

基于分期BERTopic与非平衡最优传输的我国人工智能安全政策主题演化研究

谢桂平, 郑 静

杭州电子科技大学经济学院, 浙江 杭州

收稿日期: 2026年9月23日; 录用日期: 2026年10月14日; 发布日期: 2026年10月27日

摘 要

目的/意义: 随着人工智能技术的快速发展, 其安全风险成为贯穿模型研发、部署与应用全流程的综合性问题。系统梳理我国人工智能安全治理的内容演变, 对进一步加强安全治理具有重要意义。方法/过程: 本文选取2018~2025年的人工智能安全政策文本和行业报告作为研究语料, 按两年为窗口划分为四个时间片, 采用分期BERTopic模型来识别主题内容; 结合切片Wasserstein距离和非平衡最优传输构建跨期主题流量关系, 据此识别主题演化路径。结果/结论: 研究发现, 我国人工智能安全主题呈现明显的阶段性特征, 政策关注点由早期的训练数据与算法公平逐步转移至个人信息处理、生成内容治理、模型攻击防护及综合应急处置。这些主题表现为延续、分裂与重组并存的多层次结构特征。本文所识别的多种主题结构关系, 为把握治理重心的迁移方向、完善系统性治理提供了方法支撑。

关键词

人工智能安全, 主题演化, BERTopic, 非平衡最优传输

Evolution of AI Safety Policy Topics in China: A Period-Specific BERTopic and Unbalanced Optimal Transport Approach

Guiping Xie, Jing Zheng

School of Economics, Hangzhou Dianzi University, Hangzhou Zhejiang

Received: September 23, 2026; accepted: October 14, 2026; published: October 27, 2026

Abstract

Purpose/Significance: The rapid development of artificial intelligence (AI) has made safety risks a

文章引用: 谢桂平, 郑静. 基于分期 BERTopic 与非平衡最优传输的我国人工智能安全政策主题演化研究[J]. 统计学与应用, 2026, 15(10): 133-143. DOI: 10.12677/sa.2026.1510234

comprehensive issue spanning model development, deployment, and application. A systematic examination of the evolution of AI safety governance in China is of great significance for further strengthening safety governance. Method/Process: AI safety policy documents and industry reports published between 2018 and 2025 were collected and divided into four two-year periods. BERTopic was applied separately to each period to identify topics. Sliced Wasserstein distance and unbalanced optimal transport were then combined to construct cross-period topic flows and trace topic evolution paths. Result/Conclusion: AI safety topics in China exhibit distinct characteristics across periods. Governance attention has progressively shifted from early concerns about training data and algorithmic fairness to personal information processing, generated-content governance, protection against attacks on AI models, and comprehensive emergency response. These topics exhibit a multilayered structure characterized by the coexistence of continuation, splitting, and recombination. The identified topic-structure relationships provide methodological support for understanding shifts in governance priorities and improving systematic governance.

Keywords

Artificial Intelligence Safety, Topic Evolution, BERTopic, Unbalanced Optimal Transport

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

作为新一轮科技革命和产业变革的关键引擎,人工智能技术正重塑社会运行的各个环节,相应的人工智能管理制度也在逐渐完善。2017年国务院发布《新一代人工智能发展规划》¹,将安全监管纳入总体部署;2021年、2022年《互联网信息服务算法推荐管理规定》²《互联网信息服务深度合成管理规定》³相继出台,算法与合成内容进入到监管阶段;2023年《生成式人工智能服务管理暂行办法》⁴进一步将管理范围延申到训练数据、模型输出与服务提供行为;2024年的《人工智能安全治理框架》⁵区分内生安全风险与应用安全风险,治理对象由技术系统延伸至社会应用场景。

但从整体上看,现有政策对技术变化的回应仍显滞后,分散的制度规则尚未贯通模型生命周期。早期政策主要围绕数据安全展开,随后开始关注算法运行可能产生的不公平结果。生成式人工智能广泛应用后,模型输出是否真实成为新的政策问题,模型运行过程的安全要求也随之增加。风险影响不再停留在技术系统内部,而是借助平台应用进入社会运行过程。与此同时,大模型的迭代周期逐渐加快,政策制定则需要在风险识别、部门分工、责任配置和处置程序之间形成协调,技术演进速度与制度供给节奏之间的差异由此逐渐拉大。技术风险并不会停留在单一模型内部,而可能借助平台、工具和应用系统向下游传导,在此背景下,系统梳理人工智能安全政策的主题分布及其演化脉络,识别治理重心的迁移方向与薄弱环节,成为完善治理体系的重要前提。

2. 文献综述

随着基础模型和生成式人工智能进入社会应用,相关研究内容由开始的模型训练和算法问题,逐步

¹https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

²https://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm

³https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm

⁴https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

⁵https://www.cac.gov.cn/2024-09/09/c_1727567886199789.htm

转变为模型幻觉、深度伪造、有害内容生成、知识产权侵害以及平台化传播造成的社会影响。国外学者从不同角度进行研究，Bommasani 等[1]指出，基础模型具有较强的通用性，其缺陷会随着模型迁移进入不同应用场景，并在技术链条中持续扩散。Weidinger 等[2]进一步从歧视、信息危害、恶意使用和人机交互等方面构建语言模型风险分类，说明人工智能安全已经具有跨场景传导和多主体关联的特征。Bender 等[3]将研究视线转移到模型开发阶段，指出训练数据的来源与使用方式会影响模型的表现，模型规模增长并不意味着安全能力同步提升。Raji 等[4]提出端到端的算法审计框架，将审计嵌入数据收集、模型开发、部署测试与运行监测等环节，强调人工智能风险治理不能局限于模型上线前的单次评估，应形成覆盖全生命周期的责任追踪机制。

国内学者为研究 AI 风险治理重心的变化，开始通过政策文本挖掘潜在语义主题。魏弋等[5]基于全球主要国家和国际组织发布的人工智能政策进行主题建模，发现治理关注由早期的技术发展和原则倡导逐渐转向风险防范与制度约束；李亮等[6]运用扎根理论分析 170 份政务数据安全政策，发现政策治理已不再局限于安全技术配置，而是延伸至组织责任和制度控制。张涛[7]进一步分析中央与地方政策的扩散特征和主题转移，提出政策议题会随制度扩散发生内容迁移。周文泓等[8]通过政策文本构建人工智能数据治理框架，将治理要求贯穿研发、部署与应用全过程。这些研究表明主题模型能够从大规模文本中提取潜在议题，并揭示政策关注重点的变化。

综上所述，现有研究已经形成从人工智能风险类型识别到治理体系分析、从传统文本计量到深层语义主题建模的研究脉络，为理解人工智能安全议题及其政策响应提供了重要基础。但仍旧存在一定的局限性：现有研究采用主题关键词变化来描述不同阶段政策关注的差异，同时，多以相似度来匹配跨时期主题，不能充分揭示议题之间的联系[9]。基于此，本文将研究期划分为若干等距时间片，分别使用 BERTopic 识别各阶段的主题，在保留内部文本语义分布的基础上，进一步以切片 Wasserstein 距离衡量相邻时期主题的分布差异，借助非平衡最优传输方法建立跨期主题间的流动关系。本文将研究重点由主题数量和热度变化转向议题结构重组，为解释我国人工智能安全治理由技术风险防范向系统性治理拓展提供经验证据。

3. 研究设计

3.1. 研究框架

本文的研究框架如图 1 所示。具体而言，首先将研究语料等距离划分为四个时间片，识别各个阶段的主题内容，再比较相邻时期主题内部文本的语义分布，进而通过非平衡最优传输识别主题之间的质量流动。最终根据显著流量的来源、去向及识别主题的延续、分裂、重组。

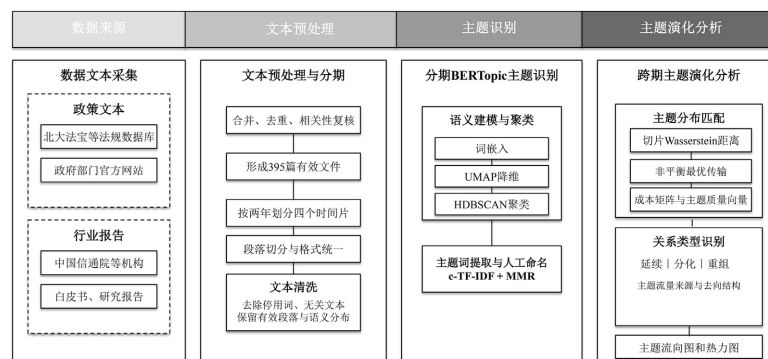


Figure 1. Research framework

图 1. 研究框架图

3.2. 数据来源与文本预处理

本文以 2018~2025 年我国人工智能安全相关政策文件和行业研究报告为研究对象。政策文件能够反映风险界定、责任配置与监管要求等制度化内容；行业报告则更接近技术研发和应用实践，可以补充正式政策所缺少的技术和场景信息。两类文本相互补充可以更加完整地观察人工智能安全议题的变化过程。政策文本主要通过北大法宝等法规数据库及政府部门官方网站检索，行业报告主要来自中国信息通信研究院等专业机构公开发布的白皮书、研究报告和年度报告。检索以“人工智能安全”为核心，并结合“数据安全”“生成式人工智能”等表述进行补充检索。对于涉及一般数字经济发展、未包含人工智能安全内容的文件予以剔除；标题相同、内容重复以及仅为转载或简短解读的文本不重复纳入。行业报告仅保留具有完整正文、能够反映风险识别或治理主张的文本。经合并、去重和相关性检查，最终获得 395 篇有效文件。最终各个阶段文件数量如表 1 所示。

Table 1. Overview of samples and topic identification for each time slice
表 1. 各时间片样本及主题识别概况

时间片	文件数	有效段落数
2018~2019	57	123
2020~2021	79	196
2022~2023	103	354
2024~2025	156	433
合计	395	1106

3.3. 分期 BERTopic 主题识别

传统主题模型根据词频共现识别主题，难以处理同一政策表述在不同语境中的语义差异，同时，在全时期文本语料上进行建模，容易将后期出现的新议题纳入既有主题，削弱阶段间的内容差别。因此，本文在每个时间片内分别采用 BERTopic 进行主题识别。该方法先以句向量表征段落语义，再经 UMAP 降维和 HDBSCAN 密度聚类形成主题，在不预设主题数量的条件下识别规模不同、边界不规则的主题。分期识别既保留了各阶段的主题特征，也为后续考察主题间的跨期关系提供了相对独立的比较单元。具体而言，设时间片 t 中的有效段落集合为 D_t ，段落文本经嵌入模型映射为语义向量 E_t ；经降维和聚类后，得到主题集合 $Z_t = \{z_{t1}, z_{t2}, \dots, z_{tK}\}$ ，其中 K_t 表示该时间片识别出的主题数。主题词由 c-TF-IDF 提取，并采用 MMR 方法进行筛选，在保证关键词与主题语义相关的同时降低词语之间的重复性，从而增强主题表示的区分度。

3.4. 切片 Wasserstein 距离与非平衡最优传输

为保留不同时间片的主题内容的分布特征，本文将每个主题表示为其内部全部段落向量构成的加权语义分布，并以切片 Wasserstein 距离计算相邻时间片主题之间的差异。Wasserstein 距离通过多个一维投影近似高维分布间的差异，既可减少直接计算高维最优传输的复杂度，也能保留主题内部段落的离散程度。如式 1 所示。

$$SWD(P_{i,t}, P_{j,t+1}) = \frac{1}{L} \sum_{\ell=1}^L W_1(R_{\theta_\ell} P_{i,t}, R_{\theta_\ell} P_{j,t+1}) \quad (1)$$

式子 1 中， $P_{i,t}$ 和 $P_{j,t+1}$ 分别表示相邻时间片中的两个主题分布； L 为随机投影次数； R_{θ_ℓ} 表示沿投影方向

θ_i 对高维分布进行的一维投影， W_i 为一维 Wasserstein 距离。对相邻时间片中的主题两两计算距离后，得到成本矩阵 C ；再以正距离的中位数进行标准化，使不同时间片之间的成本处于可比较的尺度。

经上述计算，前后主题的语义差异被转化为标准化成本矩阵。矩阵中的每一个元素对应一组主题之间的匹配代价，数值越小，表明两者的语义分布越接近。为了进一步完整呈现主题间的分裂、合并及规模增减的变化情况，本文将成本矩阵与前后主题质量共同作为输入，采用非平衡最优传输模型计算主题流量。该模型在控制总体传输成本的同时，对流量与主题质量之间的偏离施加约束，从而形成相邻时间片主题之间的整体对应关系。

$$\Gamma^* = \operatorname{argmin}_{\Gamma \geq 0} \left\{ \langle \Gamma, C \rangle + \tau D_{KL}(\Gamma \mathbf{1} \| a) + \tau D_{KL}(\Gamma^T \mathbf{1} \| b) + \varepsilon E(\Gamma) \right\}$$

(2)

其中， Γ 为待求的主题流量矩阵， Γ^* 为使目标函数最小的最优流量； τ 为质量正则参数，用于控制流量与前后主题质量偏离的惩罚程度； C 为标准化后的主题成本矩阵， a 和 b 分别表示前后两个时间片的主题质量向量， D_{KL} 项约束流量与主题质量之间的偏离， $E(\Gamma)$ 为熵正则项。质量正则参数取 0.8，用于控制主题质量变化的惩罚程度；熵正则参数取 0.05，用于保证求解过程的稳定性。模型采用非平衡 Sinkhorn 算法，通过计算，可以得到最优流量所反映主题间的对应强度。

3.5. 演化关系判定

主题关系依据 UOT 流量的来源构成和去向结构判定。若一个前期主题的显著流量主要进入一个后期主题，则识别为延续；若一个来源主题存在两条及以上显著流出路径，且这些路径覆盖其主要流量，则识别为分裂；若一个目标主题同时接收两个及以上前期主题的显著流量，则识别为重组。本文以流量占来源主题总流量的 20% 作为显著路径的基准阈值，关系判定由流量矩阵自动完成。

4. 主题识别与演化分析

4.1. 主题识别分析

通过 bertopic 对各个时间片的语料进行主题识别，得到的各个主题如表 2 所示，并对各个主题进行人工命名。

Table 2. Topic vocabulary
表 2. 主题词表

时间片	编号	主题命名	主题规模	主题词(部分)
2018~2019	T1	训练数据公平	24	机器学习、报告、平等、高风险场景、识别
	T2	信息收集规范	23	收集使用、网络、信息内容、违法
	T3	数据智能治理	18	原则、智能、数据治理、透明
	T4	计算服务应急	14	计算服务、制定、安全评估、管理责任
	T5	智能产业发展	16	人工智能、治理、发展、开放共享
	T6	数据安全控制	15	司法、训练数据、数据来源、系统
	T7	数据网络研发	13	创新、数据类型、研发支持、5g
2020~2021	T1	ai 数据治理	30	训练数据、数据治理、ai 监管、公平
	T2	数据存储合规	27	国家网信、履行、运营者处理、数据存储
	T3	个人信息保护	24	责任、处理、个人同意、非法提供

续表

	T4	算法应急防护	20	工业、应急响应、算法安全、标准体系
	T5	自动决策评估	23	自动化决策、影响评估、个人权益、处理
	T6	重要数据评估	19	风险评估、数据处理、活动、重要数据
	T7	智能治理预警	18	人工智能、治理能力、预警能力、监测预测
	T8	网络安全审查	17	网络安全、审查、影响、信息、基础设施
	T9	数据开放管理	18	公共数据、分类、开放、个人隐私
	T1	数据伦理合规	78	内容、训练数据、伦理审查、ai 治理
	T2	数据要素流通	60	数据要素、开发利用、公共数据、资产
	T3	生成内容合规	65	生成式、人工智能、内容、标注规则
2022~2023	T4	隐私计算检测	38	产品、检测评估、隐私计算、产业
	T5	算力设施建设	44	枢纽节点、算力、基础设施、训练数据
	T6	信息系统防护	32	重要、信息系统、网络安全、审查
	T7	数据权益规制	37	行政法规、同意、法律、主体
	T1	数据分级应急	98	领域、应急处置、分类分级、数据分类
	T2	模型安全防护	88	模型安全、安全风险、攻击、数据泄露
2024~2025	T3	模型数据防护	76	训练数据、模型训练、生成式 ai、来源
	T4	智能治理支撑	65	人工智能、治理、支持、网络安全
	T5	数字产业发展	61	发展、产业、数字化转型、中小企业
	T6	检索投毒防护	45	检索、rag 系统、投毒风险、诱导

2018~2019 年, 该阶段的重点在算法和数据安全等方面。数据来源、数据质量和算法公平构成早期议题的核心内容, 反映政策最先关注人工智能运行所依赖的数据条件及其可能造成的差异性影响。收集使用、信息内容和违法和等词指向数据获取与使用规则, 说明信息收集的合法依据和使用边界是这一时期的重要内容。原则、数据治理和透明解释等词项, 则体现出数据来源安全要求逐步进入人工智能治理讨论范畴。从整体上看, 政策主题仍以数据和技术安全为主, 计算服务、训练数据和 5G 等内容均与技术运行环境有关。由此可见, 该阶段尚未形成完整的人工智能安全制度体系, 政策主要针对数据来源、算法使用和网络基础等突出问题作出回应, 并为后续责任规则和审查程序的建立打下基础。

2020~2021 年, 这个阶段主题内容开始转向数据处理程序和主体责任。数据存储、个人信息处理、责任和审查等主题词出现, 反映出政策文本开始回应数据由谁处理、在何种条件下处理以及违反规则后如何承担责任等具体问题。政策关注的对象也由技术系统本身延伸到运营者、数据处理者与监管机构之间的关系, 人工智能安全在一定程度上由技术性议题转为行政管理性议题。同时, 影响评估, 风险评估和监测预测等词, 体现出政策重心开始把风险判断置于行为发生之前, 并在系统运行过程中持续监督。

2022~2023 年, 这个时期生成式人工智能的兴起, 使内容生成和伦理约束成为新的研究重点, 内容真实性、可识别性和服务责任成为这一阶段新的关注重点。训练语料的来源审查与生成结果的价值判断被纳入同一框架, 伦理审查因此不再是外部的评价活动, 而成为研发和服务流程中不可或缺的环节, 责任分配覆盖到模型研发者与服务提供者。同时, 算力设施等基础资源开始受到关注, 相关部门开始布局建设相关节点来促进相关产业进一步发展。

2024~2025 年, 该阶段主题数量有所减少, 但核心内容更加集中, 主要表现为数据分级、模型运行、训练过程、治理支撑、产业应用和检索攻击等方面。模型层面的框架安全、攻击防护、数据泄露和隐私保护成为重点, 训练数据来源、训练过程和合成内容之间的关系进一步细化。检索增强应用中的敏感信息、工具调用和数据诱导, 反映模型使用阶段出现了新的攻击方式。数据分类、场景应用、责任安排和应急处置相互结合, 表明人工智能安全治理开始进入具体场景并形成更细致的处理规范。总体来看, 该阶段政策关注点由一般性数据和算法防护进一步转向模型运行、应用攻击和综合处置, 系统性特征更加明显。

4.2. 主题演化分析

通过固定各时间片的 BERTopic 主题结果, 只改变主题匹配方式来比较不同跨期主题匹配方法对主题结构的表达能力, 如图 2 所示。

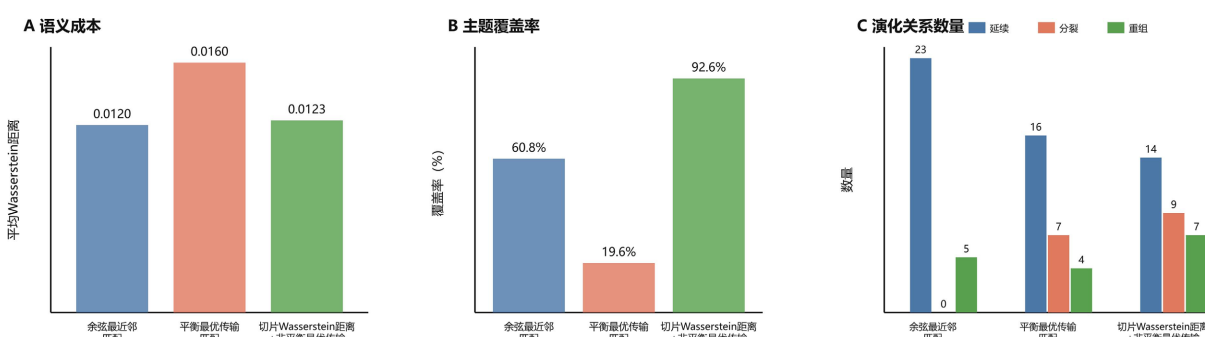


Figure 2. Comparison of cross-period relationships identified by different topic matching methods

图 2. 不同主题匹配方法的跨期关系比较

其中, 余弦最近邻匹配的平均语义成本为 0.0120, 低于切片 Wasserstein 距离与非平衡最优传输的 0.0123, 但后期主题覆盖率为 60.8%, 且未识别出分裂关系。平衡最优传输的平均语义成本为 0.0160, 后期主题覆盖率为 19.6%, 部分传输流量在显著路径阈值下未能形成清晰的主题对应。切片 Wasserstein 距离与非平衡最优传输的后期主题覆盖率达到 92.6%, 并识别出 14 个延续、9 个分裂和 7 个重组关系。

余弦最近邻匹配虽能够保持较低的局部语义成本, 但容易将主题演化简化为单一路径; 平衡最优传输可以形成多向连接, 却受到质量守恒约束的影响。切片 Wasserstein 距离与非平衡最优传输在保持语义接近性的同时, 覆盖了更多后期主题, 并保留了一对多和多对一的跨期关系, 更适合描述主题边界调整及其结构变化。因此, 本文采用 Wasserstein 距离与非平衡最优传输研究跨期主题演化。如图 3 所示, 其中, 节点高度表示主题质量, 连线宽度表示 UOT 流量。同时, 本文将一个前期主题稳定连接多个后期主题称为“分裂”, 即原有主题的不同语义成分在后期获得相对独立的表达; 将多个前期主题共同合并为一个后期主题称为“重组”, 即后期议题吸收了不同来源的主题并成为一个新的议题; 主导内容和关键词组合跨期保持稳定的关系称为“延续”。

2018~2019 年 T1 和 T3 的主要流量进入 2020~2021 年 T1。前期 T1 “机器学习、报告、平等”、T3 “原则、智能、数据治理”和 T6 “司法、训练数据、数据来源”的相关内容, 共同流向后期 T1 “训练数据、数据治理、AI 监管”。从主题内容来看, 前期分散在算法原则、数据来源和司法规范中的相关内容, 在后期逐步合并为数据治理相关主题, 说明训练数据开始成为人工智能监管的重要对象, 政策关注也由原则性要求转向具体治理活动。前期 T2 “收集使用、网络、信息内容”分别流向后期 T2 “国家网信、履行、运营者处理”和 T3 “责任、处理、个人同意”, 表明信息收集与使用主题逐步分裂为运营者责任和

个人信息保护两个方向, 数据处理主体及其责任边界开始受到重视。与此同时, 前期有关计算服务和安全评估的内容, 进一步向风险评估与网络审查主题演化, 人工智能治理和研发支持等内容也开始与产业安全及应急响应产生联系。该变化与同期人工智能技术从单一算法实验走向大规模数据训练和平台化应用的趋势相一致。BERT 等预训练模型的应用以及 GPT-2 等语言模型的发展, 使训练数据规模、数据来源和算法决策风险更加突出; 与此同时, 《数据安全法》《个人信息保护法》以及算法推荐治理规则相继出台, 政策关注由“数据能否被使用”逐渐转向“数据由谁处理、如何处理以及责任如何承担”, 反映了人工智能安全议题由技术原则向制度责任的转变。

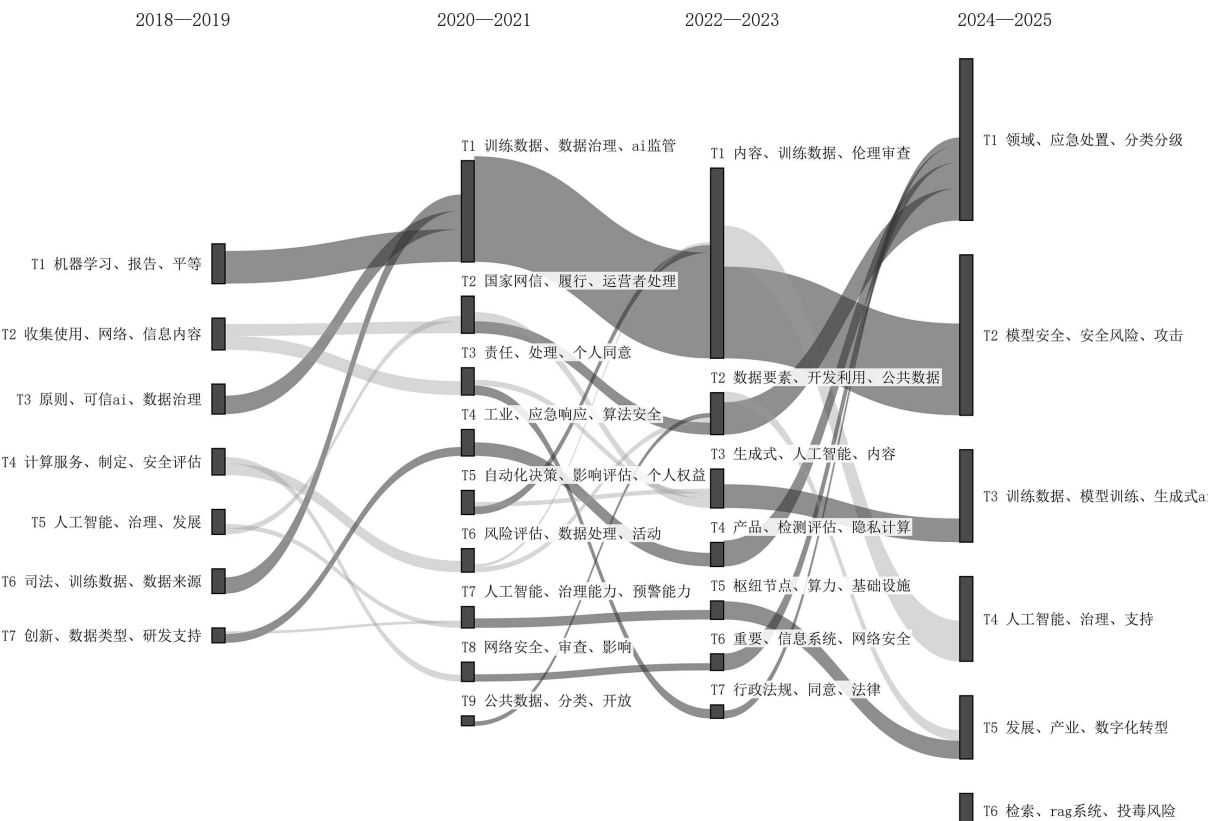


Figure 3. Cross-period topic flow diagram
图 3. 跨期主题流向图

2020~2021 年至 2022~2023 年, 人工智能安全政策主题重组关系较为明显。前期 T1 “训练数据、数据治理、AI 监管”、T3 “责任、处理、个人同意”和 T5 “自动化决策、影响评估、个人权益”的主要流量, 共同进入后期 T1 “内容、训练数据、伦理审查”和 T3 “生成式、人工智能、内容”。2022 年底 ChatGPT 的公开应用和 2023 年 GPT-4 的发布, 显著扩大了生成式人工智能的应用范围, 也使训练数据来源、生成内容真实性、模型伦理和服务提供者责任成为新的安全关注点, 但其治理要求已由数据处理延伸至内容生成过程。与此同时, 2020~2021 的 T9 “公共数据、分类、开放”与 T2 “国家网信、履行、运营者处理”的部分流量共同进入后期 T2 “数据要素、开发利用、公共数据”, 形成较为明显的重组关系。前期有关公共数据开放的内容, 与运营者责任及数据处理要求在后期相互结合, 使政策关注由数据开放进一步延伸至数据要素开发利用。这一变化表明, 数据流通与安全管理开始被纳入同一政策主题。前期有关风险评估和网络审查的内容, 则分别进入后期 T4 “产品、检测评估、隐私计算”和 T6 “重要、信息系统、网

络安全”。总体来看,这一阶段的政策关注由数据处理责任逐渐扩展至生成内容治理,公共数据开发与系统安全也得到重视,覆盖范围随之扩大。

2022~2023 年至 2024~2025 年的政策主题开始向模型安全及其应用风险集中。T3 “生成式、人工智能、内容”与后期 T3 “训练数据、模型训练、生成式 AI”保持较强的联系,两个主题都以生成式人工智能为核心,但后期关键词更加关注模型形成过程。这表明,治理重心并未停留在大模型生成内容层面,而是慢慢拓展到训练数据和模型训练环节,相关主题呈现出明显的延续特征。前期 T1 “内容、训练数据、伦理审查”的主要流量进入后期 T2 “模型安全、安全风险、攻击”,其关注重点由内容审查转向模型运行。前期主要关注生成内容是否符合伦理要求,后期则更加重视模型遭受攻击后可能产生的安全影响。模型安全由此成为连接训练阶段和运行阶段的重要议题。前期 T5 “检索节点、算力、基础设施”则几乎全部流向下一个阶段的 T5 “发展、产业、数字化转型”,说明基础设施建设在后期仍主要服务于人工智能产业发展,其议题内涵未发生显著改变。此外,后期 T6 “检索、RAG 系统、投毒风险”未与前期主题形成任何跨期流量关系,随着检索增强生成等技术进入应用后,新型人工智能安全问题也在出现,政策重心开始回应更具技术针对性的应用风险。

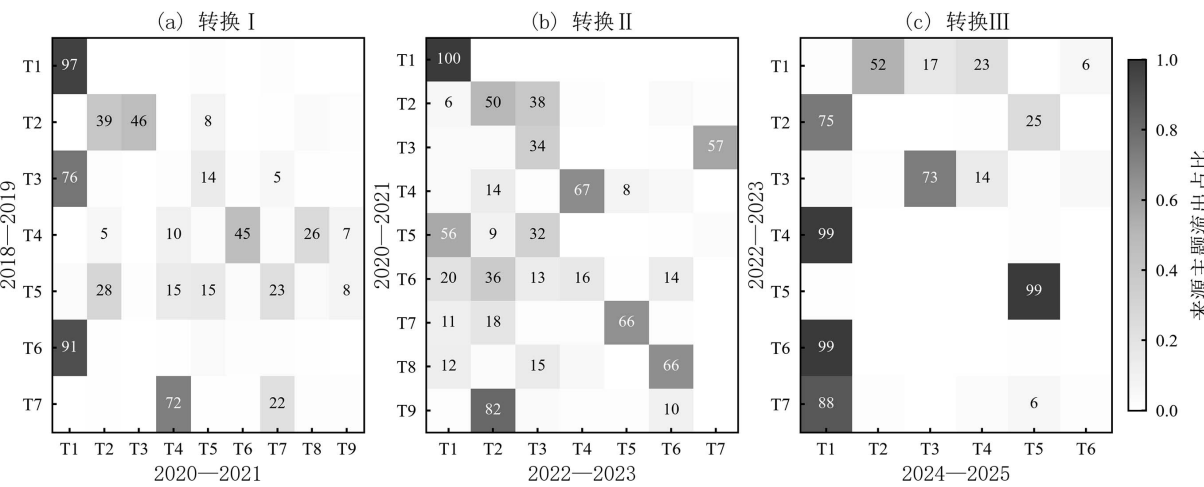


Figure 4. Heatmap of topic transitions diagram
图 4. 主题转移热力图

为进一步分析跨期主题流量的方向、集中程度及其政策含义,本文将 UOT 流量矩阵按来源主题进行归一化,结果如图 4 所示。颜色越深,表示前期主题流向后期主题的比例越高。2018~2019 年至 2020~2021 年,T2 “信息收集规范”约 39% 的流量进入 T2 “数据存储合规”,约 46% 的流量进入 T3 “个人信息保护”,表明早期围绕信息收集的原则性要求开始分化为数据处理责任和个人权益保护两个制度方向。这一变化与人工智能应用扩大后数据采集、存储和处理活动不断增加有关。随着数据安全和个人信息保护制度逐步细化,政策关注也由信息能否被收集,进一步转向由谁处理、如何存储以及个人权益如何受到保护,体现出治理对象由技术活动向责任主体和权利关系延伸。

2020~2021 年至 2022~2023 年,T3 “个人信息保护”约 57% 的流量进入 T7 “数据权益规制”,同时 T5 “自动决策评估”约 56% 的流量进入 T1 “数据伦理合规”,说明个人权益、自动化决策和影响评估等既有议题开始被纳入数据伦理和权利规制框架。其政策背景在于,人工智能系统逐渐参与信息分发、风险判断和公共服务,算法决策对个人权益和社会公平的影响更加直接。因而,政策治理不再局限于算法性能和技术风险控制,而是进一步强调数据主体权利、平台责任、算法透明和伦理审查之间的衔接,推

动人工智能治理由技术监管转向技术、责任与权利并重的综合规范。

进入 2024~2025 年后, 主题流向呈现一定集中化趋势。其中, T1 “数据伦理合规” 约 52% 的流量进入 T2 “模型安全防护”, T3 “生成内容合规” 约 73% 的流量进入 T3 “模型数据防护”, 反映出生成式人工智能快速应用后, 治理重点由生成结果审查进一步延伸至训练数据、模型开发和运行安全。其政策原因在于, 大模型风险并不只产生于输出环节, 还可能来自训练数据偏差、数据泄露、模型攻击和运行失控, 因此需要将安全要求嵌入模型研发、训练、部署和使用全过程。与此同时, T2 “数据要素流通” 约 75% 的流量进入 T1 “数据分级应急”, 表明数据开发利用开始与分类分级、风险识别和应急处置相连接, 反映出数据要素化政策需要在促进数据流通的同时强化安全边界和风险响应机制。

总体来看, 图 4 呈现出主题由前期分化、中期重组到后期集中化的演变特征。其背后的政策逻辑是, 人工智能应用范围扩大和生成式人工智能兴起不断引入新的风险环节, 推动治理政策由信息收集、个人权益等单项问题, 逐步扩展至数据伦理、内容合规、模型安全和应急处置。由此, 人工智能安全治理开始形成覆盖数据处理、模型开发、应用运行和风险处置的连续性治理框架。

5. 结论

生成式人工智能进入公共服务场景后, 训练数据、模型运行与应用交互中的风险相互叠加, 人工智能安全治理由局部技术防护转向全流程治理已成为重要议题。本文以 2018~2025 年人工智能政策文本和行业报告为样本, 运用分期 BERTopic 识别阶段主题, 并结合切片 Wasserstein 距离与非平衡最优传输分析主题关系。研究结论如下。

(1) 我国人工智能安全议题经历了由数据与算法风险、数据处理与责任配置、生成内容与伦理约束, 再到模型运行与综合处置的阶段演进。早期政策主要围绕训练数据、算法公平、信息收集和数据使用规则展开; 随后, 个人信息处理、风险评估与主体责任逐步进入政策视野, 技术问题开始被置于行政程序之中。生成式人工智能兴起后, 训练材料审查、内容标识、服务提供者责任及模型攻击防护等议题相继出现。至 2024~2025 年, 数据泄露、检索增强攻击、基础设施安全和应急处置进一步汇入模型运行治理。这表明, 人工智能安全的治理对象并非简单增加, 而是沿技术链条和应用链条持续外延, 责任主体与程序工具也随之调整。

(2) 人工智能安全主题在跨期演进中呈现延续、分化与重组并存的关系。训练数据、信息处理等基础议题在相邻时期保持稳定联系, 为后续治理调整提供问题基础。随着技术条件和应用场景变化, 原有议题又分化为数据治理、决策程序、内容生产和模型运行等方向。来自不同领域的主题, 则围绕模型安全、数据防护与应急处置重新组合。主题演进因而不是由分散走向整合的单向过程。分化说明不同风险环节需要形成相应的治理规则, 重组则说明模型安全、数据防护和应急处置已无法由单一环节独立完成。系统性治理的关键, 在于使训练、部署、运行和处置之间形成可衔接的规则与责任链条。

(3) 切片 Wasserstein 距离与非平衡最优传输能够提高跨期主题关系的识别能力。与仅依据关键词重合或主题中心相似度的匹配方法相比, 该方法以主题内部文本的语义分布计算成本, 并允许主题质量跨期增减, 因而能够同时识别一对多的分化、多对一的重组及稳定延续关系。结果显示, 本文方法在保持较低语义成本的同时提高了后期主题覆盖率。这为观察政策议题如何形成、拆分和重新组合提供了可操作的文本计算路径。

本文仍有进一步拓展的空间。一方面, 政策文件与行业报告在文本篇幅和表达方式上存在差异, 统一语料处理虽然有利于观察两类文本的共同议题, 但不同文本类型在主题规模和语言表达上存在较大的差别, 后续研究可依据文本类型、发布主体和材料篇幅设置分层权重, 以提高不同来源主题比较的区分度。另一方面, 本文依据语义分布和跨期流量识别议题的延续、分化与重组, 能够呈现主题结构如何变

化, 但对政策出台、技术突破或风险事件何以触发主题转向的具体机制, 仍有待结合事件资料、政策过程材料和典型案例进一步阐释。

参考文献

- [1] Bommasani, R., Hudson, D.A., Adeli, E., *et al.* (2021) On the Opportunities and Risks of Foundation Models. <https://arxiv.org/abs/2108.07258>
- [2] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J., *et al.* (2022) Taxonomy of Risks Posed by Language Models. 2022 *ACM Conference on Fairness, Accountability and Transparency*, Seoul, 21-24 June 2022, 214-229. <https://doi.org/10.1145/3531146.3533088>
- [3] Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 3-10 March 2021, 610-623. <https://doi.org/10.1145/3442188.3445922>
- [4] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., *et al.* (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, 27-30 January 2020, 33-44. <https://doi.org/10.1145/3351095.3372873>
- [5] 魏弋, 李刚. 人工智能治理范式变革——基于政策分析与主题建模的研究[J]. 图书与情报, 2024, 44(5): 62-76.
- [6] 李亮, 徐晓林, 明承瀚. 政务数据安全政策体系的优化路径研究——基于治理政策文本的扎根理论分析[J]. 图书馆, 2024(12): 36-45.
- [7] 张涛, 张博雅, 马海群. 我国央地政府数据安全政策扩散特征及主题转移研究[J]. 情报理论与实践, 2025, 48(5): 118-127, 137.
- [8] 周文泓, 熊小芳, 叶雅寒. 面向人工智能的数据治理框架研究——基于政策文本的构建与展望[J]. 信息资源管理学报, 2025, 15(4): 87-98.
- [9] Blei, D.M. and Lafferty, J.D. (2006). Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning*. 25-29 June 2006, 113-120. <https://doi.org/10.1145/1143844.1143859>