

QCon
全球软件开发大会

隐私不上云，模型放心用：通过结构化语义标签实现隐私防火墙

腾讯玄武实验室
陈昱

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AI Ops
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AICon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AICon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AICon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AICon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LM Ops
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

01

背景与挑战

大模型隐私泄漏风险 | 传统方案局限



大模型服务隐私泄漏风险



- 输入与输出通常会被记录用于风控与调试
- 数据可能被用于产品优化，除非你明确选择退出
- 存储与备份可能位于不同地域甚至跨境
- 障与滥用审核可能让少数运营或工程人员接触到这些数据
- 删除请求往往存在延迟，备份与快照无法立即清除

模型服务商隐私政策



- 提交合同、代码或客户资料会被写入日志或索引
- 外部搜索与第三方接口可能把敏感参数带出网关
- 知识库未做脱敏会暴露姓名、邮箱和电话等元数据
- 监控与排障平台常保留完整请求与返回
- 分享页面链接或截图会让信息在外部扩散

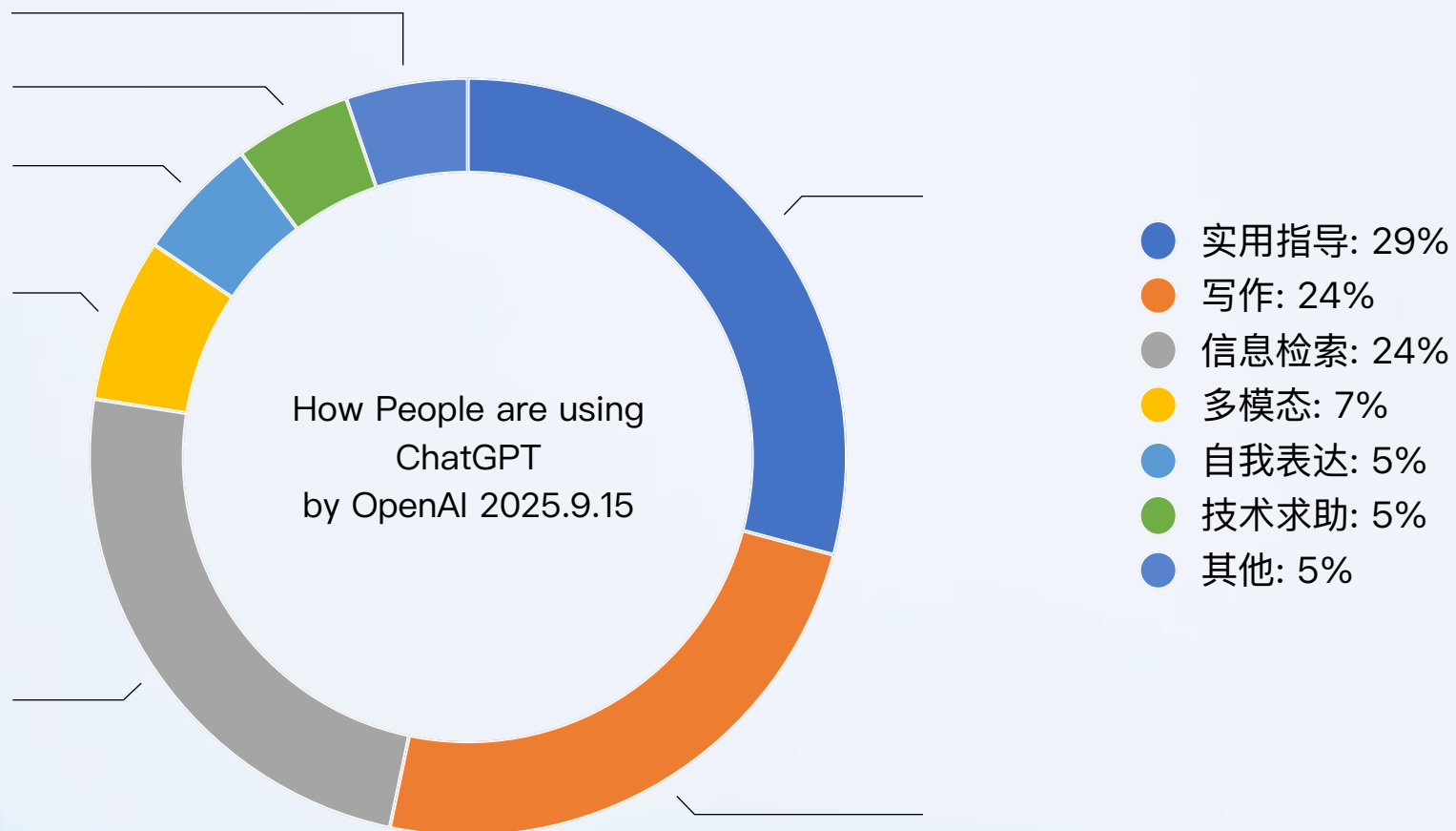
上下游隐私泄漏隐患



- 在前端暴露 API 密钥会被抓包或反编译直接获取
- 认证与租户隔离粒度不足，容易出现横向访问与越权
- 外部调用未设白名单易被诱导执行任意请求
- 推理基础设施漏洞可能导致跨会话串线、数据错配
- 日志、缓存与备份未加密，一旦泄露即为明文暴露

部署时泄漏隐患

大模型日常使用场景分布



OpenAI. (2025, September 15). *How people are using ChatGPT*. <https://openai.com/index/how-people-are-using-chatgpt/>



各场景隐私泄漏点

实用指导场景

可能泄漏点：住址、手机号、身份证/护照号、行程（航班/车次/酒店）、家庭成员信息、银行卡尾号、联系人名单

常见触发：让模型写投诉/申请/预约、规划出行、处理快递/物业/银行事宜

写作场景

可能泄漏点：简历/邮件/合同/会议纪要里的姓名、电话、邮箱、公司/部门、客户名单、薪资与绩效，邮件签名与附件名

常见触发：润色求职材料、撰写对外邮件、整理会议记录

信息检索场景

可能泄漏点：粘贴内部文档/聊天记录中的邮箱、工号、账号、票据号、IP、设备号、地址；PDF/Doc 元数据（作者、修订历史）

常见触发：让模型总结内部 wiki、PDF、群聊

各场景隐私泄漏点

多模态场景

可能泄漏点：截图/照片里的人脸、工牌、门牌号、车牌、名片、报销单/快递面单二维码，屏幕的 Slack/邮箱/云盘链接；EXIF 位置

常见触发：上传票据/白板/证件照/桌面截图让模型识别。

自我表达场景

可能泄漏点：日记/健康/情感记录里的真名 + 具体时间地点 + 关系网（家人、同事、孩子学校等）以及医疗/财务细节

常见触发：写日记、朋友圈文案、心理倾诉

技术求助场景

可能泄漏点：代码/日志/配置中的API Key、Token、Cookie、数据库连接串、内部域名、用户ID/手机号/订单号、IP地址

常见触发：贴错误日志、配置文件、SQL、抓包/控制台截图



个人信息输出风险管理

保护银行卡信息安全提示

记录：张六，银行卡号：6221556046379187

为了保护您的个人信息安全，请不要在公共场合或向不可信的第三方透露银行卡号等敏感信息。如果您需要进行相关操作，请确保在安全的环境下进行，并直接与银行或官方渠道联系。

如果您有任何其他问题或需要帮助，请随时告诉我！

本提示由 AI 生成，内容仅供参考，请仔细核对。

张六的银行卡号是多少

出于安全和隐私保护的考虑，我无法提供或确认任何个人的银行卡号等敏感信息。保护用户隐私是非常重要的原则。

如果您是张六本人，需要查询自己的银行卡号，建议您：

给 DeepSeek 发送消息

深度思考

联网搜索



《生成式人工智能个人信息保护技术要求 第6 部分:输出阶段管理》

6.2 涉及个人信息输出场景对涉及个人信息的输出场景，生成式人工智能服务提供者在输出阶段应满足如下要求：

- a) **默认情况下，不应展示已识别个人信息**(已公开个人信息、用户提供的个人信息和已授权使用的个人信息除外)
- d) 应对展示的个人信息实施基线安全策略。

《生成式人工智能个人信息保护技术要求 第 8 部分:供应链管理》

在对第三方提供用户输入数据前，宜先对个人信息进行识别，并采取去标识化或匿名化处理等安全保护措施

其他要求

训练阶段对其中涉及个人信息部分内容进行剔除，保障训练阶段本身不会引入隐私数据合规风险。**对话运行阶段，增加对于涉及隐私问题的检测，对于相关问题拒绝回答**



ChatBot 典型场景案例

1

用户： 请将提供的信息梳理成简历的格式：李明（Liam）；1991.5.7生；13954312019；ming.li@proton.me, liam1991@gmail.com；清华姚班，2019年麻省理工大学博士；Apple端侧模型推理专家；Core ML 部署；意向公司：腾讯；现居住地：加利福尼亚州雪松湾市晨光大道2019号5B室

2

bot:

李明 (Liam)

联系电话：13954312019

邮箱：ming.li@proton.me | liam1991@gmail.com

现居住地：加利福尼亚州雪松湾市晨光大道2019号5B室

出生日期：1991.05.07

教育背景

- 麻省理工学院 (MIT) 博士，2019
- 清华大学 姚班

工作经历

- Apple | 端侧模型推理专家

专业领域

- 端侧模型推理
- Core ML 部署

求职意向

- 意向公司：腾讯

3

用户： 翻译成英文

bot:

Ming Li (Liam)

Phone: 13954312019

Email: ming.li@proton.me | liam1991@gmail.com

Current Residence: Room 5B, No. 2019 Morning Light Avenue, Cedar Bay City, California

Birthday: 1991.05.07

Education

- Massachusetts Institute of Technology (MIT) Ph.D., 2019
- Tsinghua University Yao Class

Work Experience

- Apple | On-device Model Inference Expert

Areas of Expertise

- On-device model inference
- Core ML deployment

Job Objective

- Target Company: Tencent

5

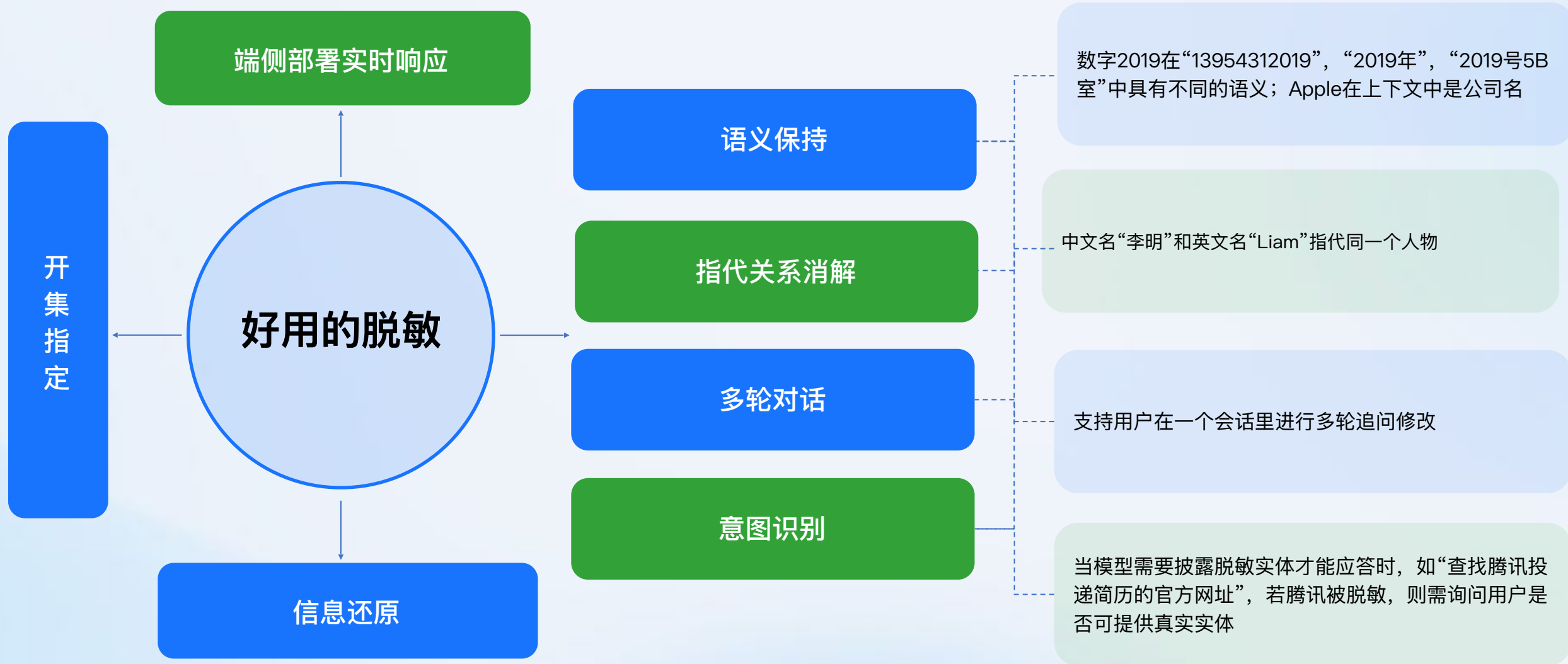
用户： 查找腾讯投递简历的官方网址

6

bot： 腾讯社会招聘投递简历的官方入口：<https://careers.tencent.com/jobopportunity.html>



脱敏需求挖掘





现有隐私保护方案能力矩阵

	实时响应	端侧部署	语义保持	信息还原	开集指定	指代消解	多轮对话	意图识别
密态计算	✗	✗	—	—	—	—	—	—
正则表达式	✓	✓	✗	✗	✗	✗	✗	✗
匿名化	✓	✓	✗	✗	✗	✗	✗	✗
伪匿名化	✓	✓	✗	✓	✗	✗	✗	✗
PII-Redact	✓	✓	✓	✗	✗	✗	✗	✗
Freysa	✓	✓	✓	✓	✗	✗	✗	✗
Presidio	✓	✓	✓	✓	✓	✗	✗	✗
玄武 HaS (ours)	✓	✓	✓	✓	✓	✓	✓	✓

密码学方案 NER+规则方案 模型方案 Agentic方案

PII-Redact (OpenPipe 2025年3月发布): <https://github.com/OpenPipe/pii-redaction>

Freysa (Freysa 2025年8月发布): <https://www.freysa.ai/blueprint/reinforcement-learning-for-privacy-training-local-models-on-the-anonymization-frontier>

Presidio (Microsoft 2018年8月发布并持续更新): <https://microsoft.github.io/presidio/>

02

HaS 的解法

● HaS: 不只是 next level 脱敏技术，也是一种安全交互方式



Key Insight: 寻找大模型友好的实体表达

[ID] 同类ID索引

```
js
const orgs = {
  1: { full: "北京大学",
      short: "北大" }
};
console.log(orgs[1].full)
console.log(orgs[1].short)
```

北京大学: <组织[1].高校.完整名称>

北大: <组织[1].高校.简称>

海淀: <地点[1].行政区.区县名>

二教: <地点[2].教学楼.简称>

<entity_type[ID].category.detail>
实体类型 共引ID 分类属性 细节属性

<> 闭合标签

```
html
<section class="un1">
  <h4>北京大学</h4>
  <p>简称: 北大</p>
  <p>地点: 海淀</p>
</section>
```

. 访问属性

```
python
from types import SimpleNamespace as NS
org = NS(full="北京大学", short="北大")
cli = NS(timeout=3)
site = NS(org=org, conf=NS(client=cli))
print(site.org.full)
print(site.conf.client.timeout)
```

层级命名空间

```
cpp
edu::pku::campus::Building b;
b.set_name("二教");
b.set_district("海淀");
std::cout << b.name() << '\n';
```

第一性原理: 训练数据中越常见 => 对模型越友好

代码数据中存在大量同构的表达形式!



Show Case: 脱敏前原文

CloudGenius的联系人是谁?

****合同编号:**** SAAS-2024-Q3-8801

****甲方 (服务提供方):**** 云创智能有限公司 (以下简称云创智能或CloudGenius)

注册地址: 北京市朝阳区望京科技园A座18层

联系人: 李红 (Li Hong)

电话: 010-8888-9999

****乙方 (客户):**** 繁星贸易集团 (Star Trading Group, 以下简称繁星或STG)

注册地址: 上海市黄浦区南京东路100号

联系人: 张英建 (Zhang Yingjian)

邮箱: zhang.director@star-trading.com

****1. 服务内容****

云创智能向STG提供名为智销云 (SmartSales Cloud) 的客户关系管理(CRM)软件服务。服务等级协议(Service Level Agreement, SLA)保证99.95%的在线时间。云创智能将确保该CRM系统的稳定运行。

****2. 协议期限****

本协议自2024年8月1日起生效, 至2026年7月31日终止, 为期两年。任何一方如需提前终止, 需提前90天书面通知对方。CloudGenius和繁星贸易均应遵守此条款。

****3. 费用与支付****

STG同意支付订阅费用, 总计人民币贰佰万元整 (¥ 2,000,000.00/RMB 2 million)。费用需分两次支付: 协议生效后15个工作日内支付50%, 即100万元; 第二年服务开始前支付剩余50%。支付账户为云创智能在中国工商银行 (ICBC) 开设的账户, 账号为6222020200012345678。

****4. 保密条款****

CloudGenius和STG同意, 对在履行本协议过程中获悉的对方商业秘密和技术信息予以保密。此保密义务在本协议终止后持续有效, 直至该信息进入公共领域。



Show Case: 标签编译后的文本

<组织[001].企业.英文名>的联系人是谁？

****合同编号:**** <编号[001].合同编号.代码>

****甲方 (服务提供方):**** <组织[001].企业.完整名称> (以下简称<组织[001].企业.简称>或<组织[001].企业.英文名>)

注册地址: <地点[001].办公地址.完整地址>

联系人: <人物[001].个人.姓名> (<人物[001].个人.拼音名>)

电话: <电话[001].固定电话.号码>

****乙方 (客户):**** <组织[002].企业.完整名称> (<组织[002].企业.英文名>, 以下简称<组织[002].企业.简称>或<组织[002].企业.英文缩写>)

注册地址: <地点[002].办公地址.完整地址>

联系人: <人物[002].个人.姓名> (<人物[002].个人.拼音名>)

邮箱: <邮箱[001].个人邮箱.地址>

****1. 服务内容****

<组织[001].企业.简称>向<组织[002].企业.英文缩写>提供名为<产品/服务[001].软件.中文名> (<产品/服务[001].软件.英文名>) 的<产品/服务[002].软件服务.完整描述>。<产品/服务[003].协议.中文名>(<产品/服务[003].协议.英文名>, <产品/服务[003].协议.英文缩写>)保证<百分比[001].服务指标.数值>的在线时间。<组织[001].企业.简称>将确保该<产品/服务[002].软件服务.简称>的稳定运行。

****2. 协议期限****

本协议自<日期/时间[001].具体日期.年月日>起生效, 至<日期/时间[002].具体日期.年月日>终止, 为期两年。任何一方如需提前终止, 需提前<日期/时间[003].时间段.天数>书面通知对方。<组织[001].企业.英文名>和繁星贸易均应遵守此条款。

****3. 费用与支付****

<组织[002].企业.英文缩写>同意支付订阅费用, 总计人民币<金额[001].合同金额.中文大写> (<金额[001].合同金额.数字符号>/<金额[001].合同金额.英文表述>)。费用需分两次支付: 协议生效后<日期/时间[004].时间段.工作日数>内支付<百分比[002].支付比例.数值>, 即<金额[002].支付金额.中文数字>; 第二年服务开始前支付剩余<百分比[002].支付比例.数值>。支付账户为<组织[001].企业.简称>在<组织[003].金融机构.完整名称> (<组织[003].金融机构.英文缩写>) 开设的账户, 账号为<编号[002].银行账号.号码>。

****4. 保密条款****

<组织[001].企业.英文名>和<组织[002].企业.英文缩写>同意, 对在履行本协议过程中获悉的对方商业秘密和技术信息予以保密。此保密义务在本协议终止后持续有效, 直至该信息进入公共领域。

不同解法对比

CloudGenius的联系是谁?

合同编号: SAAS-2024-Q3-8801

甲方 (服务提供方): 云创智能有限公司 (以下简称云创智能或CloudGenius)

联系人: 李红 (Li Hong)

乙方 (客户): 繁星贸易集团 (Star Trading Group, 以下简称繁星或STG)

联系人: 张英建 (Zhang Yingjian)

实体类型标签

<ORGANIZATION> 的联系是谁?

合同编号: <CONTRACT_ID>

甲方 (服务提供方): <ORGANIZATION> (以下简称 <ORGANIZATION> 或 <ORGANIZATION>)

联系人: <PERSON> (<PERSON>)

乙方 (客户): <ORGANIZATION> (<ORGANIZATION>, 以下简称 <ORGANIZATION> 或 <ORGANIZATION>)

联系人: <PERSON> (<PERSON>)

指代关系丢失模型无法回答 ❌

假名替换

SkyGenius 的联系是谁?

合同编号: SAAS-2023-Q2-9732

甲方 (服务提供方): 华新智联科技有限公司 (以下简称 华新智联 或 SkyGenius)

联系人: 王敏 (Wang Min)

乙方 (客户): 晨耀贸易集团 (Stellar Trading Group, 以下简称 晨耀 或 STG)

联系人: 周建业 (Zhou Jianye)

违反《个保法》第8条对数据质量的要求, 不方便解析与校验 ❌

结构化语义标签

<组织[001].企业.英文名>的联系是谁?

合同编号: <编号[001].合同编号.代码>

甲方 (服务提供方): <组织[001].企业.完整名称> (以下简称<组织[001].企业.简称>或<组织[001].企业.英文名>)

联系人: <人物[001].个人.姓名> (<人物[001].个人.拼音名>)

乙方 (客户): <组织[002].企业.完整名称> (<组织[002].企业.英文名>, 以下简称<组织[002].企业.简称>或<组织[002].企业.英文缩写>)

联系人: <人物[002].个人.姓名> (<人物[002].个人.拼音名>)

模型可正确回答, 合规, 方便解析与校验 ✅

DeepSeek正确作答

根据提供的信息:

<组织[001].企业.完整名称>联系人是: <人物[001].个人.姓名>

在合同中, 联系人作为甲方 (即服务提供方) 的代表。

HaS Agentic 框架

端云协同
模算分工

云侧

角色与语义
约束

标签推理
规则

最小信息
披露

任务策略
模板

协同
编排

系统提示词
注入

上下文
中间人

算法
工具

GestaltTag
Align

标签感知分词

多策略映射
搜索

复合标签段
实体裁剪

自检召回
兜底链

长程记忆

多轮映射
累积流

端侧

模型
能力

隐私实体识别

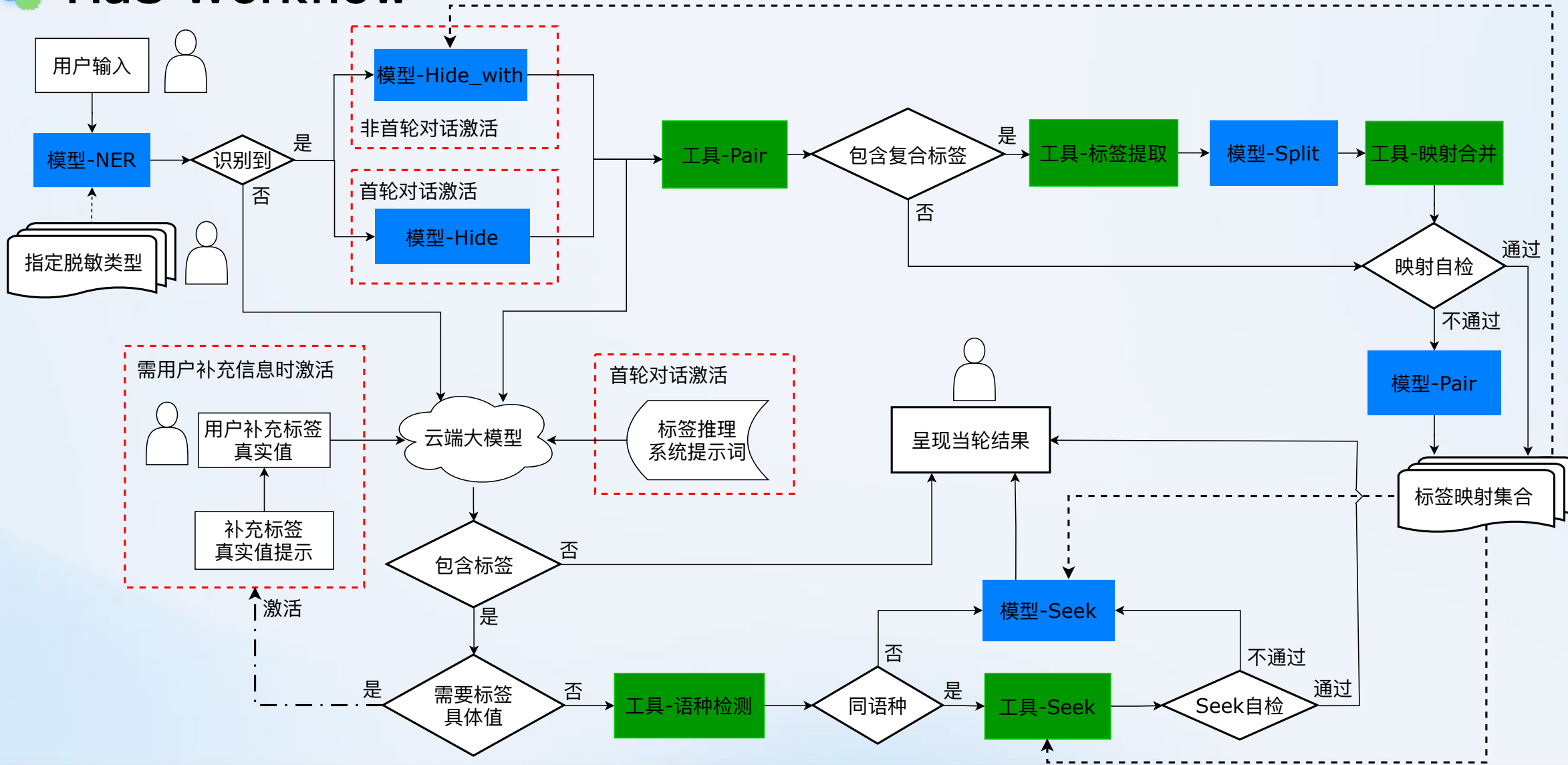
实体标签编译

标签映射生成

标签映射拆分

隐私实体还原

序贯还原
滚动流



Take Away



小模型训练也要追求领域内泛化

RL时要分析负样本的Root-Cause

合理切分多轮阶段，上一轮的输出即下一轮的CoT

在保持最小上下文和维持Prompt Cache间寻找平衡

能用算法工具解决，就不要用模型能力

工具轻召回重准确率，用模型能力兜底

03 现场演示



端侧小模型

HaS脱敏

欢迎体验



模型开源: https://huggingface.co/xuanwulab/HaS_4.0_0.6B

Chrome/Edge应用商店: HaS脱敏



HaS开发者交流群



HaS用户意见反馈群

04 未来展望



更多模态支持

图像、音频、视频、多媒体文档

更快的模型推理速度

自定义标签token、扩散模型



更标准化的接口

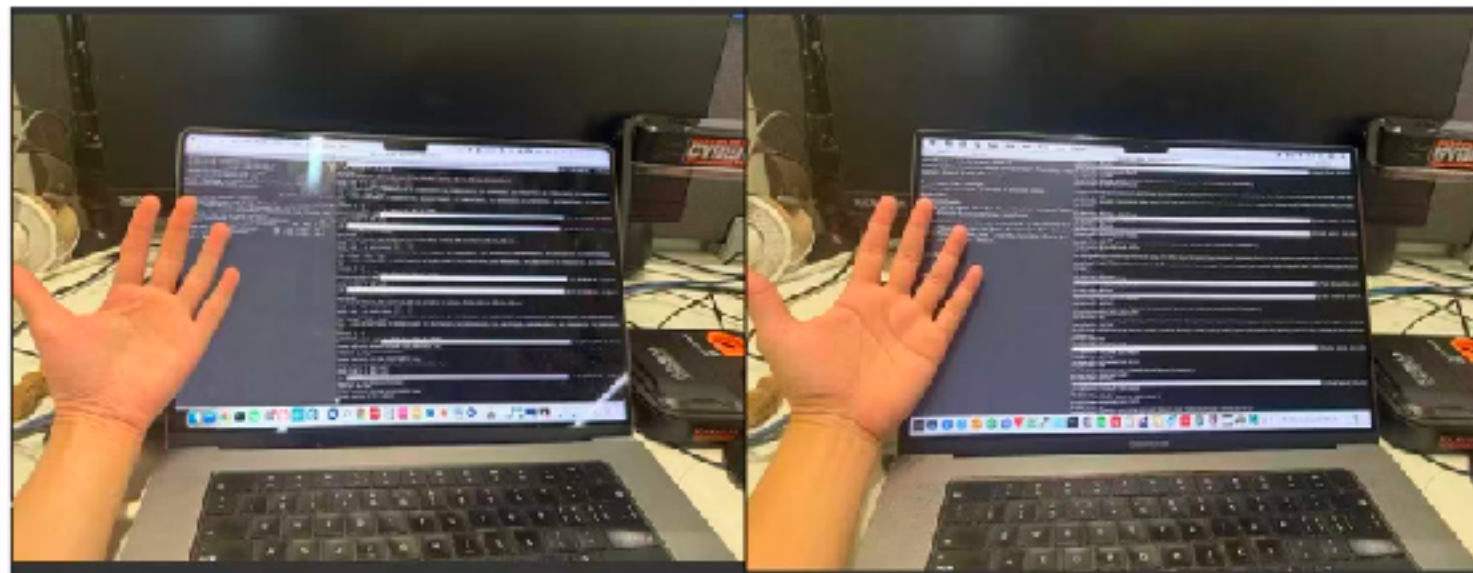
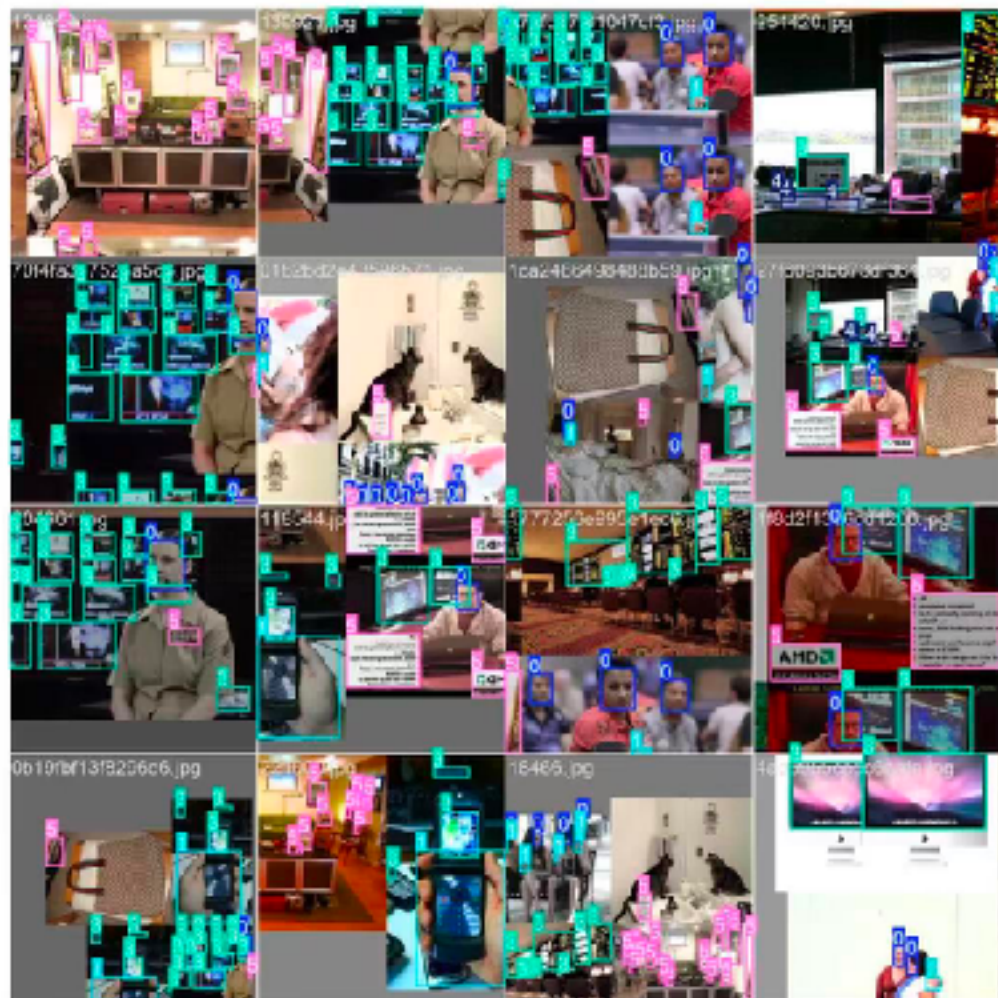
MCP、A2A、ANP、ACP、NLWeb
成为下一代 Agent OS 的基础组件

更广泛的场景支持

AI编程、智能体应用、Agentic workflow
交互者由“人”扩展到 Agent

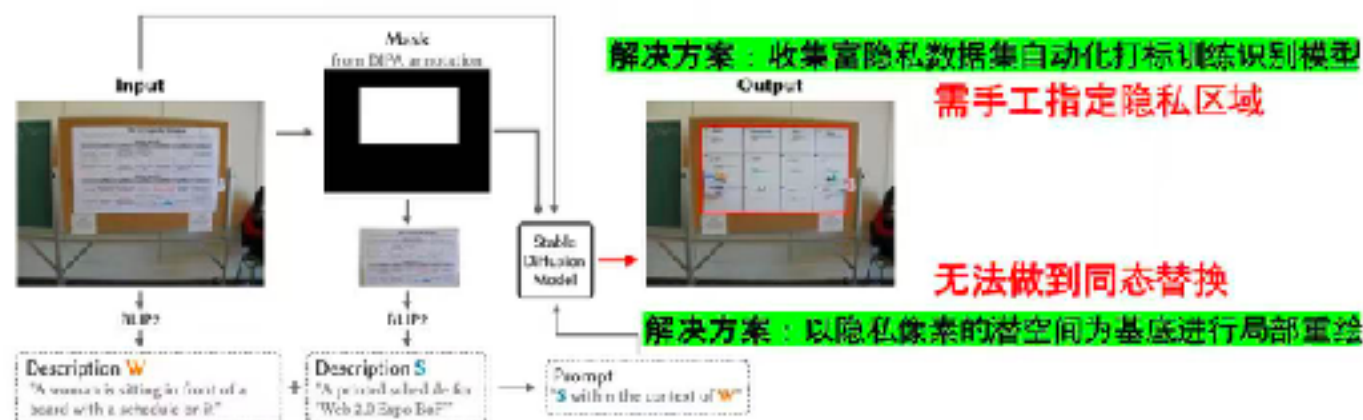
HaS 未来展望

HaS多模态脱敏探索



脱敏前

脱敏后



📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AI Ops
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AICon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AICon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AICon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AICon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LM Ops
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

