

中文个人信息（PII） 检测技术分析报告

Technical Analysis Report on Chinese PII Detection

发布方: argus-redact 项目组 日期: 2026.03 | 版本: v1.0 | 许可: CC BY 4.0

基准数据集 · 评估框架 · 参考实现

摘要

2024 年中国以泄露账户数跃居全球第一，单次事件涉及 87 亿条记录。《个人信息保护法》（PIPL）执法力度持续加码，滴滴 80 亿元罚单仅是开始。与此同时，大语言模型（LLM）的广泛部署创造了新的 PII 泄露通道——最新研究表明，非拉丁文字（含中文）的隐私保护机制效果最差。

然而，中文 PII 检测在学术界和工业界均处于被忽视的状态：开源社区无专门工具，商业方案云锁定，NER 模型精度远低于英文。本报告系统分析中文 PII 的 8 种核心类型的技术特征与检测难点，调研监管、威胁和工具现状，并基于我们创建的据我们调研为据我们调研为首个开源中文 PII 基准数据集（pii-bench-zh, 8,000 样本）给出定量评估。

核心发现：

- 结构化中文 PII（电话、身份证、银行卡、车牌、护照、邮箱）通过格式匹配 + 校验和可达 100% F1
- 中文人名检测是核心瓶颈：spaCy 中文 NER F-score 仅 71%，纯 regex 依赖上下文启发式
- 非正式场景（聊天/IM）的 PII 检测 F1 比正式场景低 15-25 个百分点
- 开源领域缺乏开箱即用的中文 PII 检测工具

1. 中国个人信息保护的监管与威胁态势

1.1 法规体系

中国已形成以《个人信息保护法》（PIPL, 2021）为核心，《数据安全法》（DSL）和《网络安全法》（CSL）为两翼的三法并立格局。

时间	法规/标准	核心要求
2021.11	PIPL 施行	罚款上限 5000 万元或年营收 5%
2025.01	网络数据安全条例 (NDSMR)	24 小时事件报告，5000万+用户平台加重责任
2025.11	GB/T 45574-2025	重新定义敏感个人信息边界
2026.01	个人信息出境认证办法	跨境传输必须安全评估/标准合同/认证三选一

执法力度：2024 年网信办约谈平台 11,159 家、罚款 4,046 家、下架 App 200 款、关闭账号 107,802 个。检察机关 2023 年办理个人信息保护公益诉讼 6,300+ 件。

1.2 重大泄露事件

时间	事件	规模	数据类型
2022.07	上海公安数据库泄露	~10 亿条, 23TB	姓名/地址/身份证/电话/犯罪记录
2024	多源数据汇编 (COMB)	12 亿条	QQ/微博/顺丰等多源合并
2025.05	Elasticsearch 暴露	87.3 亿条	身份证/手机/微信/支付宝/财务数据

据 Surfshark (2025) 统计，2024 年中国泄露账户数从全球第 12 位跃升至第 1 位，泄露速率大幅上升。

1.3 市场背景

中国网络安全市场 2025 年估值 157 亿美元，预计 2030 年达 237 亿美元 (CAGR 8.6%)。NDSMR 施行首月即收到 2,000+ 份跨境数据备案——监管驱动的合规需求正在加速释放。

2. 中文 PII 类型学：8 种核心类型的技术解剖

2.1 手机号码

格式: 1[3-9]XXXXXXXX (11 位)

变体: 138 1234 5678 | 138-1234-5678 | +8613812345678

校验: 无校验和

标准: 工信部《电信网编号计划》(2017)

中国手机号以 1[3-9] 开头，第二位 3-9 代表不同运营商段。格式固定，检测难度低。

实际挑战：聊天场景中空格和横杠分隔极为常见。我们在 3,000 条合成聊天消息的基准测试中模拟了这一现象，不支持分隔符的检测器在该场景中 recall 从 100% 降至 83.5%。

2.2 居民身份证号码

格式：AAAAAA YYYYMMDD SSSV (18 位)
结构：区划码(6) + 出生日期(8) + 顺序码(3) + 校验码(1)
校验：ISO 7064:1983 MOD 11-2
标准：GB 11643-1999《公民身份号码》

身份证号是中国最重要的个人标识符。其结构设计使得 regex + 校验和可以实现极高的检测精度：

- **区划码**：首位非 0，可排除约 10% 的随机数字串
- **日期段**：月份 01-12，日期 01-31，可排除无效日期
- **MOD 11-2 校验**：最后一位为校验码（0-9 或 X），可排除绝大多数随机 18 位数字串

实际挑战：

- 聊天中校验码 `x` 常以小写 `x` 出现
- 空格分隔（`110101 19900307 4610`）常见于手动录入
- 部分系统截取前 6 位或后 4 位传输，需处理片段匹配

2.3 银行卡号

格式：BBBBBBXXXXXXXXXX (16-19 位)
结构：BIN(6) + 账号(6-9) + 校验位(1)
校验：Luhn 算法
标准：ISO/IEC 7812, 中国银联 BIN 分配表

关键发现：在我们的合成基准测试中，仅依赖 Luhn 校验 + 上下文过滤的检测策略 recall 不足 10%，主要原因是上下文误判将大量合法卡号排除。补充已知中国银行 BIN 前缀（建行 621700、工行 622202、农行 622848 等 52 个）作为白名单后，recall 提升至 100%。

这揭示了一个普遍被忽视的问题：**国际标准在中国本地化实践中可能不够充分**。仅依赖 Luhn 等国际通用校验而不结合本地发卡行 BIN 信息的检测策略，在实际部署中可能面临 recall 不足的风险。

2.4 车牌号码

格式：省A·XXXXX（普通） / 省AXXXXXX（新能源）
结构：省简称(1) + 发牌机关(1) + [分隔符] + 编码(5-6)
标准：GA 36-2018《中华人民共和国机动车号牌》

省份简称是单个汉字，需枚举全部 31 个省级行政区（京津沪渝 + 22 省 + 5 自治区）。新能源车牌为 6 位编码（含一个字母），与普通 5 位不同。

2.5 地址

格式：[省] + [市] + 区/县 + 街道 + [门牌]
结构：层级嵌套，每层可选
标准：GB/T 2260《中华人民共和国行政区划代码》

地址是中文 PII 中检测难度最高的类型，原因有四：

- [1] **无标准格式**：同一物理位置可以有十余种文本表达
- [2] **完整**：北京市朝阳区建国路100号
- [3] **省略市**：朝阳区建国路100号
- [4] **省略区**：建国路100号
- [5] **非正式**：朝阳建国路100号（无"区"后缀）
- [6] **与普通文本高度重叠**：中山路 同时是全国最常见的街道名和普通词组
- [7] **层级组合爆炸**：全国约 2,843 个区县 × 数万条街道，穷举不现实
- [8] **非正式地址在聊天中占主导**：快递/外卖场景中"南山深南大道100号"这种省略全部上级行政区的写法极为普遍

可行方案：双轨策略——正式地址用省→市→区→街道层级匹配，非正式地址枚举主要城区名（75 个覆盖一线及新一线城市）直接匹配街道。这种方法在正式场景 recall 88.5%，非正式场景 recall 88.4%。

2.6 护照号码

格式：E/G + 8 位数字（9 位总长）
标准：中华人民共和国护照法

格式简单但容易与 Git commit hash、产品编号等冲突。通过限定首字母仅 E（因私普通）/G（因公普通）、排除更长字母数字串的负向断言来降低误报。

2.7 邮箱

RFC 5322 标准格式，跨语言通用。中文语境的特殊性在于本地化域名分布高度集中于 qq.com、163.com、126.com 等国内服务商。

2.8 人名

格式：姓(1-2字) + 名(1-2字)，总长 2-4 个汉字
姓氏池：约 500 个常见姓覆盖 99%+ 人口（本报告实现中使用 138 个高频姓 + 16 个复姓）
来源：公安部全国姓名统计

人名是中文 PII 检测的核心难题，也是与英文 PII 差异最大的类型。

为什么中文人名比英文难？

维度	英文	中文
词边界	空格天然分隔 "John Smith"	无分隔 "张三在北京"
大小写信号	首字母大写 = 可能是专有名词	无大小写概念
命名模式	Given + Family，模式固定	姓+名，但"姓"可能是普通字
歧义	较少 ("Apple" 大写=公司)	严重 ("黄山"=人/山，"小米"=人/品牌/谷物)

NER 的局限

模型	中文 NER F-score	英文 NER F-score	差距
spaCy (web_lg)	71.3%	86.4%	-15.1%
HanLP MSRA	~93%	—	学术基准高，但模型 500MB

spaCy 中文 NER 的 F-score（precision 72%、recall 68%）意味着较高的漏检和误报率，在对 recall 要求严格的 PII 保护场景中**存在明显不足**。

姓氏前缀启发式：一种不依赖 NER 的替代方案

观察到中文文本中人名几乎总伴随上下文线索出现：

- **前缀**：客户 张三、患者 李芳、联系人 王小明
- **后缀**：赵敏 女士、陈华 老师、刘伟 先生

枚举 40+ 前缀词和 15 个后缀敬称，结合 138 个常见姓氏 + 16 个复姓，可在合成基准数据上实现 **零误报** 的前提下捕获约 49% 的人名——且无需加载任何 NER 模型，延迟 <1ms。

剩余 51% 缺乏上下文线索的人名（如聊天中的"你问一下张三"）仍需 NER 模型处理。

3. LLM 时代的 PII 新风险

3.1 三条泄露路径

大语言模型为 PII 创造了三条新的泄露通道：

- [1] **提示词直传**：用户在 prompt 中直接包含 PII（“帮我分析张三的体检报告，身份证号 110101...”）。这是最常见的路径——用户不知道或不在意 prompt 会被发送到云端。
- [2] **训练数据记忆提取**：LLM 在预训练阶段记忆了包含 PII 的数据。PII-Scope（2024）系统性地证明了这一点，并发现超参数设置对攻击效果影响巨大。
- [3] **跨语言泄露**：即使训练数据仅含英文，用中文提问也可能触发英文 PII 的输出。2025 年的研究明确指出：**非拉丁文字（含中文）的隐私保护机制效果最差**（Cross-Lingual Privacy Leakage, arXiv:2506.00759）。

3.2 传统 PII 工具的根本矛盾

在 LLM pipeline 中，传统 PII 工具面临一个逻辑悖论：

永久删除方案：
张三在协和医院做了体检 → [PERSON] 在 [LOCATION] 做了体检
→ LLM 处理（无法区分不同人）
→ 输出中 [PERSON] 无法还原 → 信息永久丢失

固定假名方案：
张三 → PERSON_1（每次固定）
→ ETH Zurich 研究表明：LLM 代理可以 \$1-4/人的成本对在线用户进行去匿名化（arXiv:2310.07298）
→ 固定假名使跨请求关联成为可能

核心矛盾：删除 PII 让 LLM 丧失语义上下文，保留假名又面临关联攻击。一种有效方案是**可逆加密 + 每次随机密钥**——同一个"张三"在不同请求中被映射为不相关的假名，攻击者无法跨请求追踪。其他方向如差分隐私、联邦学习也在探索中，但在非结构化文本的实体粒度操作上尚不成熟。

3.3 跨境数据合规影响

2026 年 1 月起施行的个人信息出境认证办法要求跨境传输走三选一机制（安全评估/标准合同/认证）。这对使用海外 LLM API（如 GPT-4、Claude）的中国企业形成直接约束——**调用 API 前对 PII 进行本地脱敏成为合规刚需**。

4. 工具现状与空白分析

4.1 开源工具对中文 PII 的支持

工具	中文实体识别器	中文 PII 类型数	可逆	本地运行
Presidio (Microsoft)	无	0	否	是
Phileas	无	0	否	是
DataFog	无	0	否	是
anonLLM	无	0	是	否(OpenAI)
Google Cloud DLP	部分	~3	否	否

结论：开源社区缺乏开箱即用的中文 PII 检测方案。

Presidio 作为最成熟的开源 PII 工具，其官方内置实体列表覆盖美国（SSN、驾照）、英国（NHS、NINO）、新加坡、澳大利亚、印度等——**但未内置任何中国实体**（无身份证号、无中国手机号、无银行卡号）。其架构支持自定义 Recognizer 扩展，但需要用户自行实现全部中文 PII 模式，工程成本不低。

4.2 商业云服务

阿里云（数据安全中心 SDDP）、华为云（DSC）、腾讯云（数据安全中心）均提供中文 PII 检测能力。但存在三个结构性限制：

- [1] **云锁定**：数据必须上传到对应云平台，无法本地部署
- [2] **不可逆**：仅提供脱敏和删除，不支持可逆加密
- [3] **不可嵌入 LLM pipeline**：无 Python SDK 级集成，无法作为 LangChain/LlamaIndex 的处理节点

4.3 NER 模型的精度瓶颈

中文 NER 的学术基准（OntoNotes、MSRA）与实际 PII 场景存在显著差距：

模型	学术 F-score	PII 场景适用性	局限
spaCy zh_core_web_lg	71.3%	低	人名漏检率 ~30%
HanLP MSRA NER	~93%	中	模型 500MB，加载 2-5s
BERT-CRF (fine-tuned)	~95%	高	需标注数据和训练资源

学术 NER 优化的是 PER/LOC/ORG 三类实体的边界精度，而 PII 场景需要同时覆盖结构化类型（电话、ID）和非结构化类型（人名、地址）。单靠 NER 无法解决中文 PII 问题。

5. pii-bench-zh：据我们调研为据我们调研为首个开源中文 PII 基准数据集

5.1 为什么需要新的基准？

现有 PII 评估数据集的语言分布严重偏向英文和欧洲语言：

数据集	样本量	语言	中文 PII 类型
ai4privacy/pii-masking-400k	400K	en/de/fr/es/it/nl	0
nvidia/Nemotron-PII	100K	en	0
gretelai/synthetic_pii_finance	56K	en/de/fr/es/it/nl/sv	0
Kaggle PIILO	7K	en	0

不存在覆盖中文电话号、身份证号、银行卡号、车牌号、地址的公开标注数据集。

5.2 数据集设计

我们创建了 pii-bench-zh（Apache 2.0，HuggingFace 开放下载），包含两个子集：

子集	样本量	场景	噪音特征
formal	5,000	注册表单、病历、快递单、银行开户	标准格式
chat	3,000	微信/IM 消息、语音转文字、群聊	emoji、空格分隔、口语化、中英混排

实体分布（两个子集合计）：

类型	数量	占比
person	8,320	35.8%
phone	6,661	28.7%
id_number	2,334	10.0%
address	1,684	7.2%
email	1,424	6.1%
bank_card	1,366	5.9%
passport	723	3.1%
license_plate	694	3.0%
合计	23,206	

5.3 生成方法与质量保证

- **100% 合成数据**——模板 + 假数据生成器，不含任何真实个人信息
- **校验和有效性**：身份证号通过 MOD 11-2，银行卡号通过 Luhn 或 BIN 前缀校验
- **字符级偏移标注**：每个实体标注 start/end 偏移量，全部通过 `text[start:end] == entity.text` 验证
- **确定性复现**：seed=42，任何人可从源码完全复现

6. 定量评估

6.1 方法与局限性声明

我们使用统一评估框架对多种检测策略进行对比测试。评估指标：Precision（精确率）、Recall（召回率）、F1 分数。匹配策略为 value-level（实体文本 + 类型完全匹配）。

重要说明：pii-bench-zh 的正式场景子集（formal）使用模板 + 假数据生成器产出，其 PII 格式与检测器的 regex 模式存在对齐关系——因此结构化 PII 的高 F1 分数部分反映了模式覆盖的完备性，而非对真实世界多样性的全面应对。聊天子集（chat）通过噪音注入（空格、emoji、非标准分隔）在一定程度上缓解了这一问题，但仍属合成数据。我们同时在 ai4privacy、Kaggle PIILO 等外部真实数据集上进行了对比评估（见 6.4），以提供更客观的参照。

6.2 正式场景结果 (pii-bench-zh formal, 1,000 samples)

实体类型	Precision	Recall	F1
email	100.0%	100.0%	100.0%
id_number	100.0%	100.0%	100.0%
license_plate	100.0%	100.0%	100.0%
passport	100.0%	100.0%	100.0%
phone	98.1%	100.0%	99.1%
bank_card	100.0%	100.0%	100.0%
address	91.7%	88.5%	90.0%
person	100.0%	48.9%	65.7%
Overall	98.7%	81.7%	89.4%

发现 1：结构化 PII（电话、身份证、银行卡、护照、车牌、邮箱）通过 regex + 校验和可达近乎完美的检测精度。

发现 2：人名检测是唯一的短板（recall 48.9%），且与上下文线索可用性相关。

发现 3：地址检测 F1=90%，双轨策略（正式 + 非正式）有效。

6.3 聊天噪音场景 (pii-bench-zh chat, 500 samples)

实体类型	正式场景 F1	聊天场景 F1	退化
email	100.0%	97.5%	-2.5%
id_number	100.0%	100.0%	0%
license_plate	100.0%	100.0%	0%
passport	100.0%	100.0%	0%
phone	99.1%	97.8%	-1.3%
bank_card	100.0%	100.0%	0%
address	90.0%	93.8%	+3.8%
person	65.7%	3.5%	-62.2%
Overall	89.4%	74.2%	-15.2%

- 发现 4：** 噪音容忍处理（空格/横杠分隔、小写 x）使结构化 PII 在聊天场景中仅退化 0-2.5%。
- 发现 5：** 人名检测在聊天场景中严重退化（F1 从 65.7% 降至 3.5%），因为 IM 消息缺乏"客户XX"式上下文标记。这是 NER 模型的核心价值场景。
- 发现 6：** 非正式地址在聊天场景中反而表现更好（+3.8%），因为聊天中的地址更倾向于使用"朝阳建国路100号"这种我们的非正式模式覆盖的格式。

6.4 工具对比（英文数据集）

工具	数据集	Precision	Recall	F1
Regex + 校验和	ai4privacy (en, 500)	78.3%	30.3%	43.7%
Regex + NER	ai4privacy (en, 200)	77.4%	46.1%	57.7%
Presidio	kaggle_piilo (en, 500)	35.1%	47.1%	40.2%
Regex + NER	kaggle_piilo (en, 200)	20.8%	40.5%	27.5%

在英文人名主导的数据集（Kaggle PIILO, 85%+ 是人名）上，Presidio 凭借更成熟的英文 NER 领先。在结构化 PII 上（email: 95%+ F1），regex + 校验和方案不分语言地保持优势。

7. 技术建议

基于以上分析，我们对中文 PII 检测系统提出以下技术建议：

7.1 分层检测是必要的

单一技术无法覆盖所有中文 PII 类型：

层	技术	适用类型	延迟
Layer 1	Regex + 校验和	电话、身份证、银行卡、车牌、护照、邮箱	<1ms
Layer 2	NER 模型	人名、地名、机构名	10-100ms
Layer 3	Local LLM	隐含 PII、昵称、间接指代	~1s

Layer 1 应作为所有场景的基线——零依赖、sub-ms、确定性，可部署到任何设备（包括边缘设备和嵌入式系统）。Layer 2/3 按需叠加。

7.2 本地化校验优先于国际标准

中文 PII 的本地化特征要求检测器不能简单套用国际标准：

- 银行卡号需 BIN 前缀兜底，不能只依赖 Luhn
- 身份证号需区划码 + 日期 + MOD 11-2 三重校验
- 地址需枚举本地行政区划，不能用英文地址模式

7.3 噪音容忍是必须能力

聊天/IM 场景在中国移动互联网中占据主导地位。PII 检测系统必须处理：

- 空格/横杠分隔的数字（手机号、身份证号）
- 大小写不敏感（身份证校验码 X/x）
- emoji 穿插
- 中英文混排

7.4 可逆加密适合 LLM pipeline 场景

在需要 LLM 处理 PII 相关文本且后续需还原真实身份的场景中：

- 永久删除破坏语义，LLM 无法有效工作
- 固定假名存在关联攻击风险

-
- 可逆加密 + per-message 随机密钥可同时满足：LLM 可处理 + 输出可还原 + 跨请求不可关联

需要指出，如果业务场景不需要还原（如合规审计、数据发布），永久删除或泛化仍是更简单的方案。可逆加密的价值体现在需要"脱敏→LLM 处理→还原"完整闭环的场景中。

7.5 需要中文 PII 基准数据集

学术界和工业界需要公开、标准化的中文 PII 基准来推动技术进步。我们创建的 pii-bench-zh 是一个起点，但仍需：

- 更多真实场景的标注数据（在合规前提下）
- 更多方言和地区变体覆盖
- 更精细的实体子类型（如手机号区分移动/联通/电信）

8. 结论

中文 PII 保护面临三重困境：**监管趋严、泄露频发、工具缺位。**

本报告的核心发现：

- [1] **结构化中文 PII 可通过 regex + 校验和达到 100% F1**——这是低悬的果实，任何系统都应优先实现
- [2] **中文人名是核心瓶颈**——spaCy 中文 NER 仅 71% F-score，姓氏前缀启发式可在零误报下覆盖 49%
- [3] **聊天场景比正式场景 F1 低 15 个百分点**——噪音容忍处理可将结构化 PII 的退化控制在 2.5% 以内
- [4] **开源中文 PII 工具缺乏开箱即用方案**——这既是挑战也是机会
- [5] **LLM 时代需要可逆加密**——永久删除和固定假名在 LLM pipeline 中均有根本性缺陷

我们开源了首个中文 PII 基准数据集 [pii-bench-zh](#)（Apache 2.0, 8,000 样本, 23,206 标注实体），以及参考实现 [argus-redact](#)（MIT, 3,800 行 Python, 7 语言, 766 测试通过），希望推动中文 PII 检测技术的发展。

参考文献

- [1] 全国人民代表大会.《中华人民共和国个人信息保护法》. 2021.
- [2] 国家标准化管理委员会. GB 11643-1999《公民身份号码》.
- [3] 国家标准化管理委员会. GB/T 35273-2020《信息安全技术 个人信息安全规范》.
- [4] 国家标准化管理委员会. GB/T 45574-2025《数据安全技术 敏感个人信息处理安全要求》.
- [5] Kim et al. "PII-Scope: A Benchmark for Training Data PII Leakage Assessment in LLMs." arXiv:2410.06704, 2024.

-
- [6] Li et al. "Cross-Lingual Privacy Leakage in Large Language Models." arXiv:2506.00759, 2025.
 - [7] Staab et al. "Beyond Memorization: Violating Privacy via Inference with Large Language Models." arXiv:2310.07298, 2023.
 - [8] Cybernews. "Billions of Chinese records exposed in massive data leak." 2025.
 - [9] Surfshark. "Data Breach Recap 2024." 2025.
 - [10] 中国网络安全产业联盟. "2024年中国网络安全产业分析报告."
 - [11] DLA Piper. "China Recent Enforcement Trends in Data Protection." 2025.
 - [12] spaCy. "Chinese Models Performance." <https://spacy.io/models/zh>
 - [13] Microsoft. "Presidio: Supported Entities." https://microsoft.github.io/presidio/supported_entities/
 - [14] Ji et al. "Chinese named entity recognition: The state of the art." Neurocomputing, 2022.
-

引用

```
@techreport{chinese_pii_whitepaper_2026,  
  title={中文个人信息（PII）检测技术分析报告},  
  author={wan9yu},  
  year={2026},  
  url={https://github.com/wan9yu/argus-redact/blob/main/docs/whitepaper-chinese-pii.md}  
}  
  
@dataset{pii_bench_zh_2026,  
  title={PII Bench ZH: Chinese PII Detection Benchmark},  
  author={wan9yu},  
  year={2026},  
  url={https://huggingface.co/datasets/wan9yu/pii-bench-zh},  
  license={Apache-2.0}  
}
```