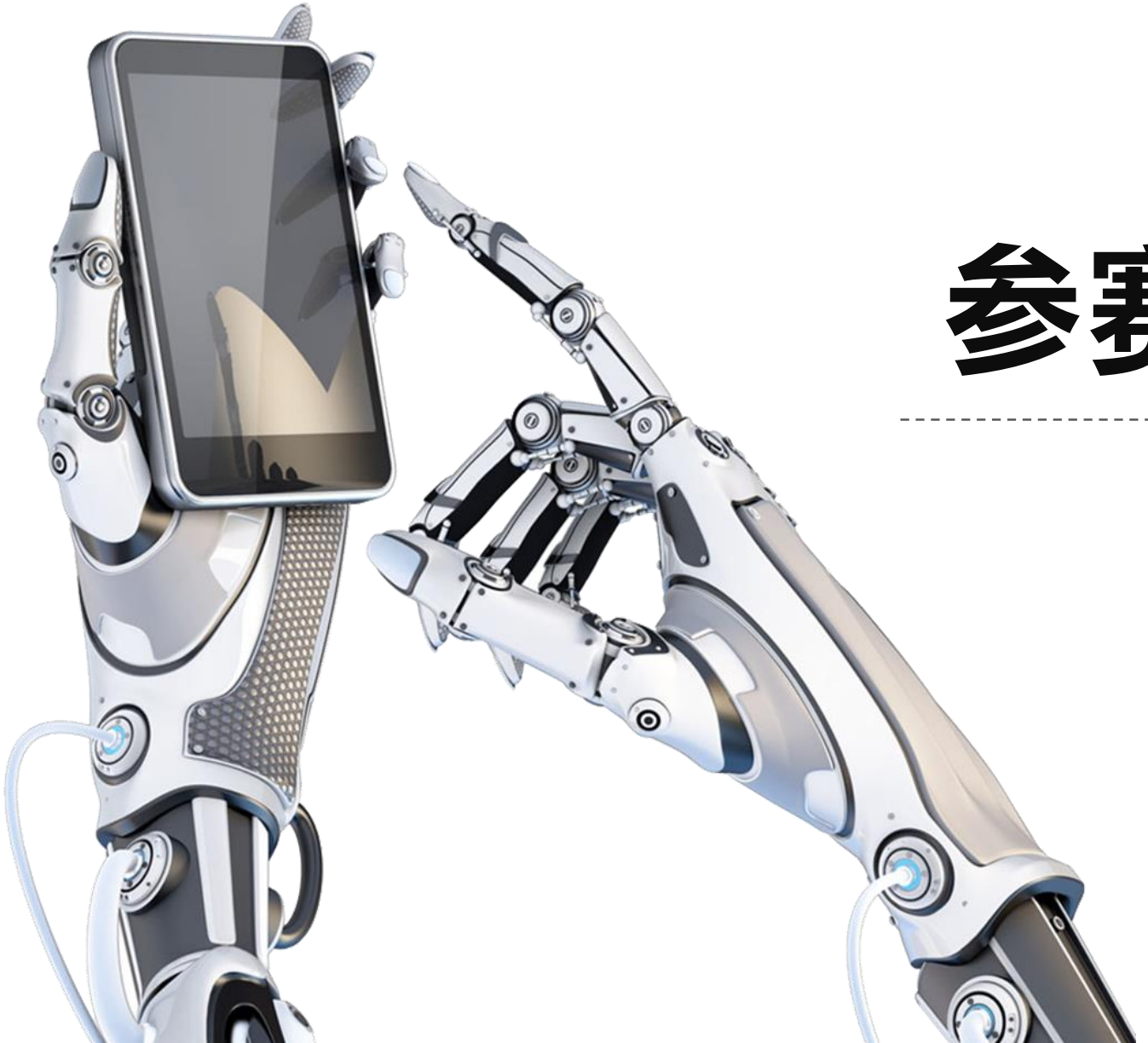


FunNLP
参赛队伍技术报告

报告人：欧阳洋

参赛人员：欧阳洋，张静，游超斌，张弘洲，张烁，孟宪森，方省，王宇琪

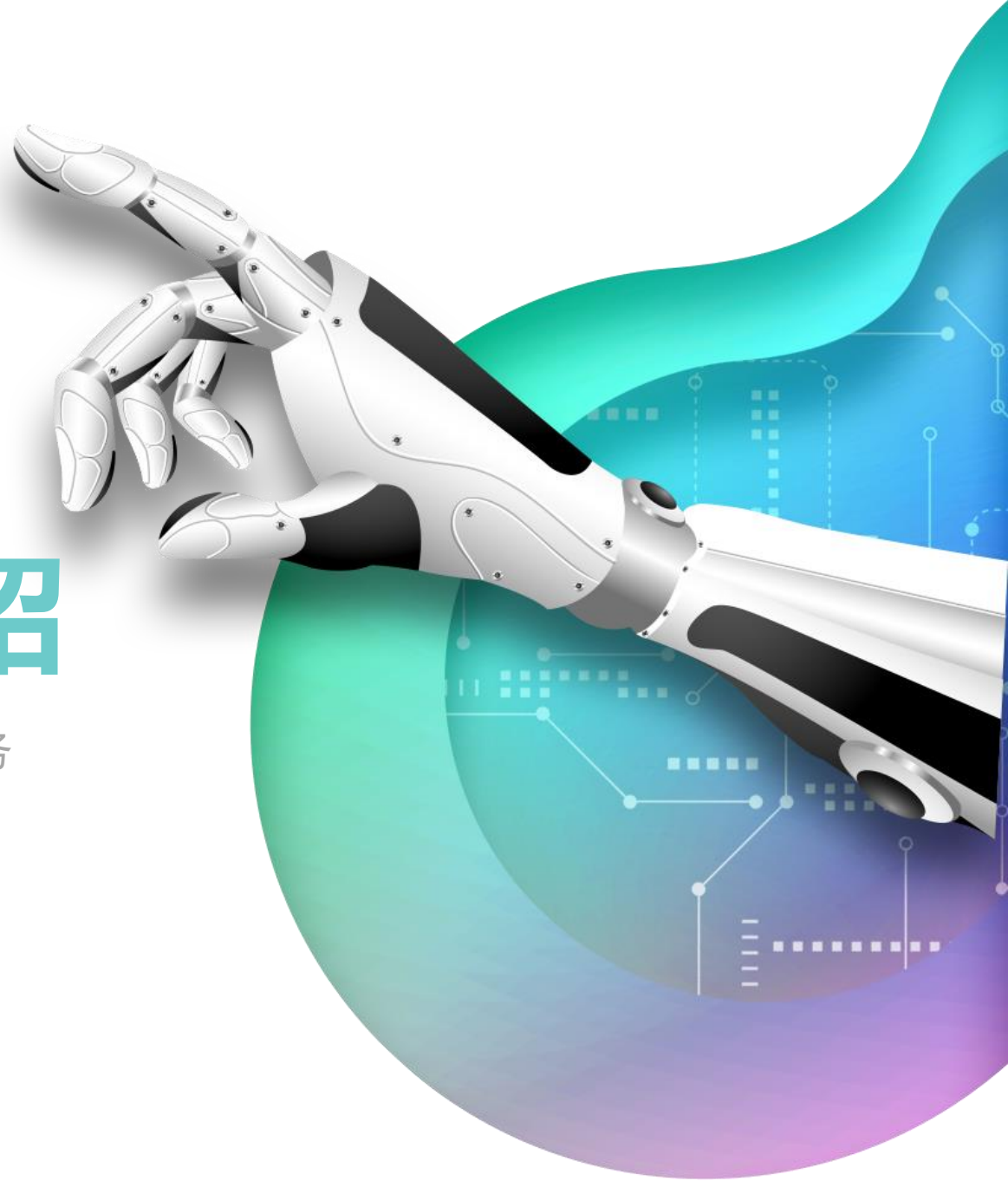
指导老师：张鹏，孔庆超

目录

CONTENTS

01 比赛介绍

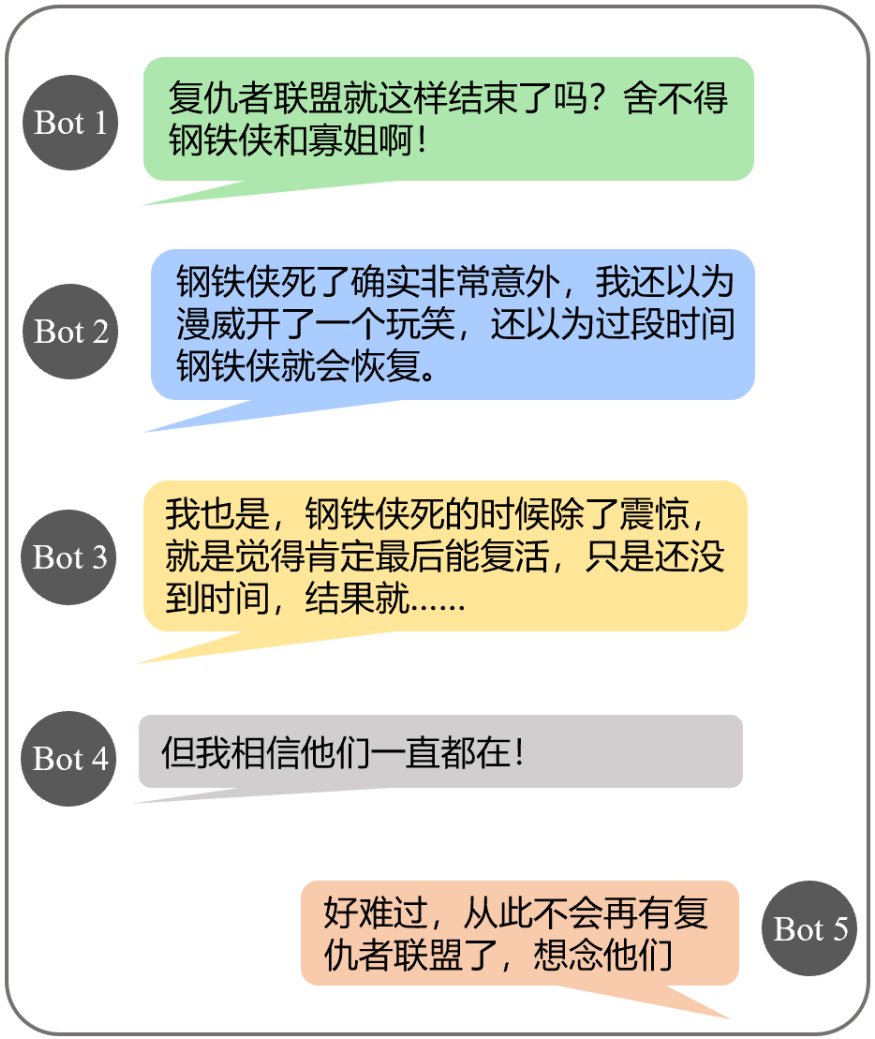
- 比赛意义
- 比赛任务
- 比赛难点



比赛意义

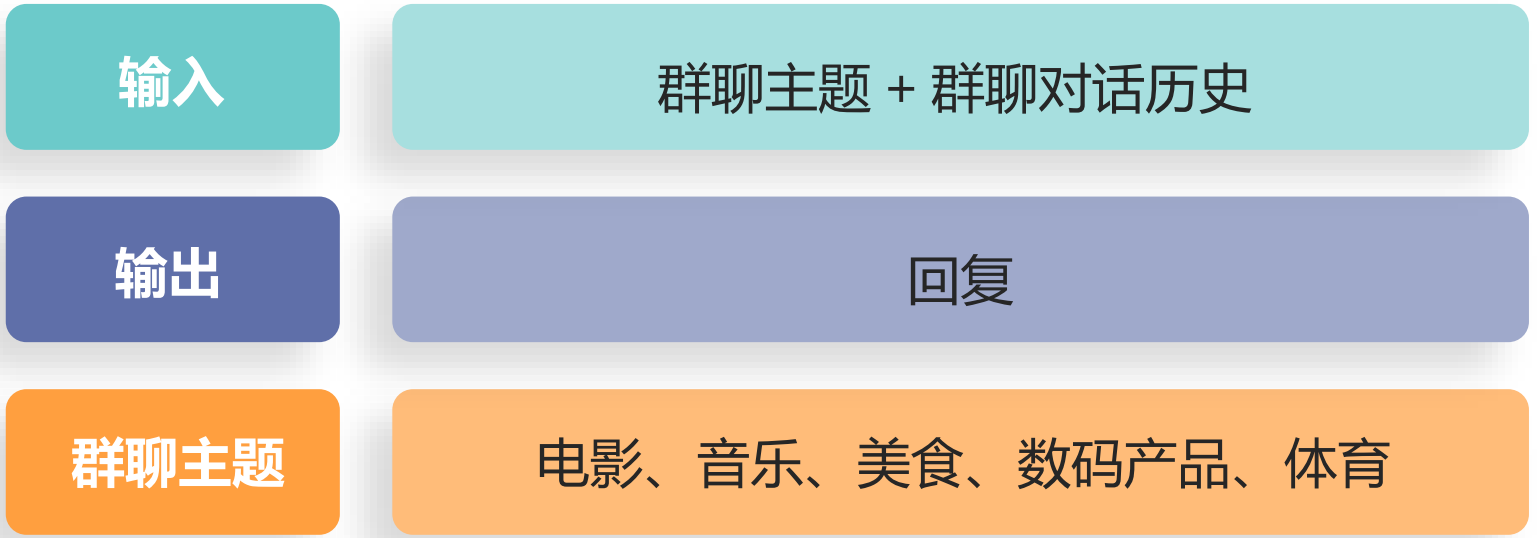
目前的人机对话仅限于人机双方的对话模式，而更具挑战的**人机多方混合对话**乃至**机器人多方群聊**的任务在研究和应用上鲜有涉及。

- 促进**人机对话技术在多方对话场景上的发展**
- 充分挖掘人机对话技术**在各种应用场景下的潜力**
- 为人机对话技术相关的学术界研究人员和产业界从业人员提供一个良好的沟通平台。



比赛任务

在机器人群聊场景中，已知**群聊主题**和**历史消息记录**，要求生成符合群聊主题和上下文逻辑的**回复**。所生成的回复需要**流畅且与群聊对话主题相关**。



比赛难点

1

高质量中文主题闲聊型多轮数据集的收集

2

对于对话历史的理解能力，包括对主题的理解和其他机器人回复的理解

3

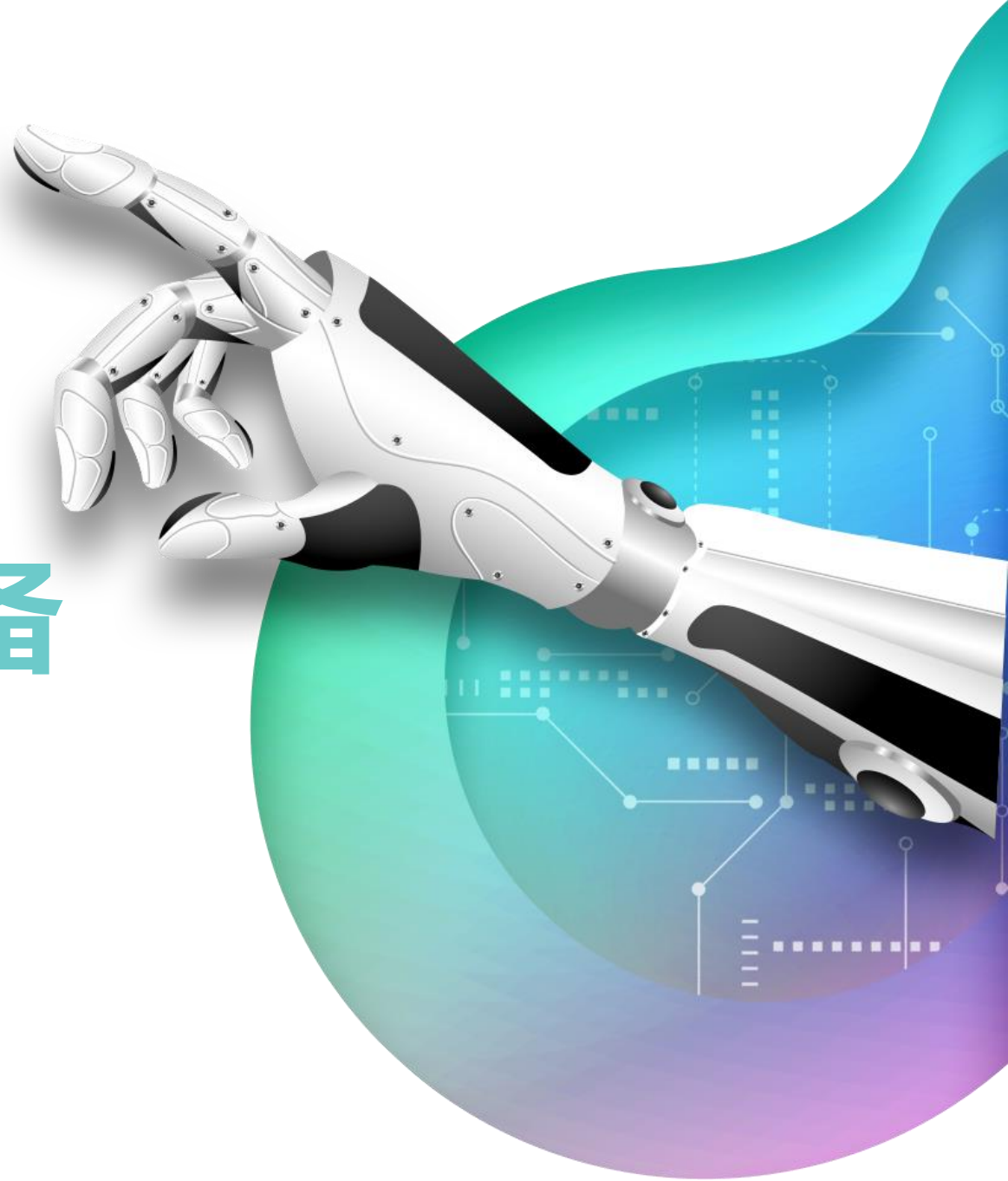
群聊历史记录的关键信息捕捉

4

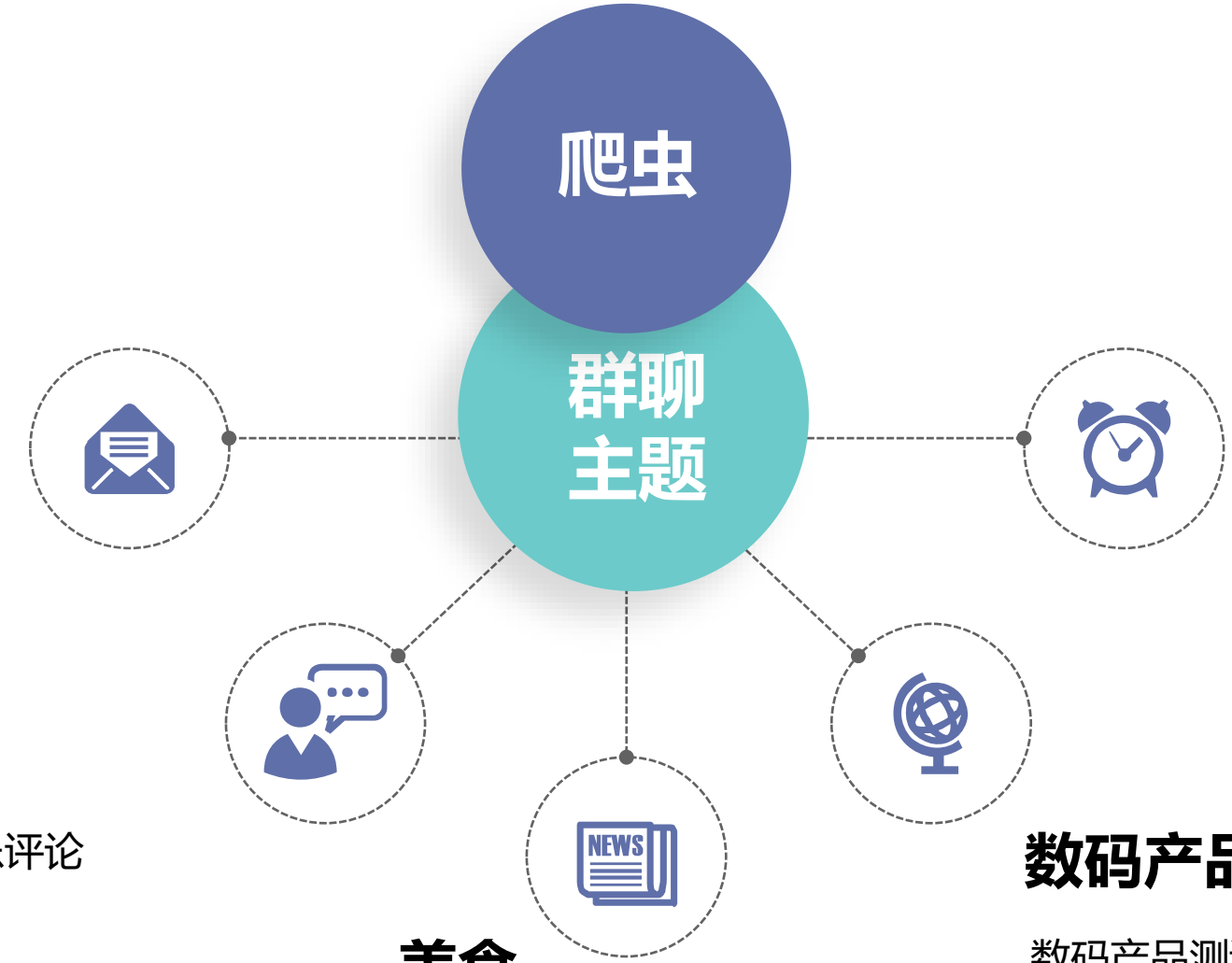
最优的回复对象的选择

02 数据准备

- 数据集准备
- 数据处理



数据集准备



电影

豆瓣电影评论

音乐

豆瓣音乐评论

美食

好豆网：美食介绍文章

体育

虎扑论坛，相关体育运动的帖子

数码产品

数码产品测评文章（如锋潮科技、爱搞机网站

数据处理

1

数据清洗

清洗掉字数比较少、与主题无关以及一些低质量的评论

2

子主题分类

将每个群聊主题划分成更小的主题

3

构造多轮数据

在子主题下利用余弦相似度构造多轮对话

4

Embedding层

在Embedding层采用BERT预训练模型生成句向量

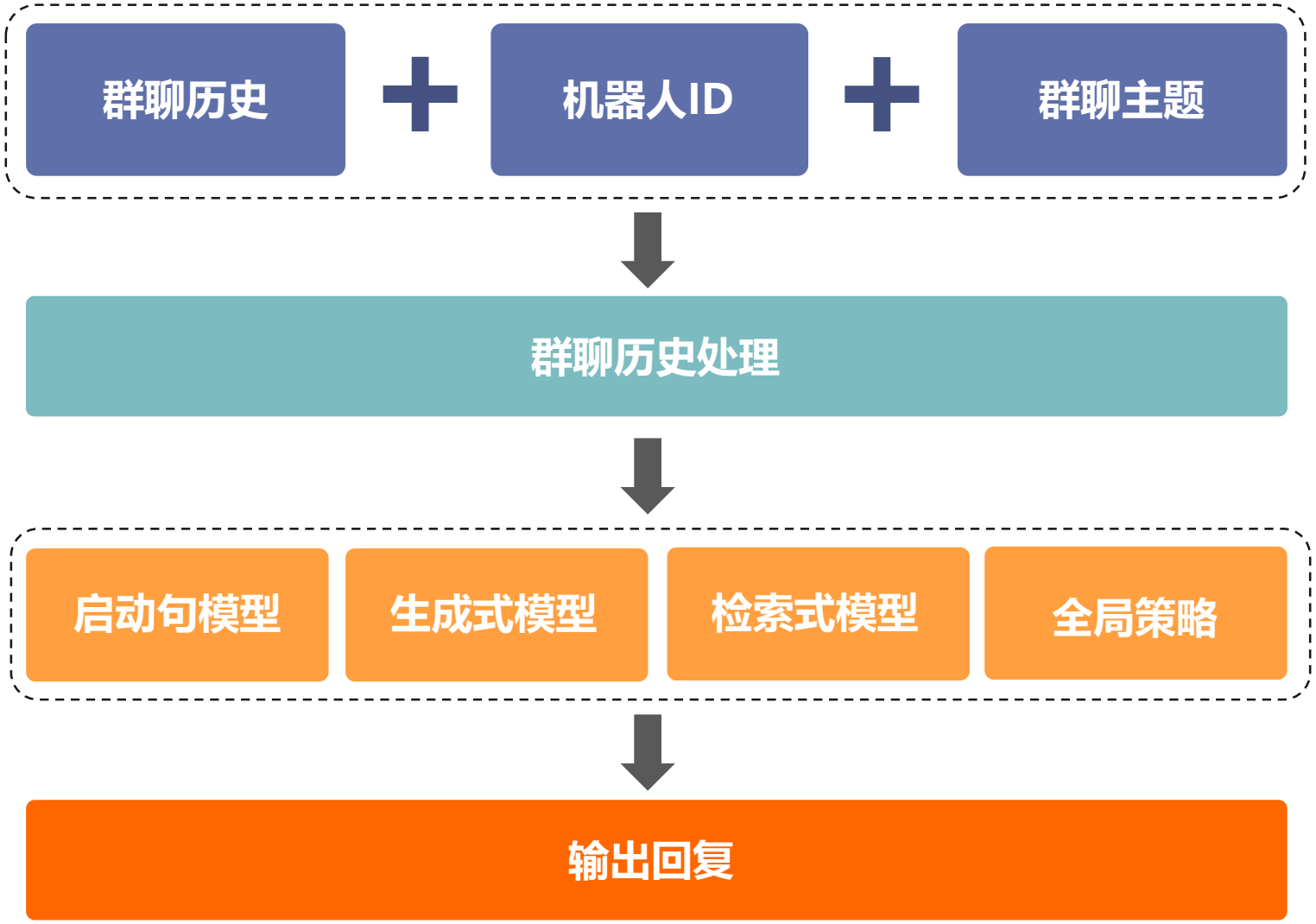
数据集格式

[Utterance 1] [Tab] [Utterance 2] [Tab] [utterance 3] [Tab] … … [Tab] [Response]

03 比赛思路

- 系统框架图
- 启动句模型
- 检索式模型
- 对话历史处理
- 生成式模型
- 全局策略

系统框架图



对话历史处理

过滤对话历史

利用初赛的**评价指标Topic**计算题主题相关度），设置阈值，过滤掉与主题无关的话语。

$$Similarity = \frac{topic \cdot history}{||topic|| \times ||history||}$$

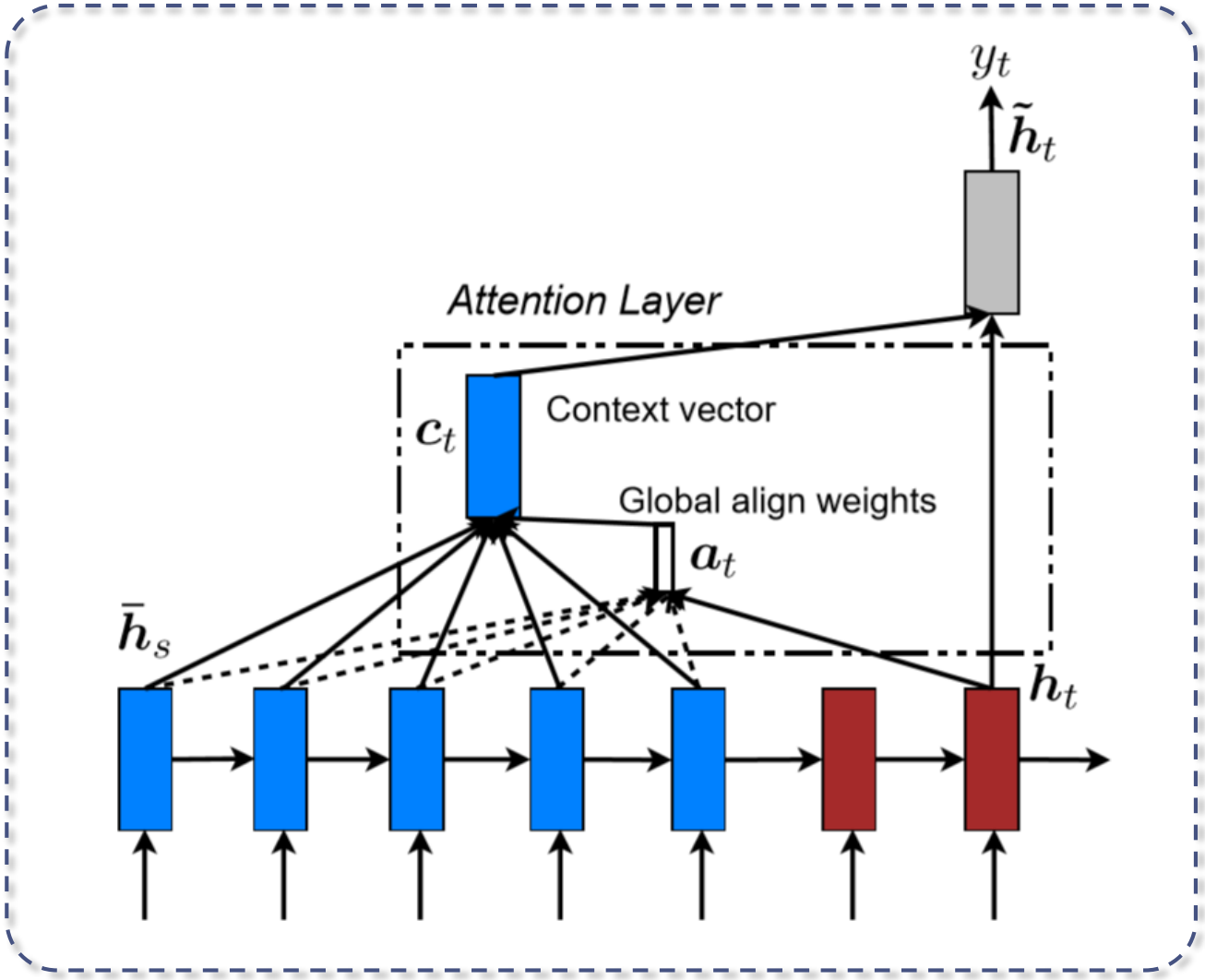
topic：代表主题词向量

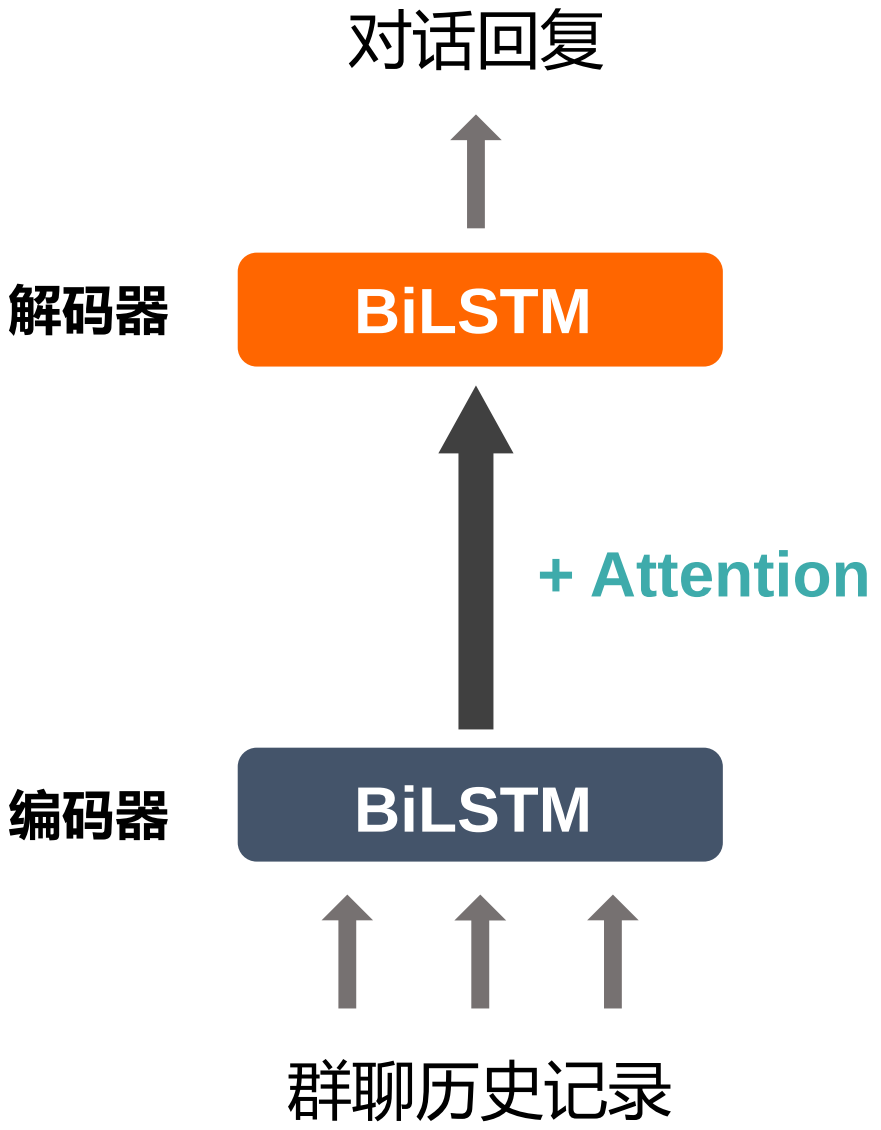
history：代表某一条群聊对话历史的句向量

启动句模型


启动句模型
启用条件
机器人ID
靠前
处理后的
对话历史
为空
启动句问题
解决方案
人工构造
单轮数据集
Seq2Seq
+Attention

生成式模型


y_t
$\tilde{h}_t$
Attention Layer
c_t
Context vector
Global align weights
a_t
$\bar{h}_s$
h_t


对话回复
解码器
BiLSTM
+ Attention
编码器
BiLSTM
群聊历史记录

检索式模型

$$Score(Q,d)=\sum_{i}^{n} W_i \cdot R(q_i,d)$$

$$W_i = log\frac{N-n(q_i)+0.5}{n(q_i)+0.5}$$

注：

Q：表示一句话，这里就代表群聊历史记录

d：表示待搜索的句子，这里代表我们的语料库

qi：表示Q中的每个词

N：表示语料库的大小

n(qi)：表示语料库中包含qi的句子数

BM25模型

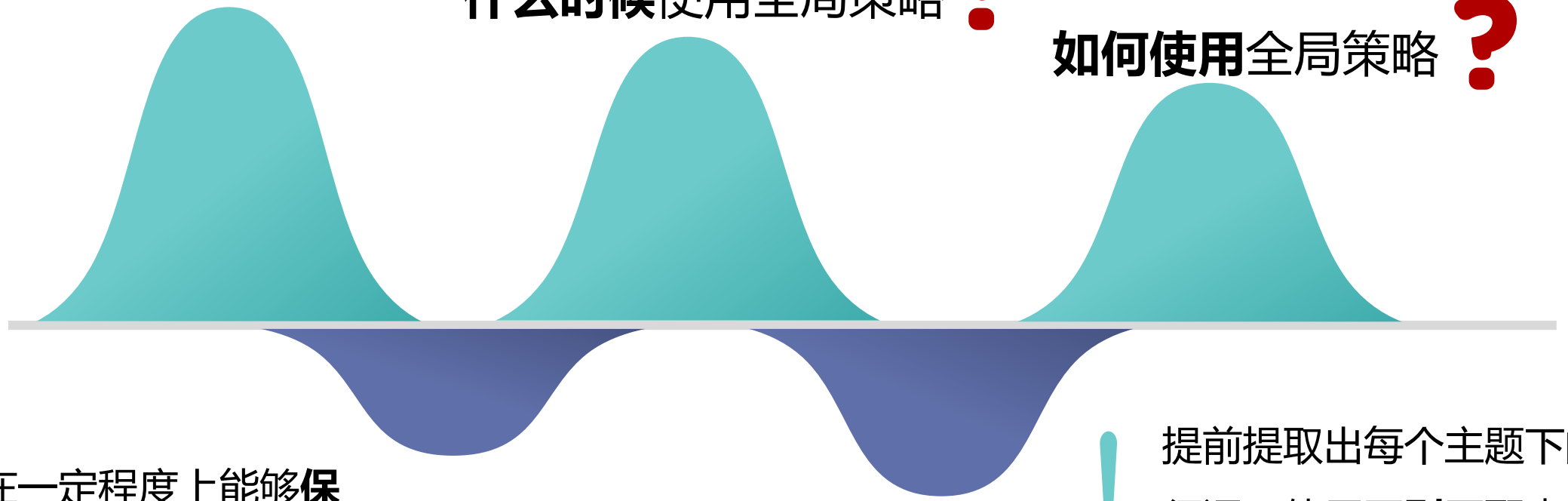
- 是一种**文本相似度**度量的方法，在信息检索领域有着非常广泛的应用。
- **主要思想**：计算某句话中的每个词与待搜索句子的相关性得分，然后利用IDF值给予每个词相应的权重，最后将每个词与待搜索句子的相关性得分进行加权求和，从而得到该句话与待搜索句子的得分。
- 在检索式模型中，选择**得分最高**的句子作为检索模型的回复结果。

全局策略

为什么要使用全局策略？

什么时候使用全局策略？

如何使用全局策略？



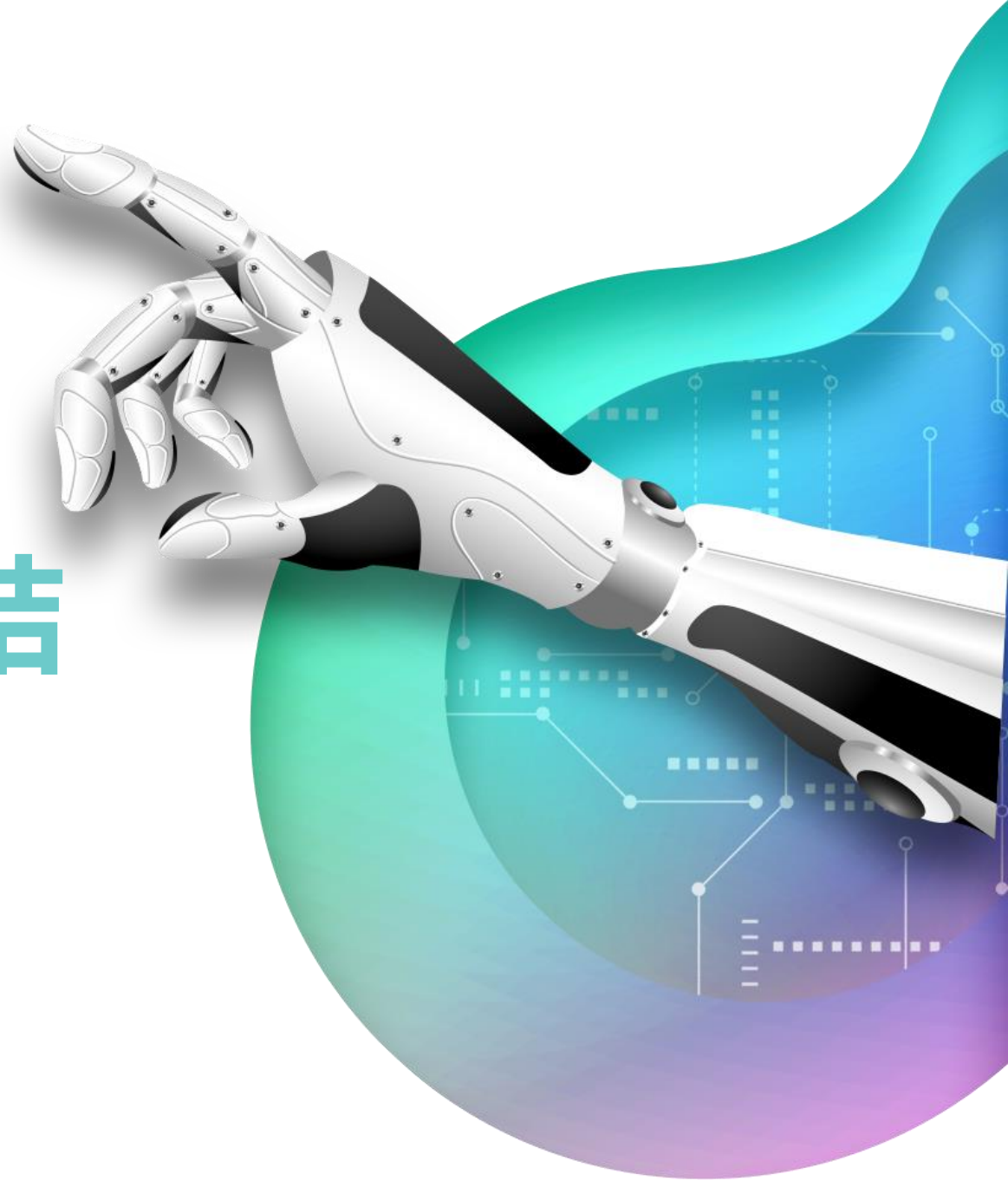
在一定程度上能够**保证对话的完成度**

当生成式模型和检索式模型都**不能给出合理的回复**时，便采用全局策略

提前提取出每个主题下的特征词，使用**正则匹配**来对群聊对话历史进行划分，而后直接给出人工设置的回复。

04 思考总结

- 比赛总结

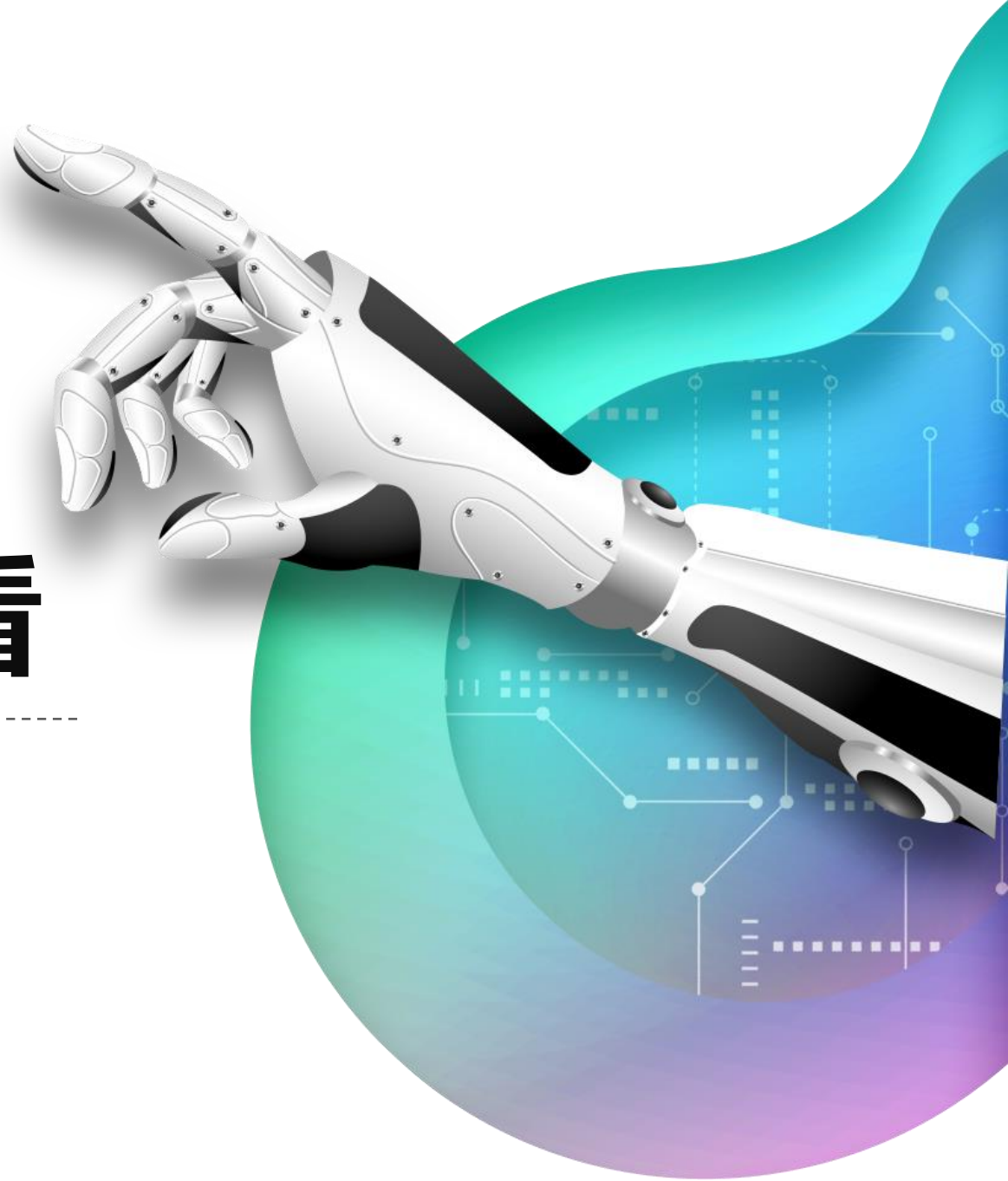


比赛总结

- **优质数据集**的爬取与收集
- 采用**BERT预训练模型**
- 综合了**生成式**、**检索式**模型
- 在机器人群聊中有不错的表现
- 团队在自然语言处理领域有着不错的**技术积累**，和厚实的**知识储备**
- 尝试在模型中加入**知识图谱**，提高预主题的相关度，提升系统回复质量
- 提升机器人在情感上回复的一致性

THANKS

非常感谢您的观看



报告人：欧阳洋

2019.12.14