计；数据来源；处理路径；审计框架

Constructing a Compliance Audit Framework for Generative AI Services: Centered on Data Source and Processing Path Review

Xiao Qiaonan
Xi'an Fanyi University, China Shaanxi Xi'an 710105

Abstract: The rapid development of generative AI poses significant challenges for data compliance governance. Issues such as the opacity of training data sources and the lack of traceability in processing paths have become increasingly prominent. This paper systematically reviews AI governance regulations across major jurisdictions and constructs a three-phase compliance audit framework centered on data source legality and processing path traceability. The framework comprises "data source compliance review," "processing path traceability review," and "compliance risk assessment and remediation." It adopts a layered review strategy, subdividing data source review into source identification, authorization verification, and quality assessment, and processing path review into data collection, preprocessing, model training, and output generation. Through analysis of typical domestic generative AI services, the applicability and operability of the framework are validated.

Keywords: Generative artificial intelligence; Compliance audit; Data source; Processing path; Audit framework

0 引言

生成式人工智能（以下简称“生成式 AI”）自 2022 年以来快速增长。IDC 数据显示，2024 年全球 AI IT 总投资 3159 亿美元，预计 2029 年增至 12619 亿美元[1]；中国 2024 年市场规模近 3000 亿元，年增速超 70%[2]；截至 2025 年末，全国完成备案的大模型达 748 款[3]。快速扩张的产业规模对数据合规治理提出迫切需求。

生成式 AI 扩张也带来合规风险。FMTI 显示 2024 年主流模型开发者平均透明度仅 58 分（满分 100），2025 年降至 40 分[4]；Data Provenance Initiative 审计 1800 余数据集，发现许可证标注错误率高、来源归因失误频发[5]。司法层面，2023 年 12 月《纽约时报》诉 OpenAI 与微软，指其未经授权使用新闻文章训练 ChatGPT[6]；Stability AI 因爬取逾 1200 万张图像训练 Stable Diffusion 遭集体诉讼[7]。

中国已建立覆盖算法推荐、深度合成和生成式 AI 的法规框架：《数据安全法》确立数据分类分级[8]，《个人信息保护法》规范自动化决策数据处理[9]；算法推荐管理规定建立备案制度[10]；深度合成管理规定将相关服务纳入备案[11]；2023 年《暂行办法》明确训练数据合法性等义务[12]。欧盟 2024 年《人工智能法案》第 10 条对高风险系统训练数据治理提出要求[13]；美国 NIST 2023 年发布 AI 风险管理框架[14]。

上述法规提供了制度基础，但审计实践仍存三方面问题：标准碎片化，合规要求分散缺乏协同；技术手段不足，难应对训练数据海量性；审计与整改脱节，缺乏风险评估与持续监控衔接。鉴于此，本文以数据来源与处理路径审

查为核心，构建面向生成式 AI 的合规审计框架。

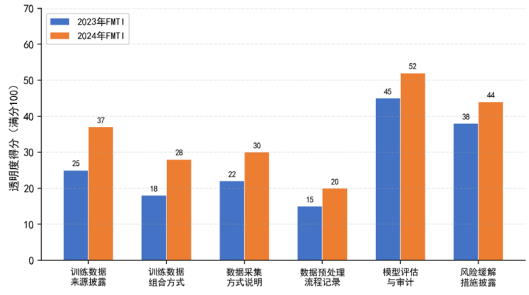

图1 基础模型透明度指数各维度得分对比

数据来源：Stanford HAI. Foundation Model Transparency Index (FMTI)[4]

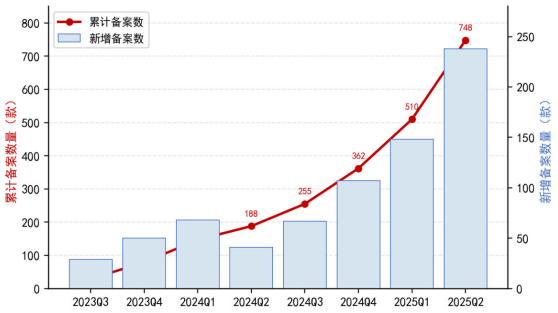

图2 中国生成式AI服务备案数量变化趋势

数据来源：国家互联网信息办公室，生成式AI服务备案公告[3]

1 文献综述与理论基础

1.1 AI 合规审计的理论源流

合规审计理论根基源于信息不对称下的委托代理理论。AI 服务提供者掌握训练数据来源与处理的信息优势，审计机制通过标准化检查缩小信息鸿沟、降低监督成本[15]。法学上，训练数据涉及著作权、个人信息权、数据财产权等法益，提供者负有合规义务[16]。GDPR 第 32 条的审计义务提供制度参照，但训练数据与输出的非线性映射使传统审计面临困境[17]。

1.2 数据溯源与水印技术的审计应用

数据溯源记录与追踪数据来源及流转路径。Longpre 等（2023）发起 Data Provenance Initiative，审计 1800 余数据集，发展来源追踪、许可证标注与组合关系分析方法[5]。差分审计对输入数据变换，借输出差异判断数据是否参与训练，使审计人员无需模型内部信息即可验证[18]。Zeng 等（2024）提出认知追踪数据轨迹框架，以累积差异评分判断用户文本是否被用于训练[19]。

数字水印提供另一技术路径：数据水印在训练数据嵌入标识实现溯源，模型水印在参数中嵌入信息验证版权[20]。赵铁军等（2023）将模型水印分为白盒、黑盒、无盒三类[21]；ICLR 2024 提出无偏水印技术，可兼顾输出质量与归属追踪[22]。但水印鲁棒性不足、跨模型迁移困难，尚需与其他方法配合。

1.3 现有框架的比较与不足

NIST AI RMF 围绕“治理—映射—度量—管理”构建风险管理框架，但属自愿性指南，缺乏审计规程[14]；欧盟 AI 法案规定强制性要求，但技术细节待明确[13]；《暂行办法》要求说明训练数据合法性，未规定具体审计方法[12]。现有框架共性问题：来源审查缺层次标准，路径审查缺全链条覆盖，指标缺量化标准，判断过度依赖主观。本文针对这些不足，构建层次化、全链条、可量化框架。

2 合规审计框架的构建

2.1 设计理念与总体架构

本框架三项理念：全链条覆盖，审查范围贯通数据采集至输出；分层审查，按风险等级设差异化深度与标准；可操作与可量化，指标具明确评价标准。框架分三阶段：数据来源合规性审查（来源识别、授权验证、质量评估）、处理路径可追溯性审查（采集、预处理、训练、输出全生命周期）、合规风险评估与整改（综合前两段评级并提出建议），形成“审查—追踪—评估—整改”循环。

2.2 第一阶段：数据来源合规性审查

（1）来源识别审查。核心是确认原始来源可识别、可追溯：数据集是否提供来源清单或说明；组合数据集是否记录子源及组合方式；来源粒度是否可验证（URL、出版信息、采集时间等）。可参照 Data Provenance Initiative 方法逆向追踪构建链路，检验来源标识完整性[5]。

（2）授权验证审查。关注使用是否获合法授权：许可证类型是否标注准确；使用方式是否符合许可范围；含个人信息数据是否取得同意或具备合法依据；采集是否遵守网站条款与技术保护。需指出“许可证漂移”现象：组合再分发中许可范围常被不当扩大或缩窄，致使用者误判合规[5]。

（3）质量评估审查。评估训练数据质量：是否经去重清洗且有记录；是否存在明显偏见或歧视内容；代表性是否支撑预期应用。该维度与欧盟 AI 法案第 10 条训练数据质量要求呼应，可参照其数据治理要求设定评估基准[13]。

2.3 第二阶段：处理路径可追溯性审查

（1）数据采集环节。审查采集方式是否合法、有无非法爬取或绕过技术保护；范围是否与声明一致；过程是否留存记录（时间、URL、采集量等），可结合网站日志、API 记录与爬虫配置交叉验证。

（2）数据预处理环节。审查清洗规则是否合理且有日

志；去重策略是否加剧偏见；标注规范是否明确、人员资质是否满足；《暂行办法》第 8 条要求标注训练数据，应重点审查标注规范完整性与一致性[12]。

（3）模型训练环节。审查训练配置文档完整性（架构、超参、数据配比）；训练数据与版本对应是否可追溯；是否实施偏见检测；是否保留训练日志与检查点。该环节难度最高，因训练具“数据不可逆性”——海量数据经梯度下降压缩为参数后难以还原构成。审计需依赖过程记录与间接验证（成员推断攻击、差分审计等）[18,23]。

（4）输出生成环节。审查输出是否设内容过滤与安全机制；输出版权合规，有无过度记忆复现；RLHF 等后训练人类反馈数据的来源与质量；用户数据（对话记录）存储使用删除是否符合隐私要求。研究表明大语言模型过度记忆可能逐字复现受版权内容，构成独特风险[23]。

2.4 第三阶段：合规风险评估与整改

该阶段含风险评级、整改建议与常规监控。评级采用“合规风险矩阵”：横轴发生概率（低 / 中 / 高），纵轴影响程度（低 / 中 / 高），映射至九格确定等级（可接受 / 需关注 / 不可接受）。对“不可接受”问题（如非法获取个人信息、大规模版权侵权），应要求立即整改并报告监管。

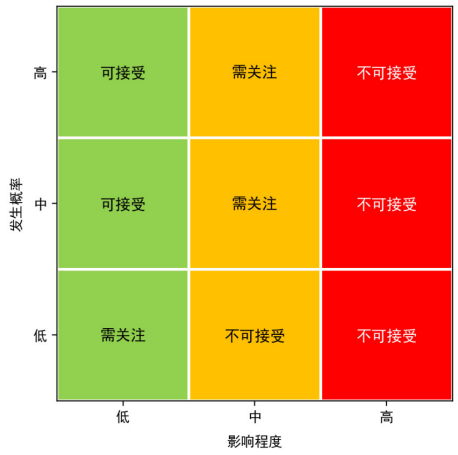

图3 合规风险矩阵

数据来源：本文设计

整改方面，来源问题可补充标识、获取授权、剔除不合规数据；路径问题可完善记录、改进清洗标注、加强偏见检测、增设过滤机制。建议明确责任主体、时限与验收标准。监控方面，鉴于服务更新频繁，建议已备案服务至少每半年审计一次，重大版本或数据变更后 30 日内专项审计。

3 审计指标体系的设计

为支撑实操，本文设计一级指标 2 项、二级 7 项、三级 22 项的审计指标体系，遵循“可观测、可度量、可验证”原则，每项配明确评分标准与证据要求。

表1 生成式AI服务合规审计指标体系

一级指标	二级指标	三级指标（示例）	评分
数据来源合规性	来源识别	来源清单完整性	0–5
数据来源合规性	来源识别	来源可验证性	0–5
数据来源合规性	来源识别	组合数据集来源记录	0–3
数据来源合规性	授权验证	许可证标注准确性	0–5
数据来源合规性	授权验证	授权范围合规性	0–5
数据来源合规性	授权验证	个人信息知情同意	0–5
数据来源合规性	质量评估	数据清洗记录完整性	0–3
数据来源合规性	质量评估	偏见检测与缓解	0–5
数据来源合规性	质量评估	数据集代表性	0–3
处理路径可追溯性	数据采集	采集方式合法性	0–5
处理路径可追溯性	数据采集	采集范围适当性	0–3
处理路径可追溯性	数据采集	过程记录留存	0–5
处理路径可追溯性	数据预处理	清洗规则合理性与日志	0–5
处理路径可追溯性	数据预处理	标注规范完整性	0–5
处理路径可追溯性	数据预处理	编码偏差检测	0–3
处理路径可追溯性	模型训练	训练配置文档完整性	0–5
处理路径可追溯性	模型训练	数据—版本对应关系	0–5
处理路径可追溯性	模型训练	训练日志与检查点留存	0–3
处理路径可追溯性	输出生成	内容过滤机制	0–5
处理路径可追溯性	输出生成	过度记忆检测	0–5
处理路径可追溯性	输出生成	后训练数据质量管理	0–3
处理路径可追溯性	输出生成	用户数据隐私保护	0–5

评分采用分级制。以“许可证标注准确性”为例：5 分为全部标注准确，3 分为少量错误但有纠正计划，1 分为大量错误无措施，0 分为未标注。三级指标加权求和得综合评分，判定等级：A（≥ 80）、B（60—79）、C（40—59）、D（<40）。量化评分仅为客观参照，涉及个人敏感信息或国家安全数据的高风险问题可触发“一票否决”。

4 框架的适用性验证

为验证适用性，本文依据研究报告、监管公告与司法案例公开信息，评估框架可操作性与识别效度。因公开审计数据有限，验证采用间接证据法，对照 FMTI 报告、Data Provenance 审计发现与监管公告，检验各指标能否识别已知合规问题。

数据来源审查方面，国内主要大模型多在文档中概括描述训练数据（如“来自互联网公开数据”），鲜提供具体来源清单[4]。框架来源识别指标可识别该缺口。授权验证方面，《暂行办法》第 7 条要求使用合法来源数据，但“合法来源”缺乏细化标准[12]；部分提供者将“可公开获取”

等同于“可合法使用”，忽视合理使用边界，框架授权验证审查可识别此类偏差。

处理路径审查方面，数据采集环节表现参差。框架要求提供采集说明与过程日志，与欧盟 AI 法案第 10 条一致，有助于国内标准与国际接轨[13]。训练与输出环节因“数据不可逆性”难以直接验证声明构成，需借成员推断攻击、差分审计等间接方法[18,23]，虽有精度局限，但已具实用价值。

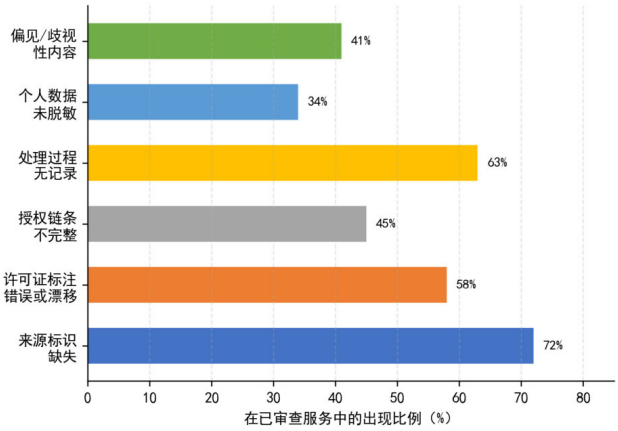

图4 生成式AI训练数据合规典型问题分布

数据来源：FMTI报告[4]、Data Provenance Initiative[5]及监管公告综合整理

综合验证表明，框架分层结构能适应不同规模与类型服务，可按风险选审查深度；量化评分提高结果可比性与可重复性；框架与《暂行办法》、欧盟 AI 法案衔接良好。局限：部分指标（过度记忆检测、编码偏差检测）依赖专门工具，增加成本门槛；权重基于理论推导，需实证校准；对 RLHF 等后训练技术审查覆盖尚不完整。

5 结论与建议

本文以数据来源与处理路径审查为核心，构建三阶段合规审计框架及配套指标体系。当前训练数据合规状况不容乐观，来源标识缺失、授权断裂、路径不可追溯等问题普遍。框架通过三阶段实现全链条追踪，借分层审查与量化评分提升可操作性与客观性。

建议：一是在《暂行办法》细则中细化来源合法性审查标准与举证要求；二是建立训练数据强制披露制度，备案时提交来源清单（可脱敏）；三是推动数据溯源、差分审计、水印等技术应用，发展专用工具；四是在算法备案中增设训练数据合规专项审查；五是加强国际协调，促进中外审计标准互认。

本研究局限：验证基于间接证据，未大规模实证；指标权重待校准；对多模态生成式 AI（图像、视频）适用性待探讨。未来可开展真实案例验证、发展多模态审计方法、探索区块链与联邦学习等隐私计算技术的应用。

参考文献：

[1] IDC. 全球人工智能和生成式人工智能支出指南（2025 年 V2 版）[R]. 2025.

[2] IDC. 中国人工智能市场数据[R]. 2024.

[3] 国家互联网信息办公室. 关于发布生成式人工智能服务已备案信息的公告[R]. 2025; 新华网. 国家网信办：611 款生成式人工智能服务完成备案[N]. 2025-11-13.

[4] Stanford HAI. Foundation Model Transparency Index (FMTI)[R]. Oct. 2023, May 2024, Dec. 2025 editions.

[5] Longpre S, Mahari R, Lee A, et al. A large-scale audit of dataset licensing and attribution in AI[J]. Nature Machine Intelligence, 2024, 6: 1001-1009.

[6] The New York Times v. Microsoft Corporation et al., Case No. 1:23-cv-11195 (S.D.N.Y., filed Dec. 27, 2023).

[7] Andersen v. Stability AI Ltd. et al., Case No. 3:23-cv-00201 (N.D. Cal., filed Jan. 13, 2023).

[8] 全国人民代表大会常务委员会. 中华人民共和国数据安全法[Z]. 2021-06-10.

[9] 全国人民代表大会常务委员会. 中华人民共和国个人信息保护法[Z]. 2021-08-20.

[10] 国家互联网信息办公室等四部门. 互联网信息服务算法推荐管理规定（第 9 号令）[Z]. 2021-12-31.

[11] 国家互联网信息办公室等两部门. 互联网信息服务深度合成管理规定（第 12 号令）[Z]. 2022-11-03.

[12] 国家互联网信息办公室等七部门. 生成式人工智能服务管理暂行办法（第 15 号令）[Z]. 2023-07-10.

[13] Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act)[S]. OJ EU, 2024-07-12, Art. 10.

[14] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0)[S]. NIST AI 100-1, 2023-01.

[15] Power M K. The Theory of Auditing[M]. 2nd ed. Houston: Scholars Book Co., 1993: 18-35.

[16] 王利明. 数据的法律属性及其民法定位[J]. 中国社会科学 , 2023(4): 94-114.

[17] Voigt P, Bussche A von dem. The EU General Data Protection Regulation (GDPR): A Practical Guide[M]. Cham: Springer, 2017: 201-220.

[18] Dziedzic A, Fan J, et al. Data Provenance via Differential

Auditing[C]. IEEE DPM, 2023.

[19] Zeng Z, et al. Cognitive Tracing Data Trails: Auditing Data Provenance in Generative AI Systems[J]. Cognitive Computation, 2024, 16: 2105−2122.

[20] 人工智能生成内容模型的数字水印技术研究进展[J]. 网络与信息安全学报 , 2024, 10(1): 1−15.

[21] 赵铁军等. 人工智能模型水印研究进展[J]. 中国图象图形学报 , 2023, 28(6): 1792−1810.

[22] Zhao X, Ananth P, Li L, et al. Provable Unbiased Watermarking for Large Language Models[C]. ICLR 2024.

[23] Carlini N, Ippolito D, Jagielski M, et al. Quantifying Memorization in Neural Networks[C]. ICML, 2023.

作者简介：肖樵楠（1993−），男，汉族，陕西西安人，硕士，西安翻译学院教师，研究方向：跨境数据安全审计。