QCon
全球软件开发大会

Serverless助力大语言模型工程化实践

演讲人：姬军翔

亚马逊云科技 / 高级解决方案架构师

目 录

01 LMSI模型

02 LLM应用案例分析

03 Serverless最佳实践

04 回顾和总结

关于我

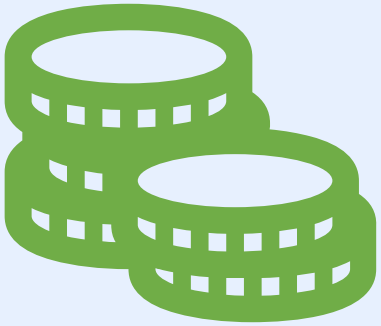
架构师
解决方案
原型
AWS
搜索
开发
客户
微服务
工程师
LLM
实施
团队
技术
项目
设计
物流
机器学习
架构
聊天机器人
跨境
Git
NLP
AI
索引
协同过滤
奢侈品
状态机
API
CTO
Scrum
Java
Solr
RAG
BM25
TF-IDF
ELK
推荐系统
文本到SQL
知识库
内容审核
版本控制
长词优先
呼叫中心
GITEX18
AVAYA
Farfetch
Social Bots

01 LMSI

语言模型系统接口模型

构建大语言模型应用的常见挑战

准确度，性能不达标



训练和部署昂贵

合规，构建复杂

一些应对方法

准确度，性能不达标	部署成本高	构建复杂
↓	↓	↓
提示工程	使用托管模型	Serverless/框架

语言模型系统接口模型(LMSI)

高

抽象程度

低

OSI 7层网络模型

1. 应用层 (Application Layer)
2. 表示层 (Presentation Layer)
3. 会话层 (Session Layer)
4. 传输层 (Transport Layer)
5. 网络层 (Network Layer)
6. 数据链路层 (Data Link Layer)
7. 物理层 (Physical Layer)

LMSI v2

1. 应用层（安全和用户管理，部署）
2. 优化层（提示词的自动优化）
3. 控制层 （合理的分片逻辑/并发推理等）
4. 提示约束层 （单独的前置/后置合规模型）
5. 提示层（多种提示模块的组合）
6. 神经网络层（单个或多个大语言模型）
7. 算力层（GPU/CPU）

https://www.twosigma.com/articles/a-guide-to-large-language-model-abstractions/

LMSIv2和大语言模型框架的对应关系

		AutoGen	LangChain	LlamaIndex	DSPY	HF Transformers	云厂商
1	应用层	😀	😀	😀			😀
2	优化层				😀		😀
3	控制层		😀	😀	😀		😀
4	提示约束层		😀	😀	😀		😀
5	提示层		😀	😀	😀		😀
6	神经网络层				😀	😀	😀
7	算力层						😀

02

案例分析

基于 Serverless 的大语言模型翻译应用

基于大语言模型的游戏内容翻译

业务声明

某游戏公司需要翻译游戏的多语言版本，因为存在较多游戏中特有的地名，人名，机器翻译的效果不好，主要是人工翻译为主，翻译的时间根据工作量的不同从数天到数周不等，业务团队希望利用大语言模型加速翻译过程并降低翻译成本。

```
"mapping" : {
        "CHS" : "奇怪的渔人吐司",
        "CHT" : "奇怪的漁人吐司",
        "DE" : "Misslungene Fischerschnitte",
        "EN": "Suspicious Fisherman’s Toast",
```

快速发现挑战和对应的解决方案

专有名词的翻译

长文本翻译

准确度提升

翻译风格

合规合法

部署

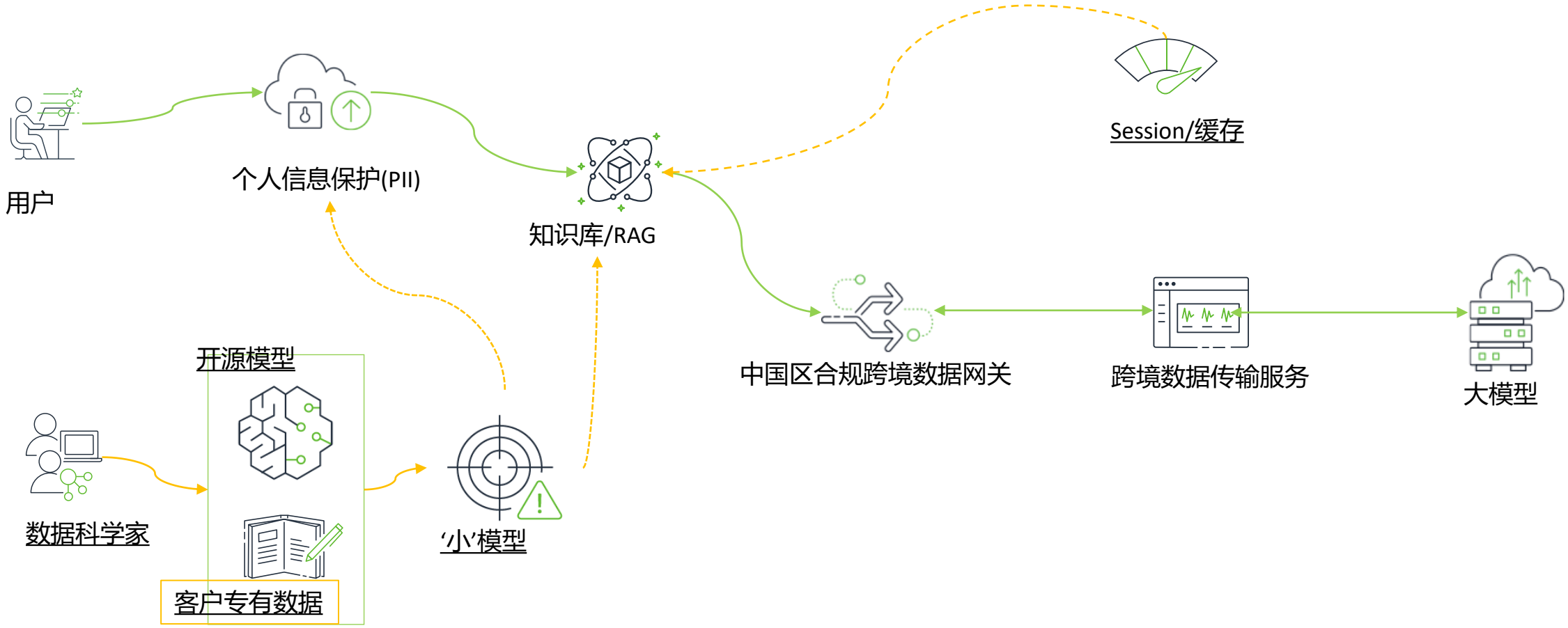
应用评估

效果反馈

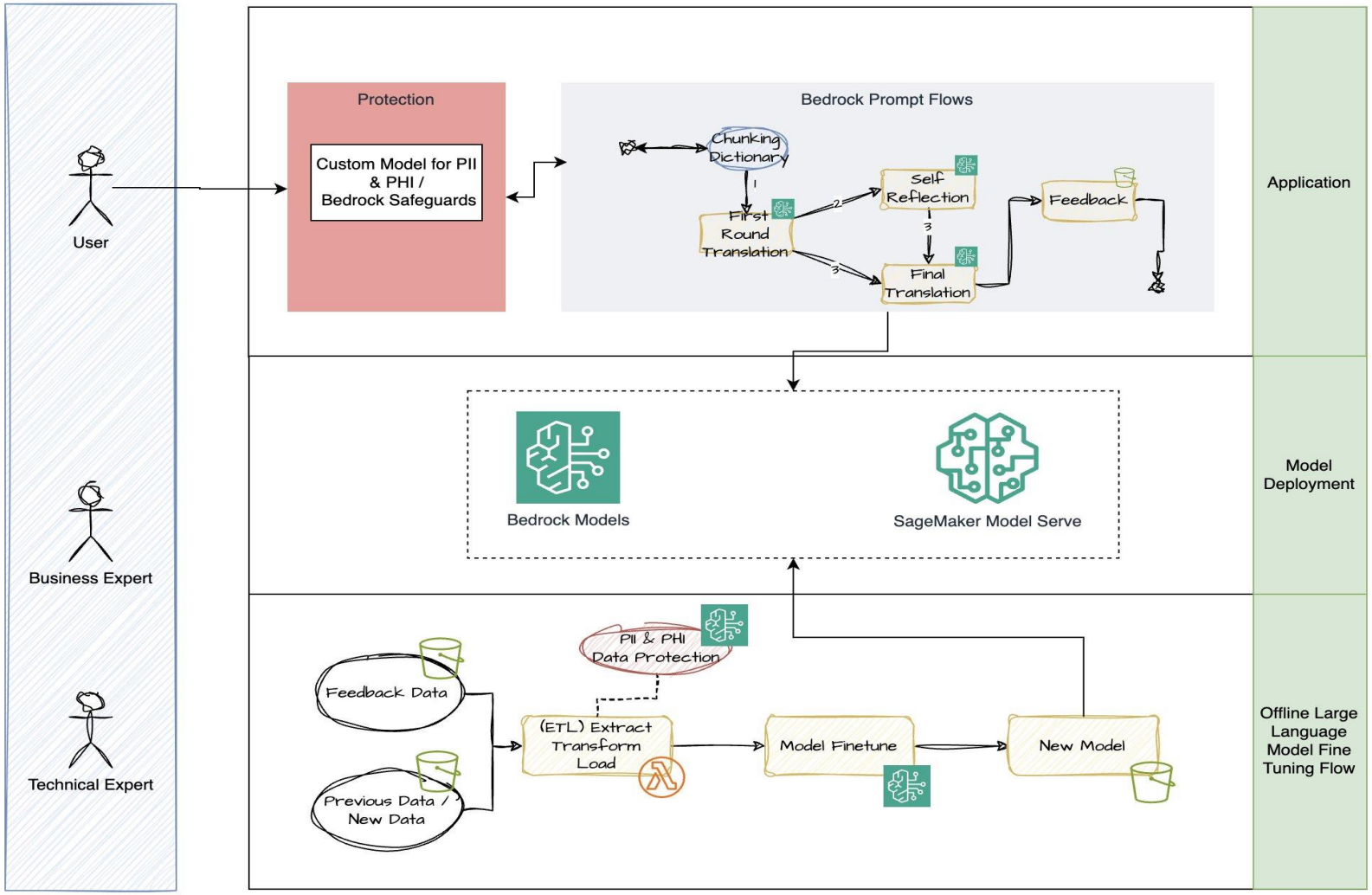
AI相关法规

名称	《互联网信息服务深度合成管理规定》	《生成式人工智能服务管理暂行办法》
信息安全要求	第七条 深度合成服务提供者应当落实信息安全主体责任，建立健全用户注册、算法机制机理审核、科技伦理审查、信息发布审核、数据安全、个人信息保护、反电信网络诈骗、应急处置等管理制度，具有安全可控的技术保障措施。	输入 Input 第十一条提供者对使用者的输入信息和使用记录应当依法履行保护义务，不得收集非必要个人信息，不得非法留存能够识别使用者身份的输入信息和使用记录，不得非法向他人提供使用者的输入信息和使用记录。 输出 Output 第九条提供者应当依法承担网络信息内容生产者责任，履行网络信息安全义务。 涉及个人信息的，依法承担个人信息处理者责任，履行个人信息保护义务。

完整的大模型应用的复杂性



通用的大语言模型应用的解决方案



03

最佳实践

用Serverless承载大模型应用