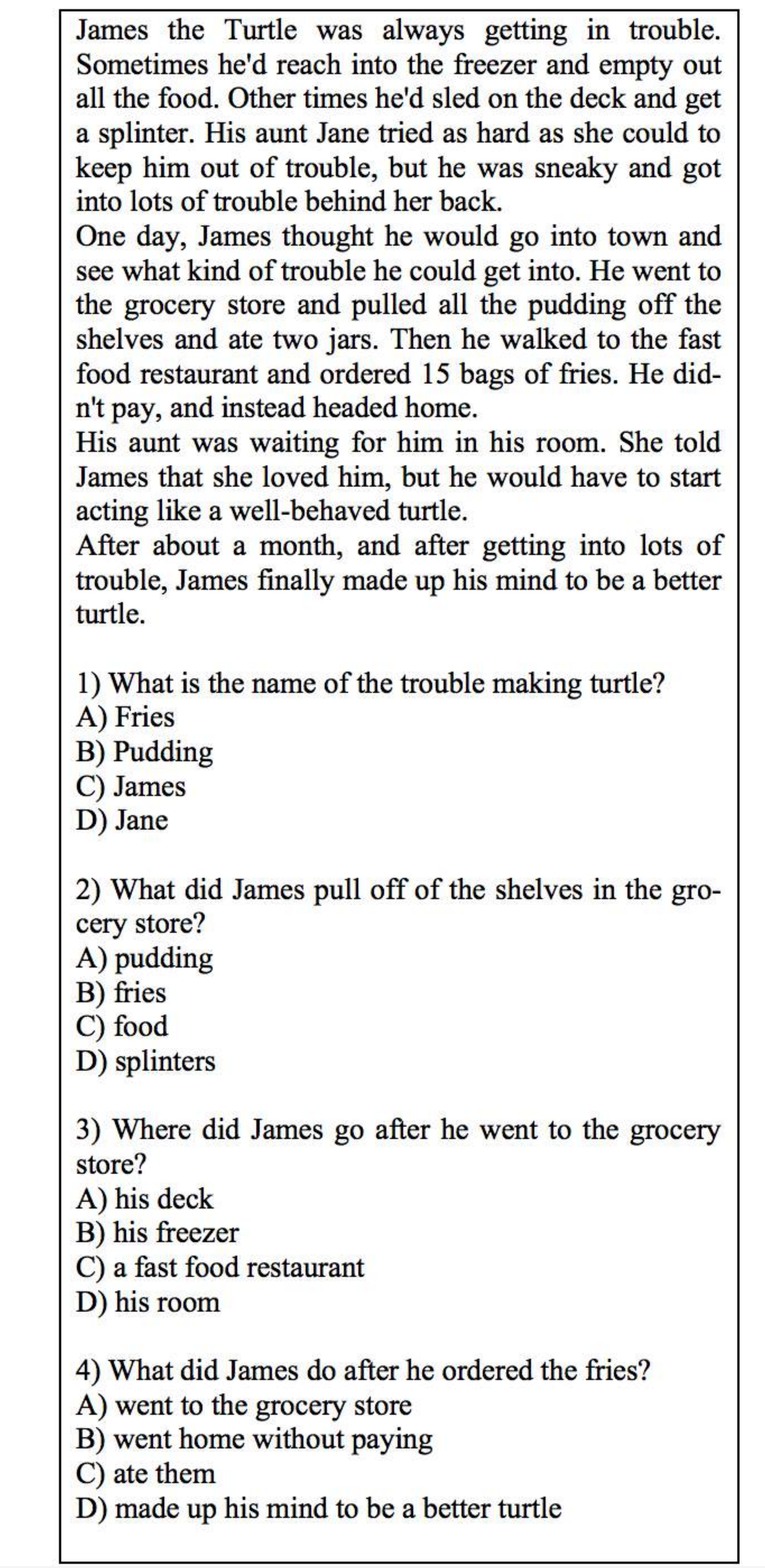
公开数据集-选择题

- 选择题：微软的MCTest
  - ✓ 真实英文儿童读物
  - ✓ 每篇150-300词
  - ✓ 要求从4个选项中选出正确答案
  - ✓ 数量较少，分160篇和500篇两种



James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

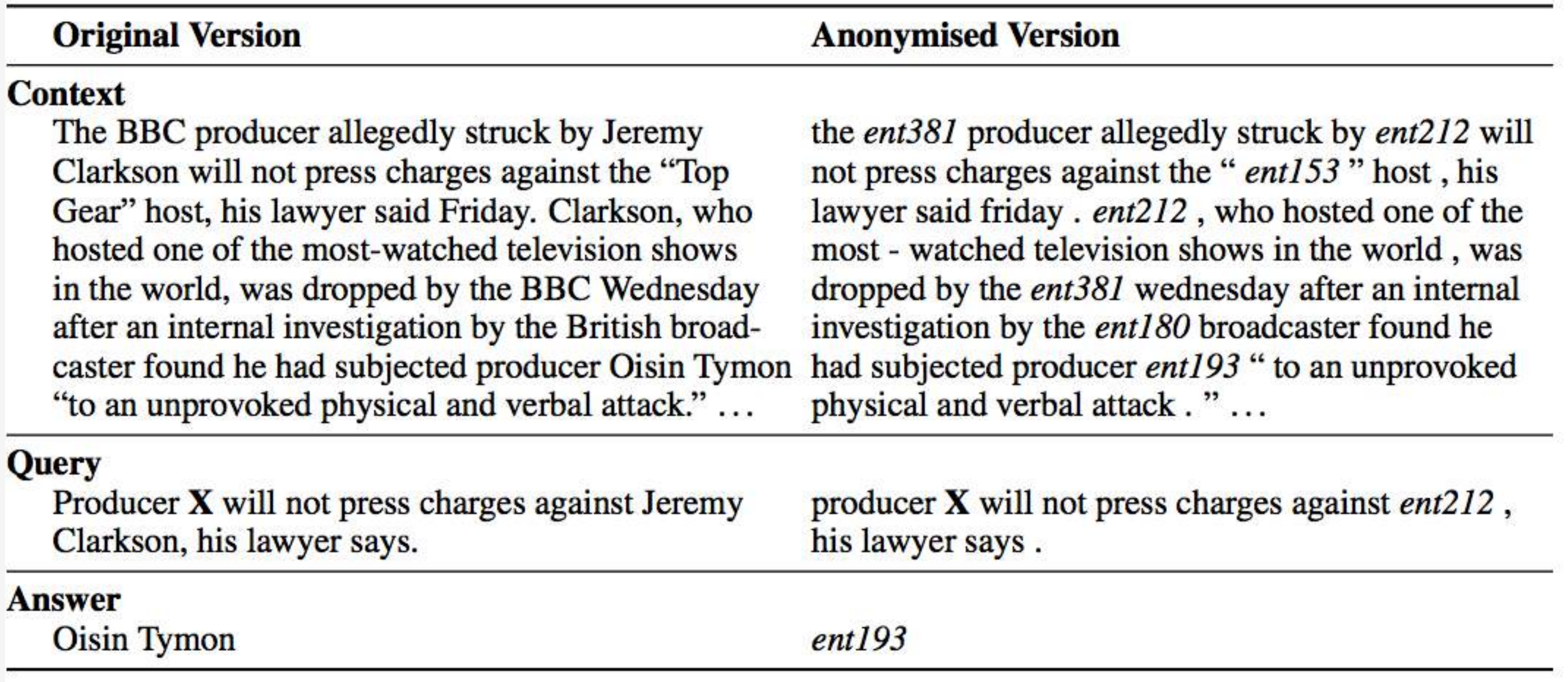
2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

公开数据集-完形填空

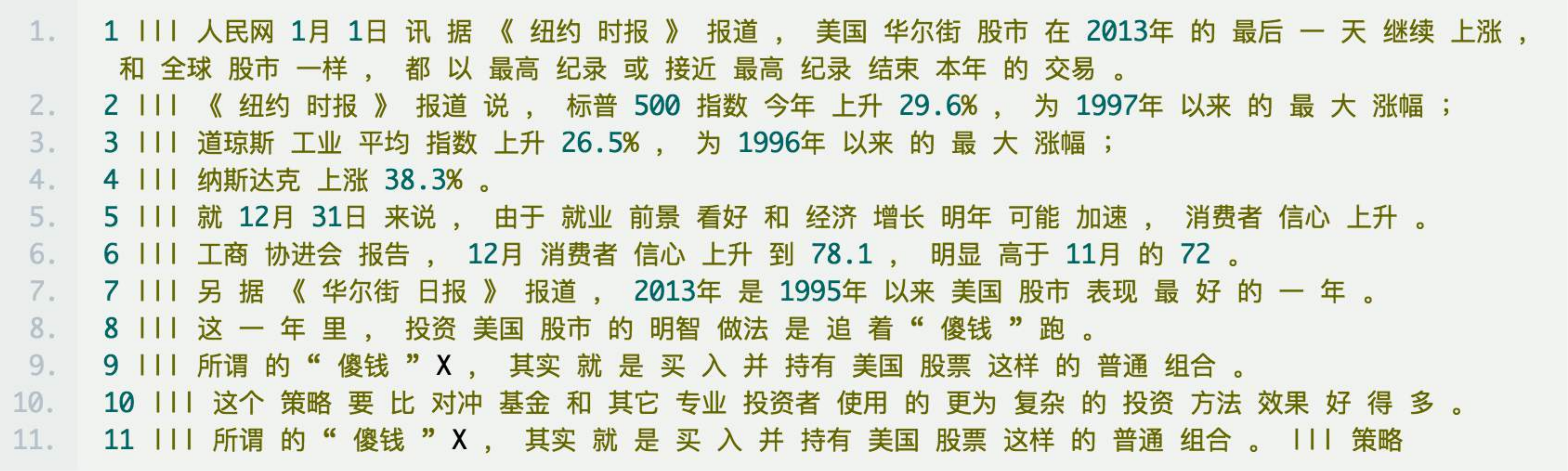
- 完形填空：DeepMind的CNN和DailyMail数据集
  - ✓ 真实新闻数据
  - ✓ 自动标注产生
  - ✓ 要求回答被抽掉的实体，实体在文中出现过
  - ✓ 数量较大，CNN9万篇，DailyMail 22万篇



	Original Version	**Anonymised Version**
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” . . .	the *ent381* producer allegedly struck by *ent212* will not press charges against the “ *ent153* ” host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* “ to an unprovoked physical and verbal attack . ” . . .
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer		
	Oisin Tymon	*ent193*

公开数据集-完形填空

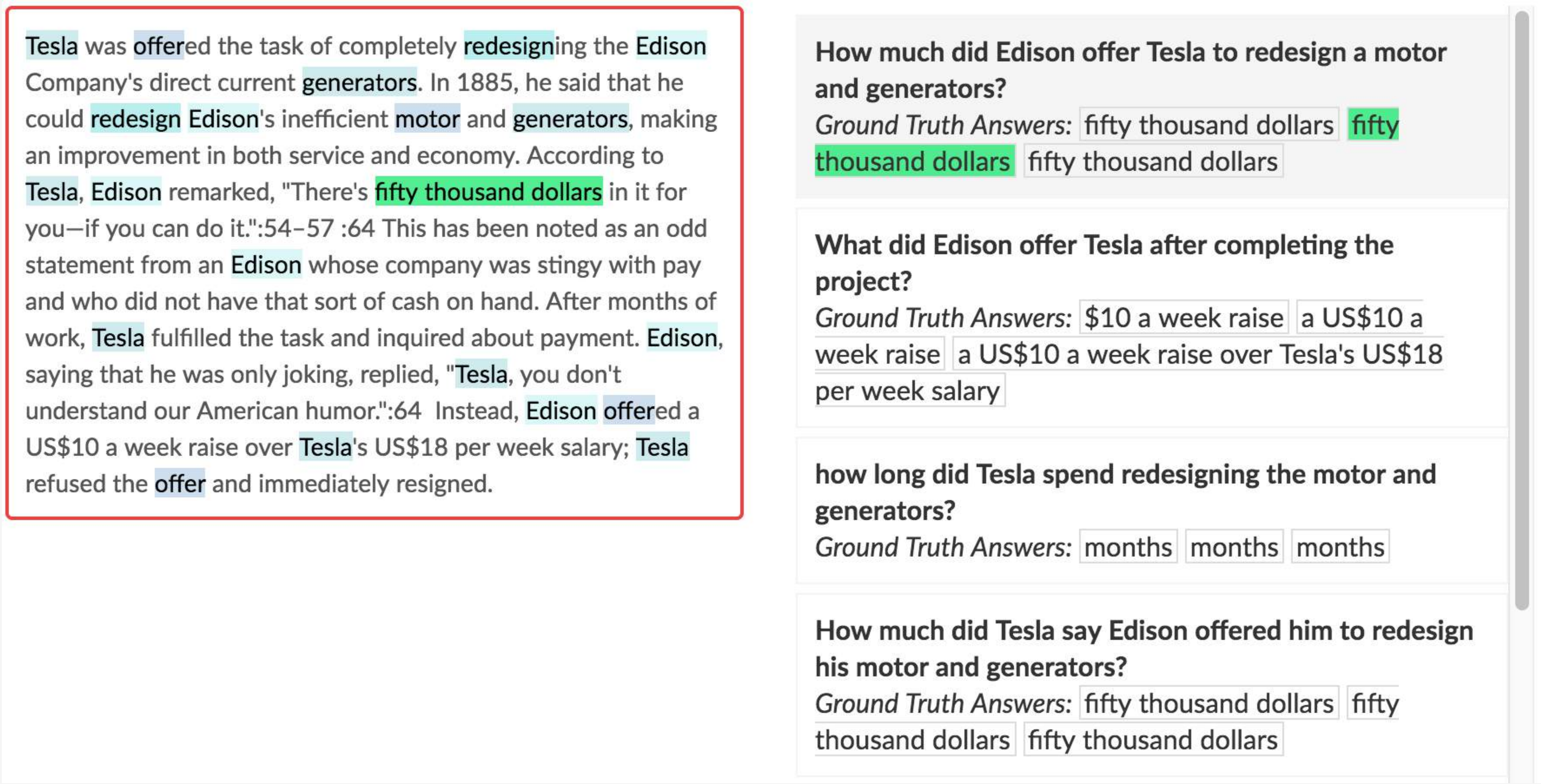
- 完形填空：讯飞和哈工大的中文数据集
  - ✓ 真实新闻数据
  - ✓ 自动标注产生
  - ✓ 要求回答被抽掉的实体
  - ✓ 数量较大，共87万篇



```
1 ||| 人民网 1月 1日 讯 据 《 纽约 时报 》 报道 ， 美国 华尔街 股市 在 2013年 的 最后 一 天 继续 上涨 ， 和 全球 股市 一样 ， 都 以 最高 纪录 或 接近 最高 纪录 结束 本年 的 交易 。
2 ||| 《 纽约 时报 》 报道 说 ， 标普 500 指数 今年 上升 29.6% ， 为 1997年 以来 的 最 大 涨幅 ；
3 ||| 道琼斯 工业 平均 指数 上升 26.5% ， 为 1996年 以来 的 最 大 涨幅 ；
4 ||| 纳斯达克 上涨 38.3% 。
5 ||| 就 12月 31日 来说 ， 由于 就业 前景 看好 和 经济 增长 明年 可能 加速 ， 消费者 信心 上升 。
6 ||| 工商 协进会 报告 ， 12月 消费者 信心 上升 到 78.1 ， 明显 高于 11月 的 72 。
7 ||| 另 据 《 华尔街 日报 》 报道 ， 2013年 是 1995年 以来 美国 股市 表现 最 好 的 一 年 。
8 ||| 这 一 年 里 ， 投资 美国 股市 的 明智 做法 是 追 着 “ 傻钱 ” 跑 。
9 ||| 所谓 的 “ 傻钱 ” X ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。
10 ||| 这个 策略 要 比 对冲 基金 和 其它 专业 投资者 使用 的 更为 复杂 的 投资 方法 效果 好 得 多 。
11 ||| 所谓 的 “ 傻钱 ” X ， 其实 就 是 买 入 并 持有 美国 股票 这样 的 普通 组合 。 ||| 策略
```

公开数据集-SQuAD

- 可变长答案数据集：斯坦福的SQuAD
  - ✓ 答案是文章中出现的任意长度片段
  - ✓ Wiki文章为主
  - ✓ 众包人工标注产生
  - ✓ 每个问题3人标注， 降低人工标注误差
  - ✓ 数量较大：500多篇文章， 2万多个段落，10万个问题
  - ✓ 鼓励用自己的语言提问，增加多样性



Tesla was offered the task of completely redesigning the Edison Company's direct current generators. In 1885, he said that he could redesign Edison's inefficient motor and generators, making an improvement in both service and economy. According to Tesla, Edison remarked, "There's fifty thousand dollars in it for you—if you can do it.":54–57 :64 This has been noted as an odd statement from an Edison whose company was stingy with pay and who did not have that sort of cash on hand. After months of work, Tesla fulfilled the task and inquired about payment. Edison, saying that he was only joking, replied, "Tesla, you don't understand our American humor.":64 Instead, Edison offered a US$10 a week raise over Tesla's US$18 per week salary; Tesla refused the offer and immediately resigned.

How much did Edison offer Tesla to redesign a motor and generators?
Ground Truth Answers: fifty thousand dollars | fifty thousand dollars | fifty thousand dollars

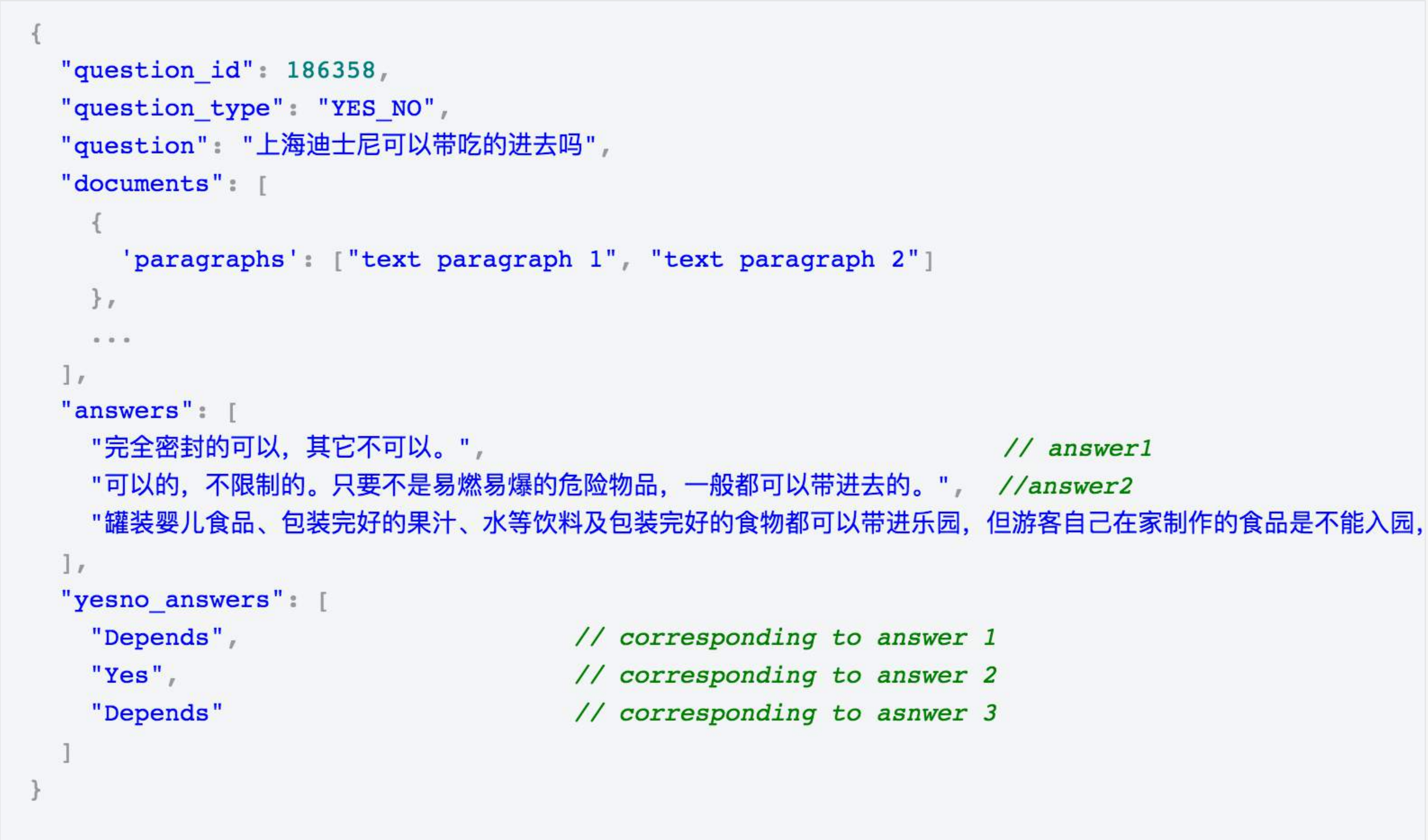
What did Edison offer Tesla after completing the project?
Ground Truth Answers: $10 a week raise | a US$10 a week raise | a US$10 a week raise over Tesla's US$18 per week salary

how long did Tesla spend redesigning the motor and generators?
Ground Truth Answers: months | months | months

How much did Tesla say Edison offered him to redesign his motor and generators?
Ground Truth Answers: fifty thousand dollars | fifty thousand dollars | fifty thousand dollars

公开数据集-DuReader

- 多任务中文数据集：百度DuReader
  - ✓ 多个任务：Description、Entity、Yes_No
  - ✓ 问题来自真实的user query
  - ✓ 文档和答案从百度搜索和百度知道中获得
  - ✓ 答案可以不在文章中出现
  - ✓ 大型数据集：20w问题、94万文章、42万答案



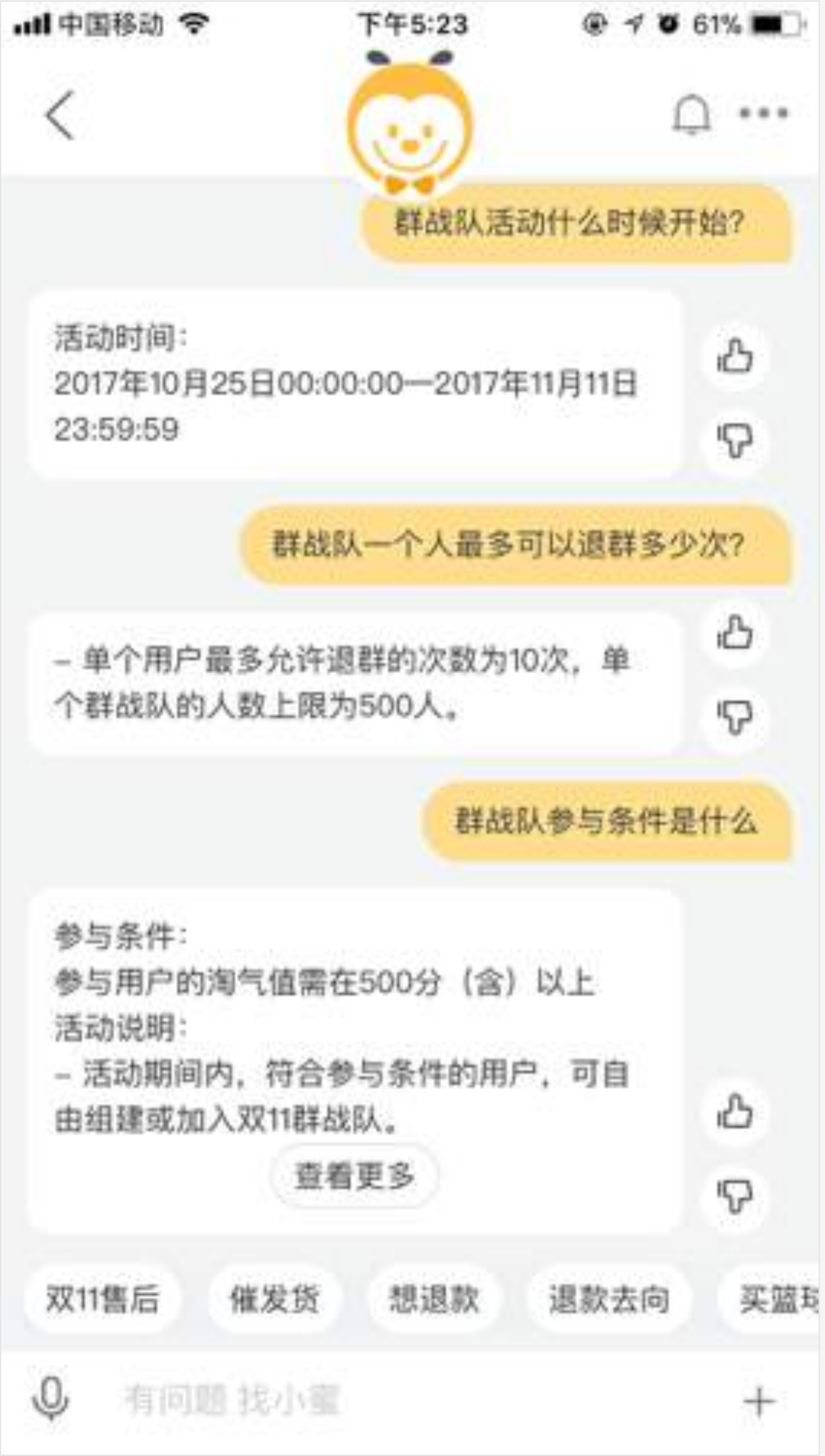
```
{
  "question_id": 186358,
  "question_type": "YES_NO",
  "question": "上海迪士尼可以带吃的进去吗",
  "documents": [
    {
      'paragraphs': ["text paragraph 1", "text paragraph 2"]
    },
    ...
  ],
  "answers": [
    "完全密封的可以，其它不可以。",                                  // answer1
    "可以的，不限制的。只要不是易燃易爆的危险物品，一般都可以带进去的。",  //answer2
    "罐装婴儿食品、包装完好的果汁、水等饮料及包装完好的食物都可以带进乐园，但游客自己在家制作的食品是不能入园，
  ],
  "yesno_answers": [
    "Depends",                      // corresponding to answer 1
    "Yes",                          // corresponding to answer 2
    "Depends"                       // corresponding to asnwer 3
  ]
}
```

TABLE OF CONTENTS 大纲

- 阿里小蜜平台介绍
- 机器阅读理解技术概览
- 业务场景及技术实践
- 挑战与展望

活动规则解读场景

- 阿里小蜜活动规则解读
  - ✓ 服务双十一等线上电商活动
  - ✓ 每个活动都有活动规则文档
  - ✓ 活动频繁且生效时间短
  - ✓ 替代人工配置FAQ



群战队活动什么时候开始?

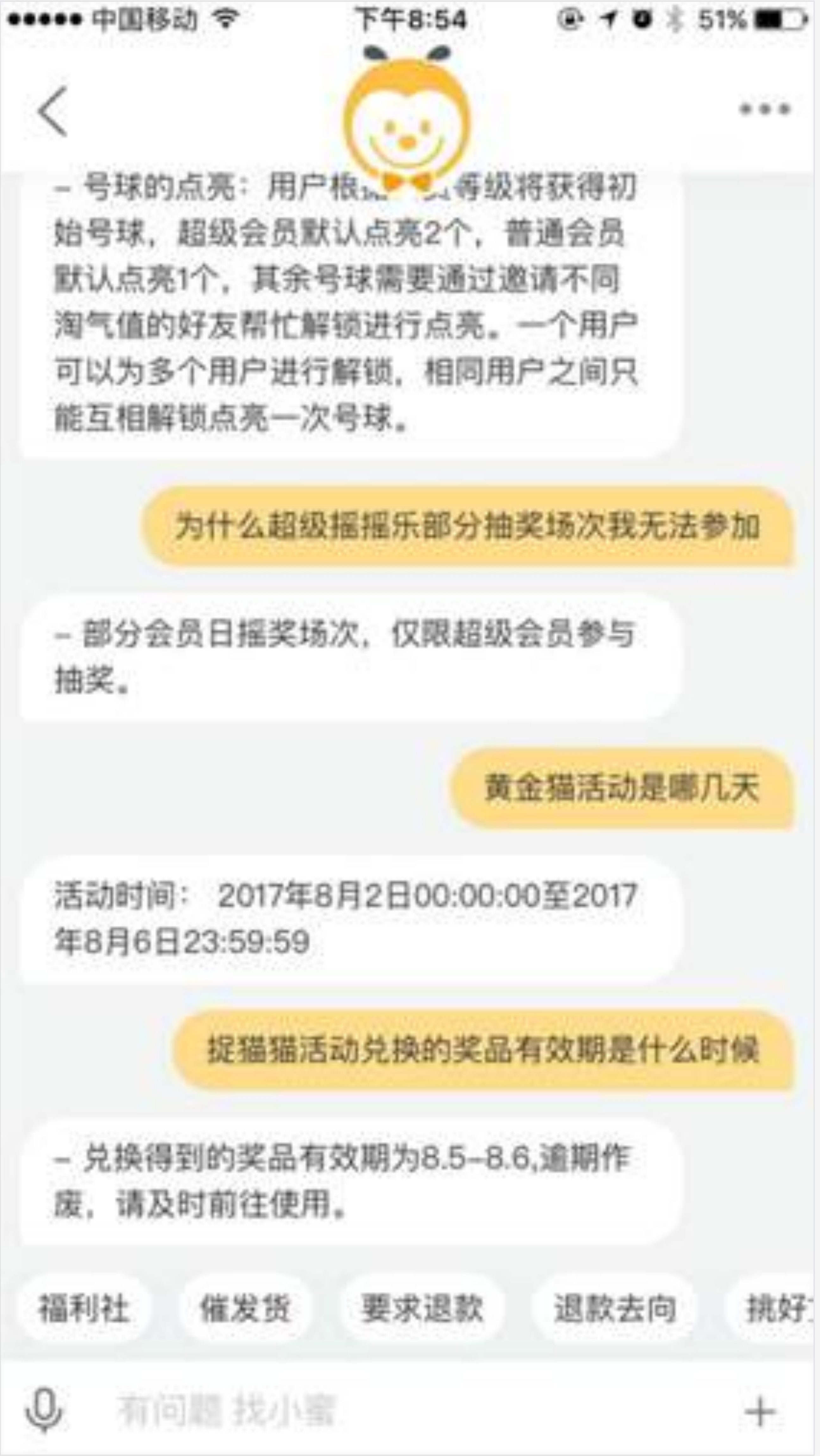
活动时间:
2017年10月25日00:00:00—2017年11月11日 23:59:59

群战队一个人最多可以退群多少次?

– 单个用户最多允许退群的次数为10次，单个群战队的人数上限为500人。

群战队参与条件是什么

参与条件:
参与用户的淘气值需在500分（含）以上
活动说明:
– 活动期间内，符合参与条件的用户，可自由组建或加入双11群战队。



– 号球的点亮：用户根[illegible]等级将获得初始号球，超级会员默认点亮2个，普通会员默认点亮1个，其余号球需要通过邀请不同淘气值的好友帮忙解锁进行点亮。一个用户可以为多个用户进行解锁，相同用户之间只能互相解锁点亮一次号球。

为什么超级摇摇乐部分抽奖场次我无法参加

– 部分会员日摇奖场次，仅限超级会员参与抽奖。

黄金猫活动是哪几天

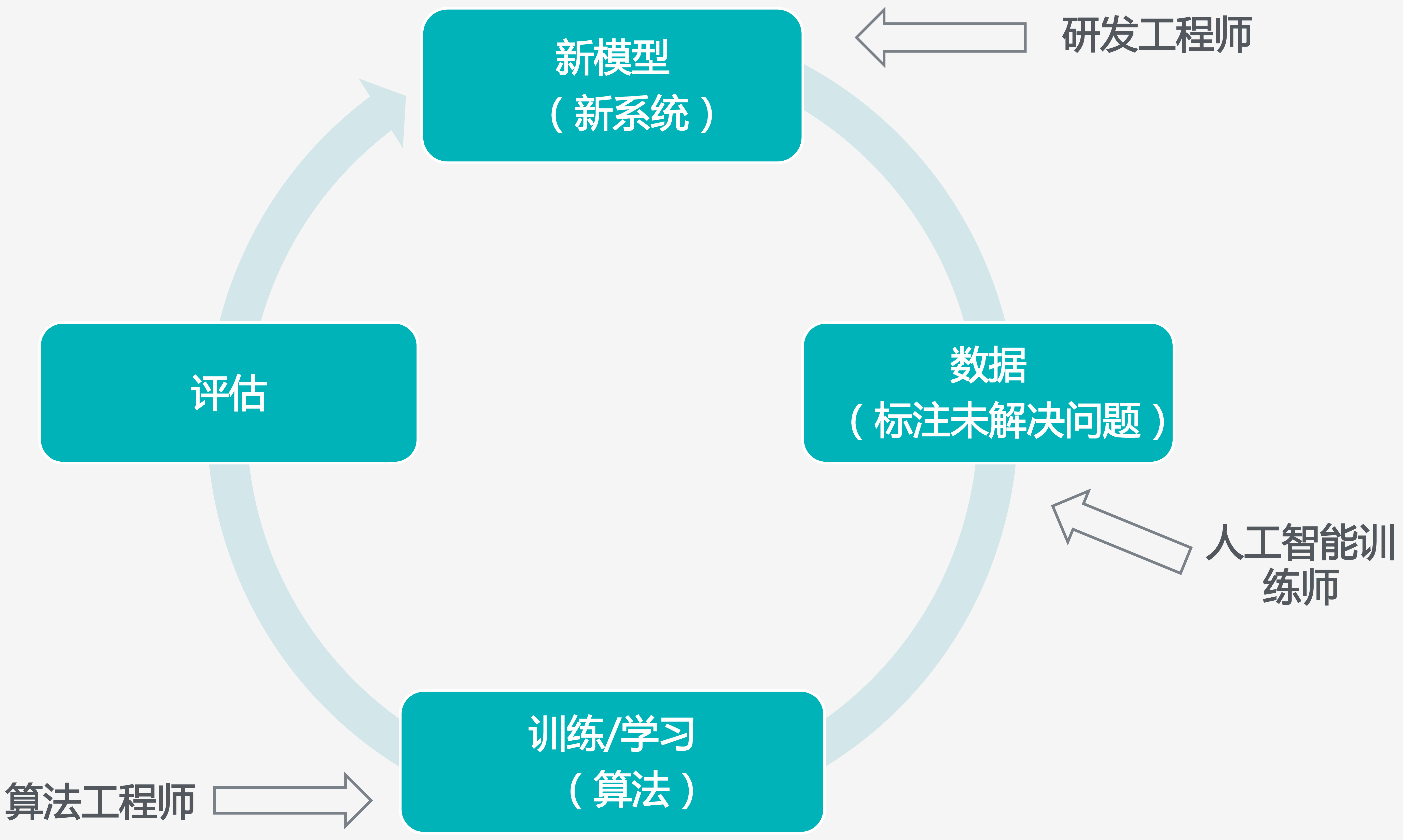
活动时间：2017年8月2日00:00:00至2017年8月6日23:59:59

捉猫猫活动兑换的奖品有效期是什么时候

– 兑换得到的奖品有效期为8.5-8.6,逾期作废，请及时前往使用。

业务数据集的构建

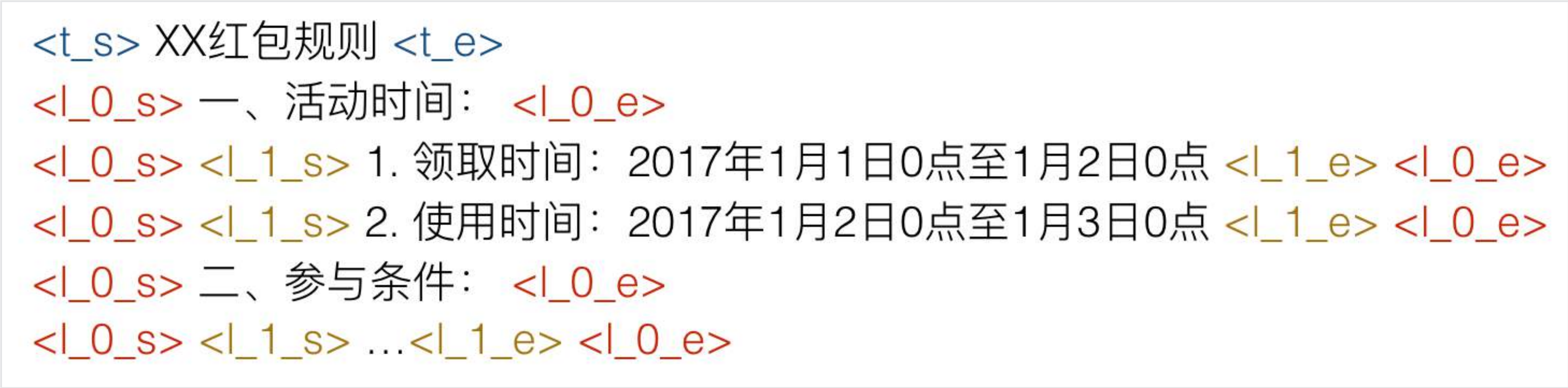
- 阿里小蜜活动规则解读
  - ✓ 采用真实问题的数据分布
  - ✓ 初始标注业务数据6.5w+
  - ✓ 构建AI Boost数据模型闭环



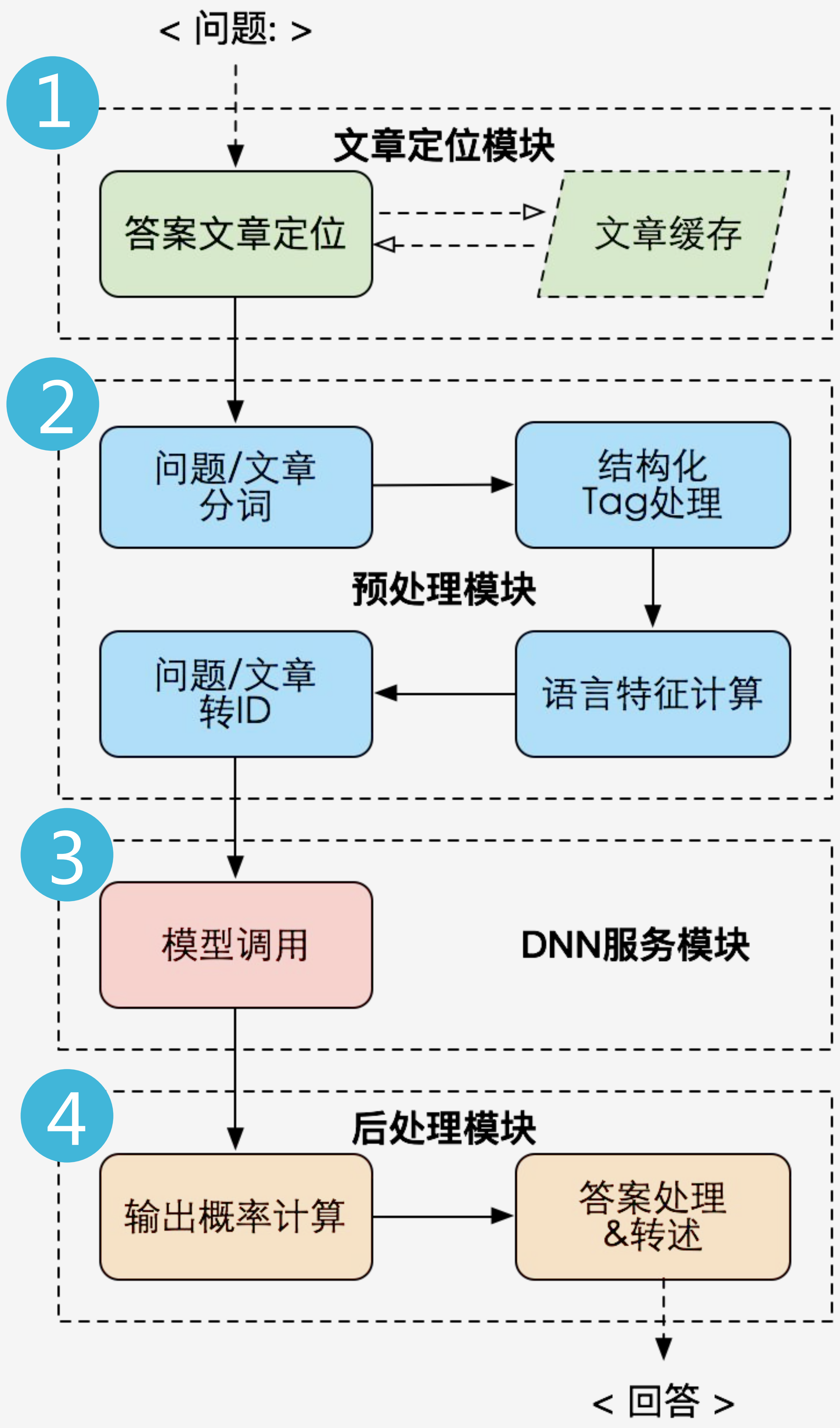
AI Boost数据模型闭环

基于机器阅读的问答处理流程

- 1. 文章片段定位
  - ✓ 针对用户问题，召回候选文档段落集合
  - ✓ 借助文本分类、检索或者问题模板辅助
- 2. 输入预处理
  - ✓ 格式归一，特征预计算
  - ✓ 问题及相应段落向量表征
  - ✓ 生成文档结构标签

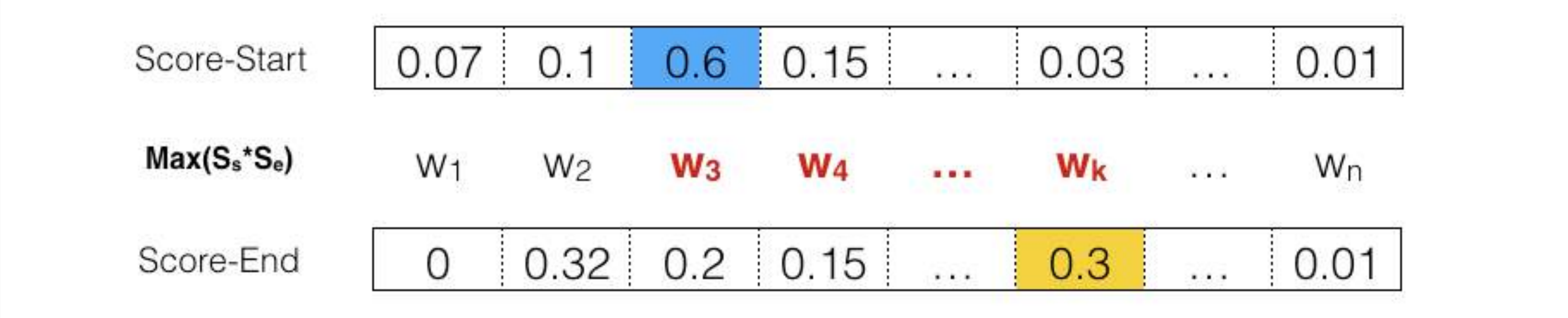


<t_s> XX红包规则 <t_e>
<l_0_s> 一、活动时间： <l_0_e>
<l_0_s> <l_1_s> 1. 领取时间：2017年1月1日0点至1月2日0点 <l_1_e> <l_0_e>
<l_0_s> <l_1_s> 2. 使用时间：2017年1月2日0点至1月3日0点 <l_1_e> <l_0_e>
<l_0_s> 二、参与条件： <l_0_e>
<l_0_s> <l_1_s> …<l_1_e> <l_0_e>

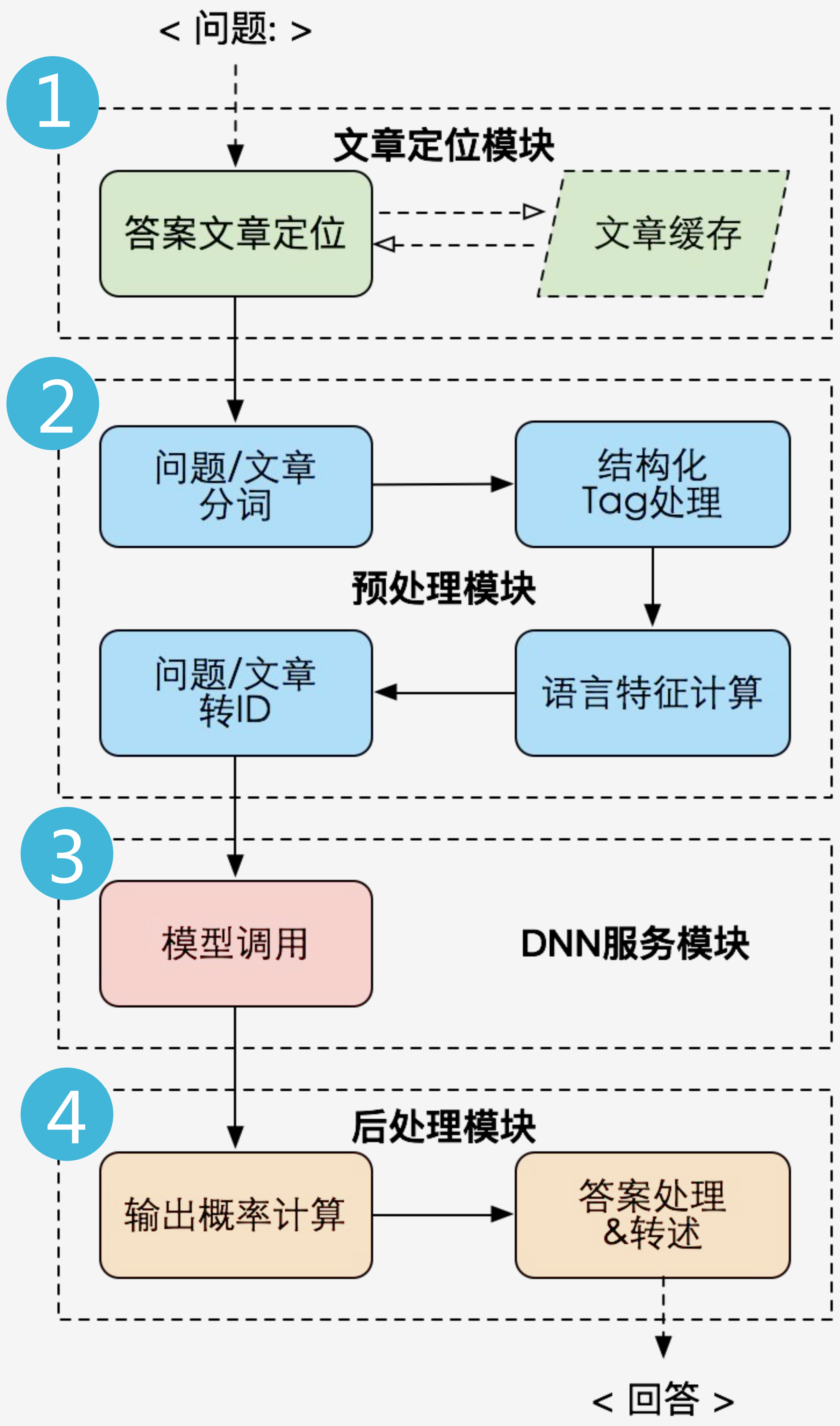


基于机器阅读的问答处理流程

- 3. 在线预测服务
  - ✓ GPU-Based 模型加载及服务驱动
  - ✓ 预测段落中词或符号得分
- 4. 后处理机制
  - ✓ 基于动态规划选取最佳文本短语作为输出

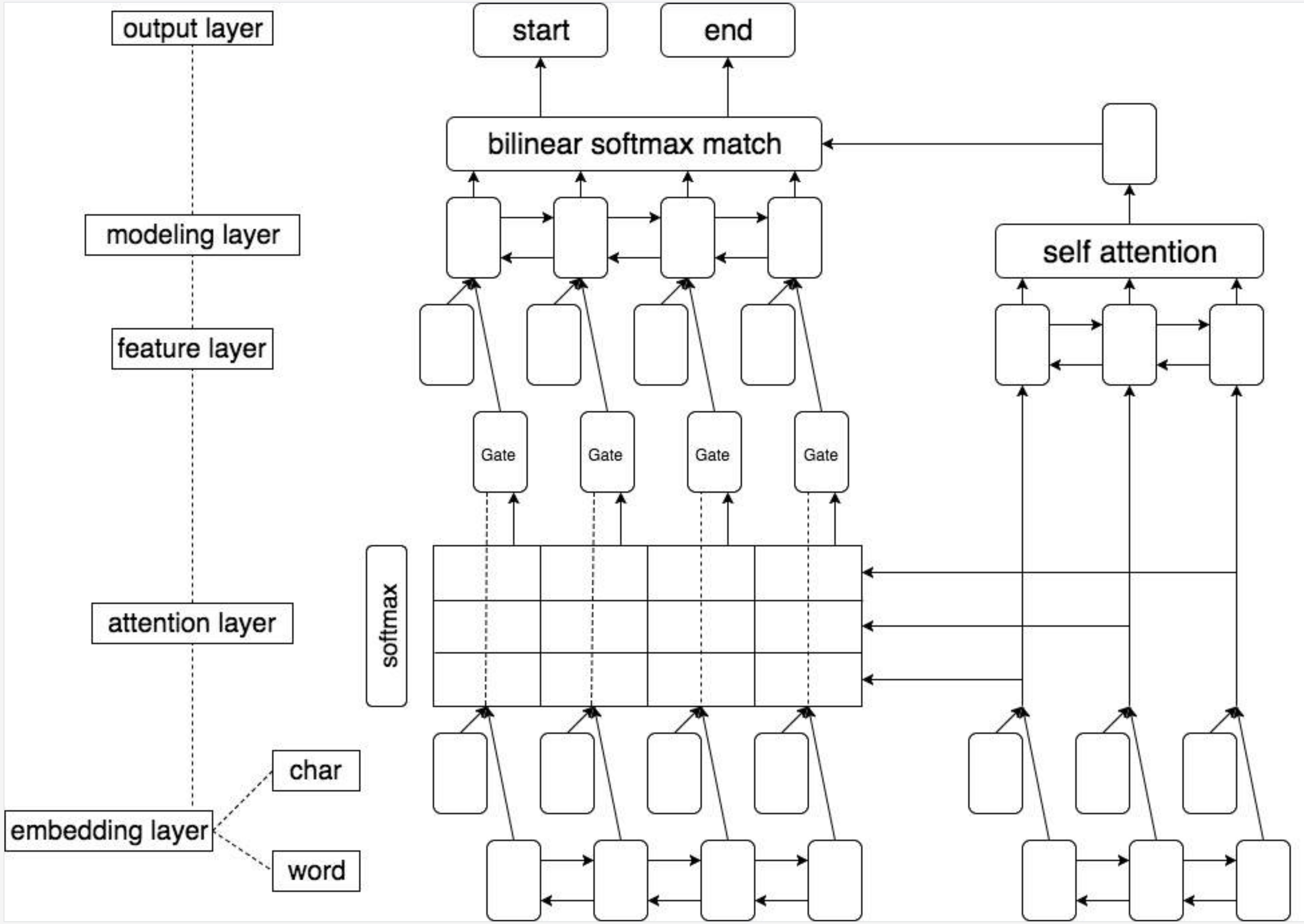


Score-Start	0.07	0.1	0.6	0.15	…	0.03	…	0.01
$\text{Max}(S_s{*}S_e)$	W_1	W_2	W_3	W_4	…	W_k	…	W_n
Score-End	0	0.32	0.2	0.15	…	0.3	…	0.01



业务模型结构

- Embedding Layer
  - ✓ 问题及篇章中词向量表示
  - ✓ RNN网络捕捉语序间依赖
- Attention Layer
  - ✓ 对齐问题和篇章，语义相似性计算
  - ✓ 引进注意力机制，带着问题找答案
- Modeling Layer
  - ✓ Question-Aware篇章建模
  - ✓ 充分利用问题中信息
- Output Layer
  - ✓ 基于问题和篇章匹配预测答案位置

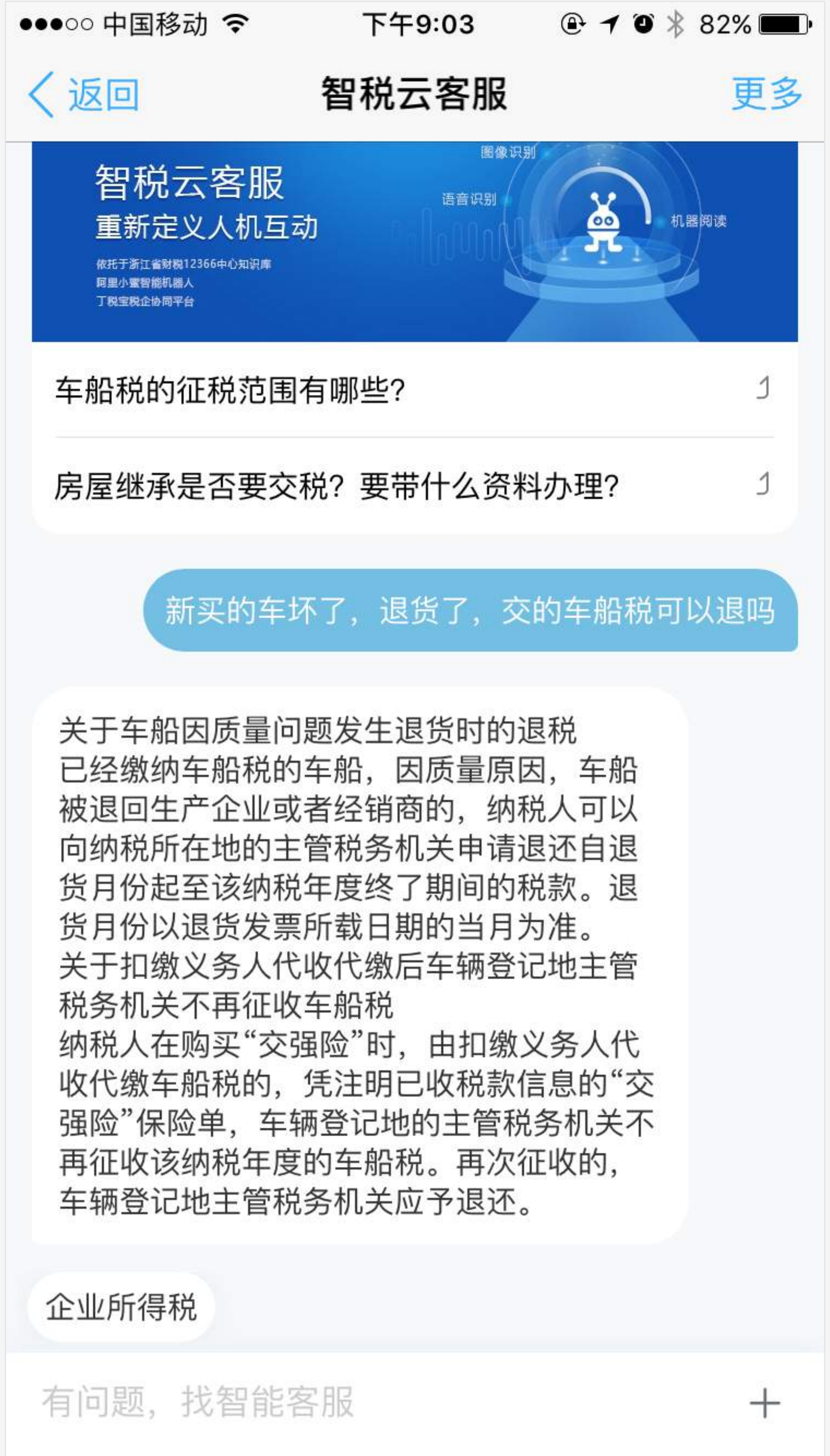


线上模型表现

- 阿里小蜜活动规则场景
  - ✓ 10%+的回答率情况下，准确率90%+
  - ✓ Exact Match Score>78.5%
  - ✓ F1 Score>87.8%
  - ✓ 单次在线服务调用响应时间<70ms

税务法规解读场景

- 企业小蜜税务法规解读
  - ✓ 服务于企业缴税咨询场景
  - ✓ 税法种类多且长
  - ✓ FAQ构建代价非常大
  - ✓ 数据标注成本高



智税云客服

智税云客服
重新定义人机互动

依托于浙江省财税12366中心知识库
阿里小蜜智能机器人
丁税宝税企协同平台

车船税的征税范围有哪些?

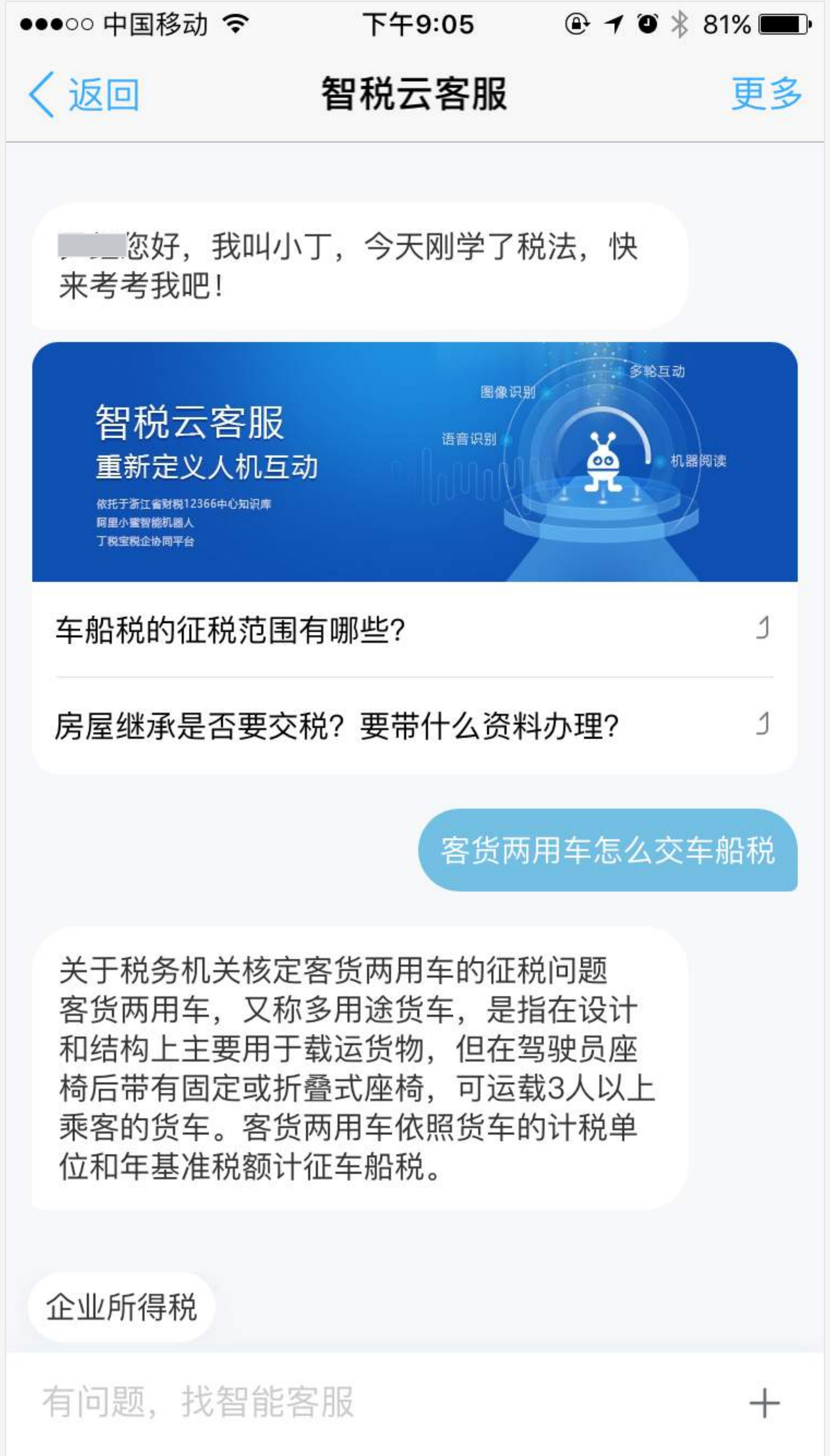
房屋继承是否要交税？要带什么资料办理?

新买的车坏了，退货了，交的车船税可以退吗

关于车船因质量问题发生退货时的退税
已经缴纳车船税的车船，因质量原因，车船被退回生产企业或者经销商的，纳税人可以向纳税所在地的主管税务机关申请退还自退货月份起至该纳税年度终了期间的税款。退货月份以退货发票所载日期的当月为准。
关于扣缴义务人代收代缴后车辆登记地主管税务机关不再征收车船税
纳税人在购买“交强险”时，由扣缴义务人代收代缴车船税的，凭注明已收税款信息的“交强险”保险单，车辆登记地的主管税务机关不再征收该纳税年度的车船税。再次征收的，车辆登记地主管税务机关应予退还。

企业所得税

有问题，找智能客服



智税云客服

[redacted]您好，我叫小丁，今天刚学了税法，快来考考我吧！

智税云客服
重新定义人机互动

依托于浙江省财税12366中心知识库
阿里小蜜智能机器人
丁税宝税企协同平台

车船税的征税范围有哪些?

房屋继承是否要交税？要带什么资料办理?

客货两用车怎么交车船税

关于税务机关核定客货两用车的征税问题
客货两用车，又称多用途货车，是指在设计和结构上主要用于载运货物，但在驾驶员座椅后带有固定或折叠式座椅，可运载3人以上乘客的货车。客货两用车依照货车的计税单位和年基准税额计征车船税。

企业所得税

有问题，找智能客服

模型的场景迁移

- 税法阅读和活动规则阅读存在一定相似性
  - ✓ 都是中文，有共同的语言特性
  - ✓ 都是规则类文本，对于答案定位有一定的共性
  - ✓ 答案粒度类似，都是以句为粒度
  - ✓ 持续学习：将模型过去学到的知识运用在新的学习场景中

活动规则 → 税务法规 → ……

模型的场景迁移

- 迁移学习的应用
  - ✓ 分层特征表达
  - ✓ 随网络层加深，面向特定任务
  - ✓ 低级语义 -->高级语义
  - ✓ 低级语义特征复用性强
  - ✓ 高级语义特征场景化强

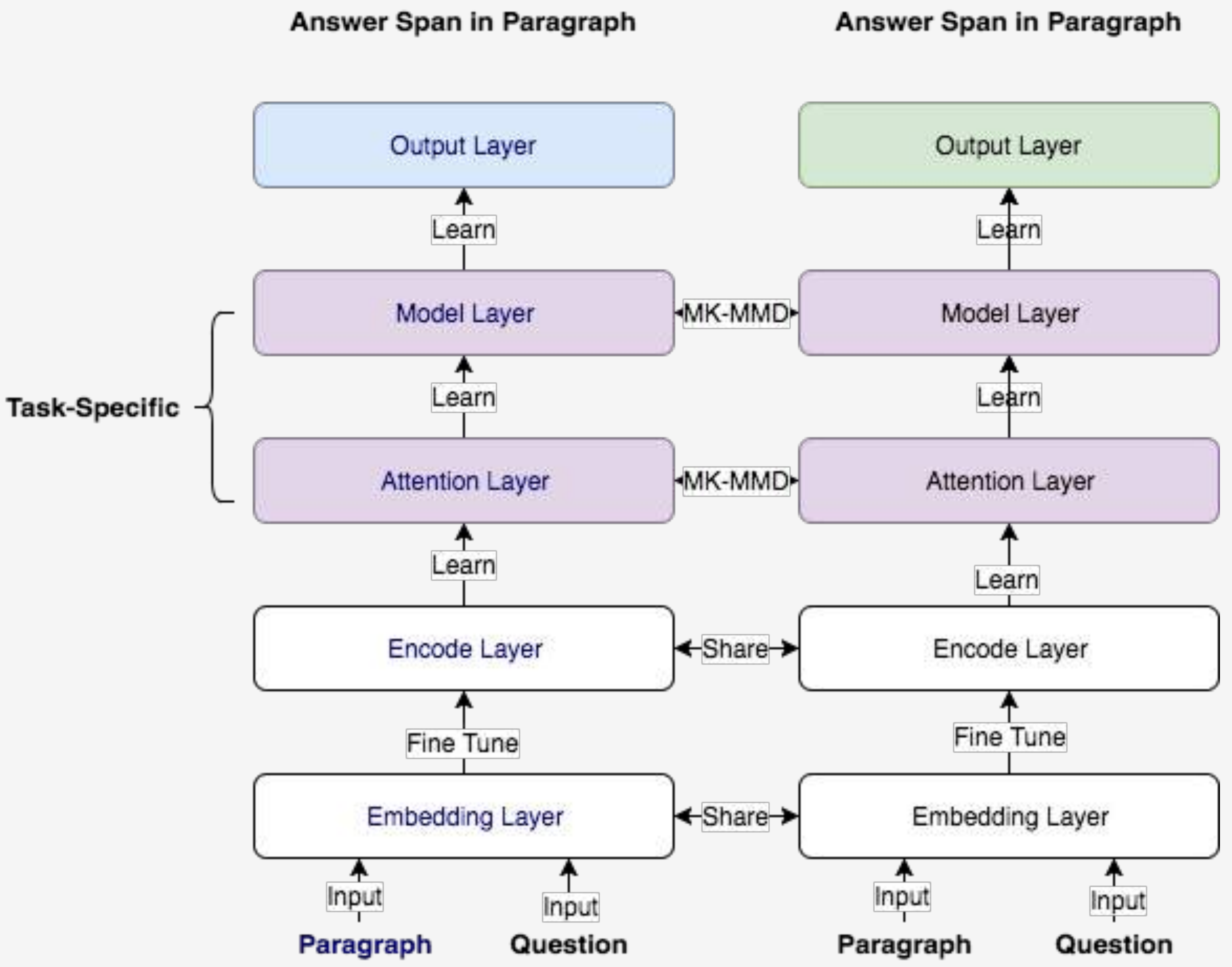


TABLE OF
CONTENTS 大纲

- 阿里小蜜平台介绍
- 机器阅读理解技术概览
- 业务场景及技术实践
- 挑战与展望

技术挑战

- Pretend to Understand
  - ✓ 推理总结问题，如How类问题
  - ✓ 干扰性问题或者文本中干扰性文字
- 知识的运用
  - ✓ CommonSense Knowledge
  - ✓ 业务知识的融入

- 线上服务性能
  - ✓ 模型复杂导致计算量过大
  - ✓ 预计算模型设计
- 情感化
  - ✓ 目前回答较为生硬
  - ✓ 结合生成技术使答案更个性化和情感化

经验总结

- 数据标注时遵循真实的数据分布
- 数据回流形成闭环
- 数据质量比模型本身更重要
- 模型性能比准确率更值得关注
- 让模型不断积累过去学到过的知识，而不是每次重新训练

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit
全球架构师峰会 2017