

面向问答的语言模型 信息对齐机制研究

汇报人：罗志一

浙江理工大学 计算机科学与技术学院（人工智能学院）

基本信息（摘要、关键词、申请代码）

摘要：问答是NLP领域的重要研究任务。传统方案受限于模型的语义理解与信息感知能力不足，难以准确回答用户提问。语言模型以其强大的信息存储和感知能力，成为突破当前问答研究局限的重要手段。然而，由于语言模型与人类问答需求间存在信息不对齐的内在特性，回答的安全性和真实性饱受争议，必须设计有效的信息对齐方案以满足用户问答的诉求。本项目拟根据语言模型的不同演化阶段，递进式设计：1) **预训练目标适配机制**，增强模型对问答任务的自适应性、提高模型参数效率；2) **基于动态混合提示精调的意图对齐机制**，采用梯度搜索和梯度下降的方法学习动态混合提示，为不同问题生成定制提示内容；3) **基于结构化知识注入的知识对齐机制**，通过编码异构结构化知识，增强模型对事实知识的感知。最终，本项目将形成一套面向问答的语言模型信息对齐机制，有力提升模型表现性能。本研究对语言模型发展和问答系统构建意义重大，具有重要的科学意义和广泛的应用价值。

关键词：问答；语言模型；提示精调；知识注入；信息对齐

申请代码：**F0211**. 信息检索与社会计算

科学问题属性（“聚焦前沿，独辟蹊径”）

问答是NLP领域的重要研究任务。传统方案受限于模型的语义理解与信息感知能力不足，难以准确回答用户提问。语言模型以其强大的信息存储和感知能力，成为突破当前问答研究局限的重要手段。然而，由于语言模型与人类问答需求间存在信息不对齐的内在特性，回答的安全性和真实性饱受争议，必须设计有效的信息对齐方案以满足用户问答的诉求。

本项目立足于解决面向问答的语言模型中信息对齐关键科学问题，形成针对语言模型在预训练、提示精调和微调的不同阶段，对齐人类问答需求的系统解决方案。在考虑模型性能和算力资源的情况下，申请者针对目标对齐、意图对齐和知识对齐三个方面开展更具创新性的研究工作：(1)提出了基于预训练目标适配的目标对齐方案。基于掩码内容粒度、掩码率的动态变化策略与预训练目标的适配，提升模型的收敛速度和收敛效果。这种方案优化了模型的预训练目标，提高了模型参数效率。(2)提出了基于动态混合提示精调的意图对齐方案。基于连续提示和离散提示的混合提示问答模板可以为不同输入定制提示内容，从而使模型更好地理解问题背后的用户意图，以获取更准确、符合用户需求的回答。这是一种新颖的提示精调方案。(3)提出了基于结构化知识注入的知识对齐方案。在知识结构中引入混合提示得到异构结构化知识，丰富语言模型对事实知识的感知，确保回答的真实性。此方案还可应用在领域知识图谱中以实现模型的专业化，具有很好的可扩展性。

最终，本项目将形成一套独创的、新颖的、面向问答的语言模型信息对齐机制，主要包括基于预训练目标适配的目标对齐机制、基于动态混合提示精调的意图对齐机制和基于结构化知识注入的知识对齐机制等一系列研究方案，有力提升语言模型在问答任务中的表现性能。该研究对语言模型发展和问答系统构建意义重大，具有重要的科学意义和广泛的应用价值。

研究动机

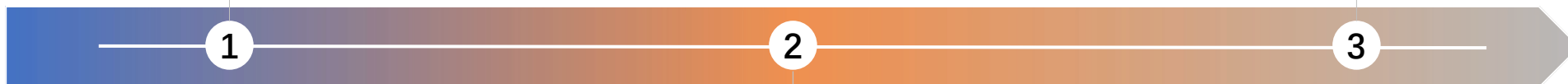
- 问答任务期望机器理解并回答自然语言问题，以使得人类以自然的方式（即提问）与机器进行交互，从机器存储、感知的海量信息中快速、准确地获取想要的信息（即问题的答案）。
- 问答系统的构建范式与信息在机器中的承载、组织形式密切相关。

基于信息检索的问答范式

例：搜索引擎、检索器-阅读器

基于语言模型的问答范式

例：BERT、GPT系列、ChatGPT等



信息主要以非结构化和结构化的形式存储与机器之中。

基于知识的问答范式

例：KBQA

构建基于语言模型的新信息存储与感知机制。

研究动机

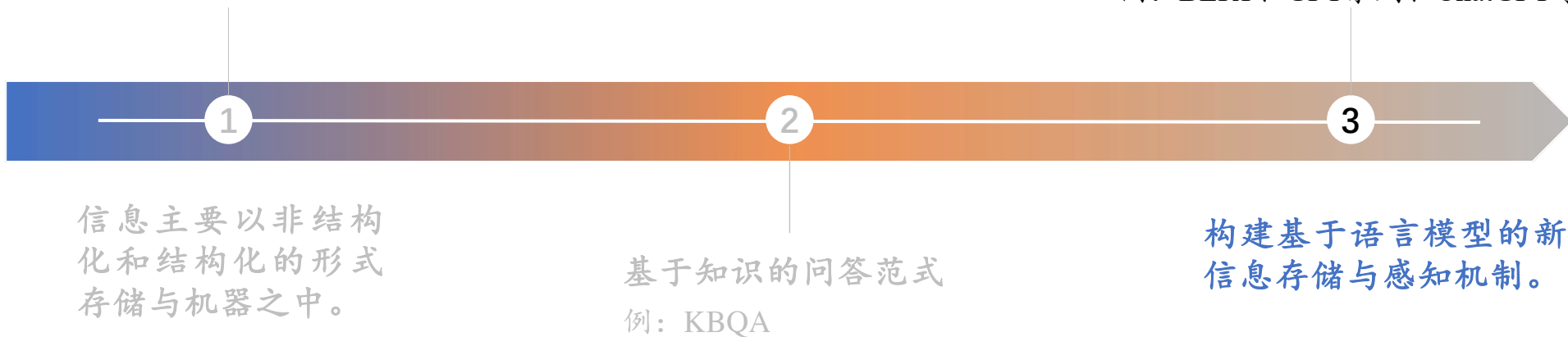
- 问答任务期望机器理解并回答自然语言问题，以使得人类以自然的方式（即提问）与机器进行交互，从机器存储、感知的海量信息中快速、准确地获取想要的信息（即问题的答案）。
- 问答系统的构建范式与信息在机器中的承载、组织形式密切相关。
- 语言模型经过不断的演化，已经具备了理解和生成自然语言的能力。然而，由于语言模型与人类问答需求间存在信息不对齐的内在特性，基于语言模型的问答研究仍然面临诸多挑战。

基于信息检索的问答范式

例：搜索引擎、检索器-阅读器

基于语言模型的问答范式

例：BERT、GPT系列、ChatGPT等



研究动机 (续)

2023年2月间的真实交互实例



| | | | |
|------|------------------|-------|---------------------|
| 中文名 | 蒙特雷斯·哈雷尔 | 毕业院校 | 路易斯维尔大学 |
| 外文名 | Montrezl Harrell | 运动项目 | 篮球 |
| 国籍 | 美国 | 所属运动队 | 费城76人队 [56-57] |
| 出生地 | 美国北卡罗来纳州塔珀伯勒 | 专业特点 | 得分能力强 |
| 出生日期 | 1994年1月26日 | 主要奖项 | 2019-20赛季最佳第六人 [3] |
| 身高 | 2.01 m [4] | | 2019-20赛季最佳拼搏奖 [25] |
| 体重 | 108.9 公斤 | 球衣号码 | 5号 [45] [58] |
| | | NBA选秀 | 2015年次轮第32位被火箭队选中 |

69 哪个nba球星是8月25日的生日?

8月25日出生的NBA球星是大前锋、中锋，目前效力于洛杉矶快船队的蒙特雷兹·哈雷尔 (Montrezl Harrell)。他于2015年的NBA选秀中第32顺位被休斯顿火箭队选中，之后曾效力于洛杉矶湖人队等多支球队。他在场上以能量十足的突破和强力扣篮而著称，是一位备受球迷喜爱的球员。

图 1: ChatGPT 生成的回答违反事实



69 罗志一的名作《群虾》的创作过程

《群虾》是中国著名作家罗志一创作的小说，是他代表作之一。据罗志一在接受采访时透露，《群虾》是在1986年开始构思的，历经四年完成，最终于1990年出版。

罗志一在创作过程中，进行了大量的实地调研和采访，深入了解了渔村的生活和人物的性格特点。他还研究了相关的历史文献和地理资料，为小说的背景和情节提供了真实的历史和地理依据。在构思和写作的过程中，罗志一一直保持着高度的创作热情和耐心，不断打磨和完善小说的情节和人物形象，直到最终完成了这部优秀的作品。

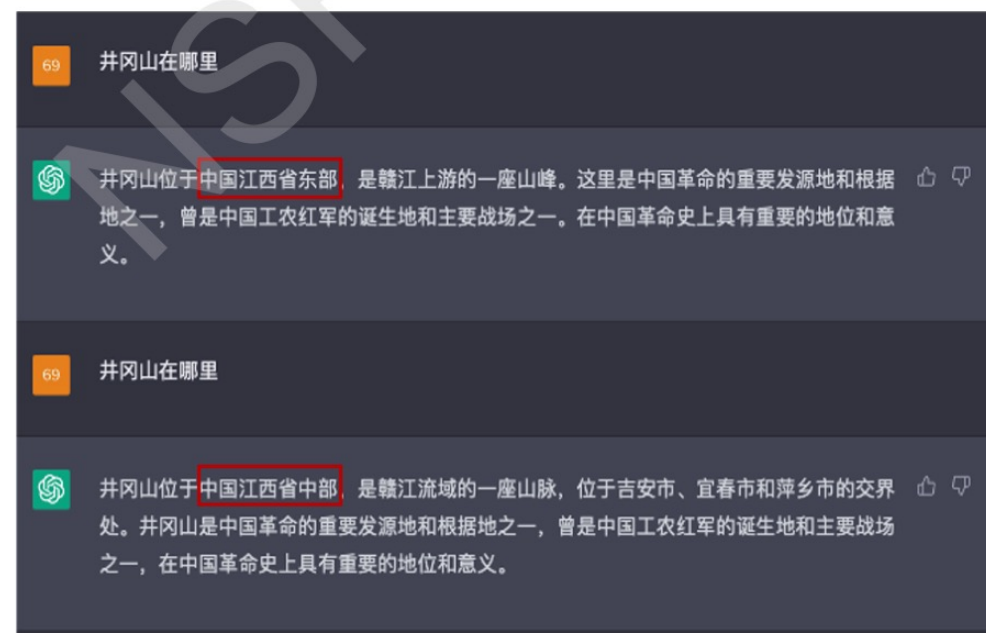
69 齐白石的作品《群虾》的创作过程

请您再次确认，齐白石的作品名是否为《群虾》？因为《群虾》是罗志一的作品，齐白石的作品中并没有这个名字。

| | | | |
|-----|-----|------|--------------|
| 中文名 | 群虾 | 规格 | 100.8×33.9cm |
| 作者 | 齐白石 | 材质 | 纸本 |
| | | 画作类型 | 水墨画 |



图 3: ChatGPT 生成的回答易被篡改



69 井冈山在哪里

井冈山位于中国江西省东部，是赣江上游的一座山峰。这里是中国革命的重要发源地和根据地之一，曾是中国工农红军的诞生地和主要战场之一。在中国革命史上具有重要的地位和意义。

69 井冈山在哪里

井冈山位于中国江西省中部，是赣江流域的一座山脉，位于吉安市、宜春市和萍乡市的交界处。井冈山是中国革命的重要发源地和根据地之一，曾是中国工农红军的诞生地和主要战场之一，在中国革命史上具有重要的地位和意义。

图 2: ChatGPT 生成的回答前后矛盾

研究动机（再续）

挑战：①语言模型的学习目标与生成安全真实回答的诉求不一致；
②语言模型的感知信息模式与用户意图的表达之间有差距；
③语言模型感知的信息内容与事实知识不匹配。

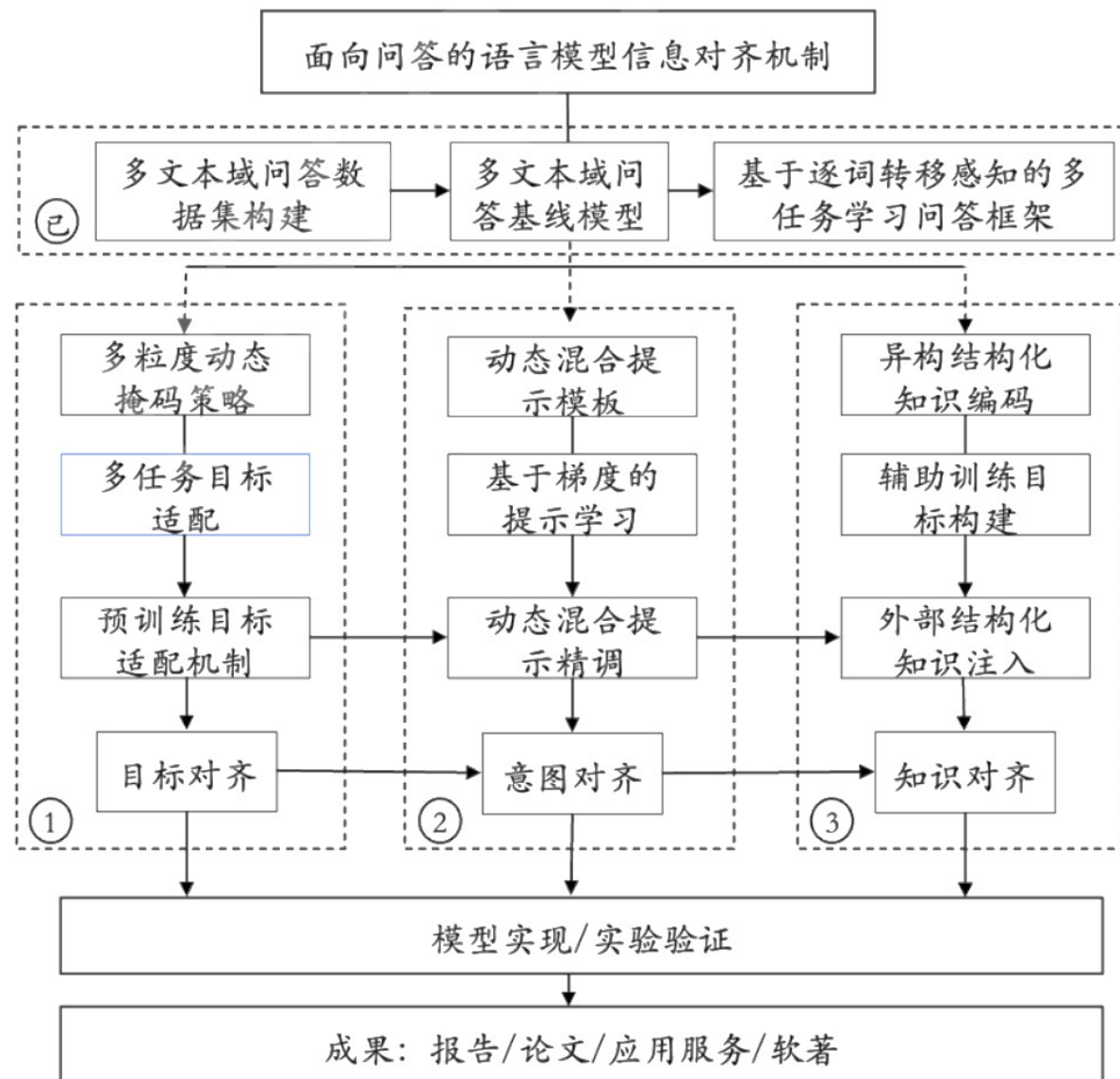
尽管大模型能够生成流畅、自然的语言，但是其生成内容的真实性和安全性难以得到保证，经常会出现编造的、与事实不符、前后矛盾不合理甚至是易被篡改的回答。因而，**语言模型的信息对齐机制**成为问答技术中亟需研究的重要课题。

本项目拟从以下三个方面入手研究面向问答任务的语言模型信息对齐机制。

- ①**目标对齐**：研究语言模型学习目标与生成安全真实回答诉求对齐；
- ②**意图对齐**：研究语言模型感知信息模式与用户意图对齐；
- ③**知识对齐**：研究语言模型感知信息内容与事实知识对齐。

研究内容

本项目立足于解决面向问答的语言模型中信息对齐关键科学问题，形成针对语言模型在预训练、提示精调和微调的不同阶段，对齐人类问答需求的系统解决方案，收集大规模自监督语料、构建高质量问答数据集及实验验证平台，并评估对齐模型的整体性能，最终为面向问答任务的实际应用提供理论支撑与优化技术。本项目的研究课题是对前期工作的深入和延续。

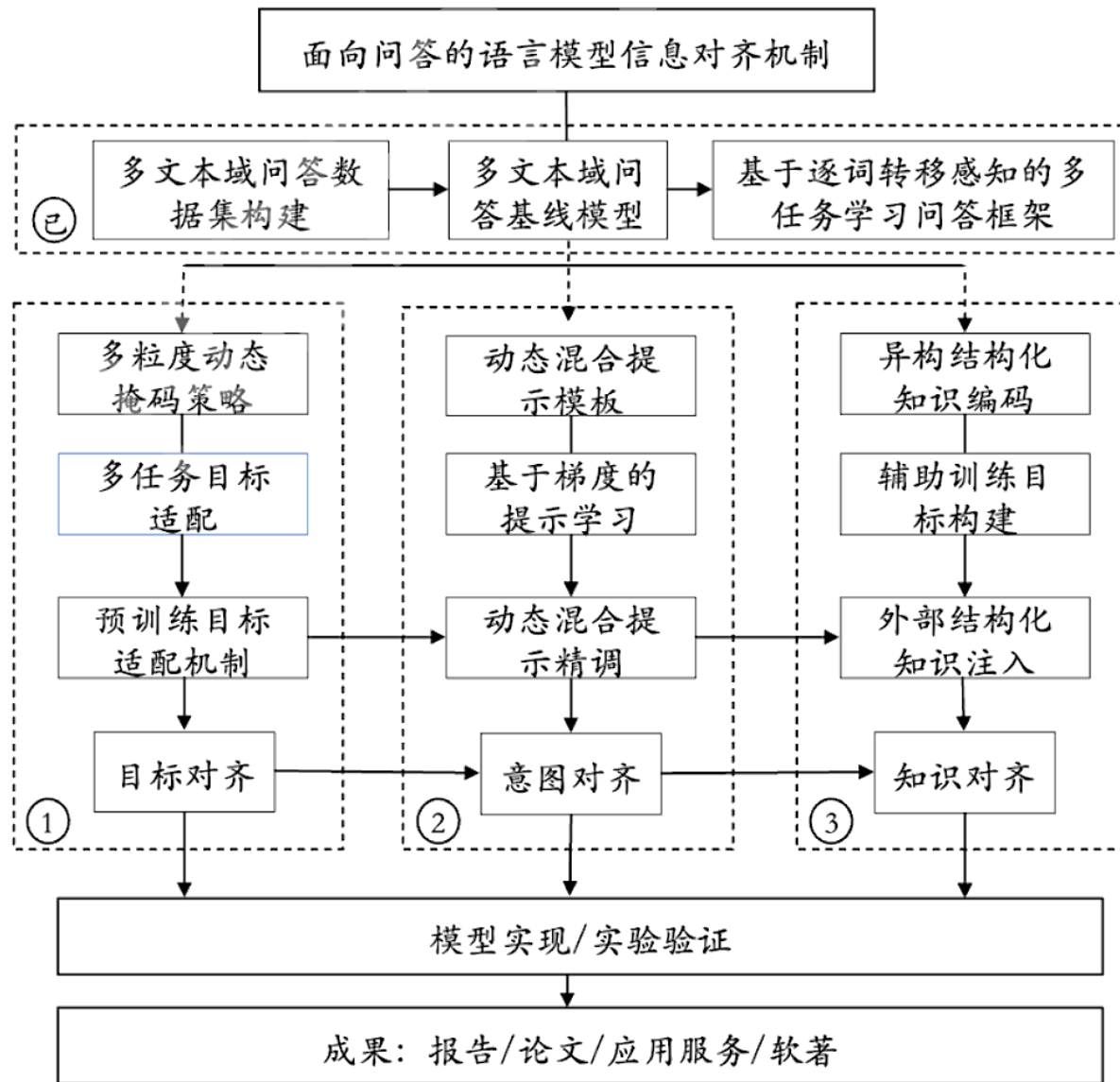


本项目的研究脉络图

研究内容 ①

基于预训练目标适配的目标对齐机制研究：

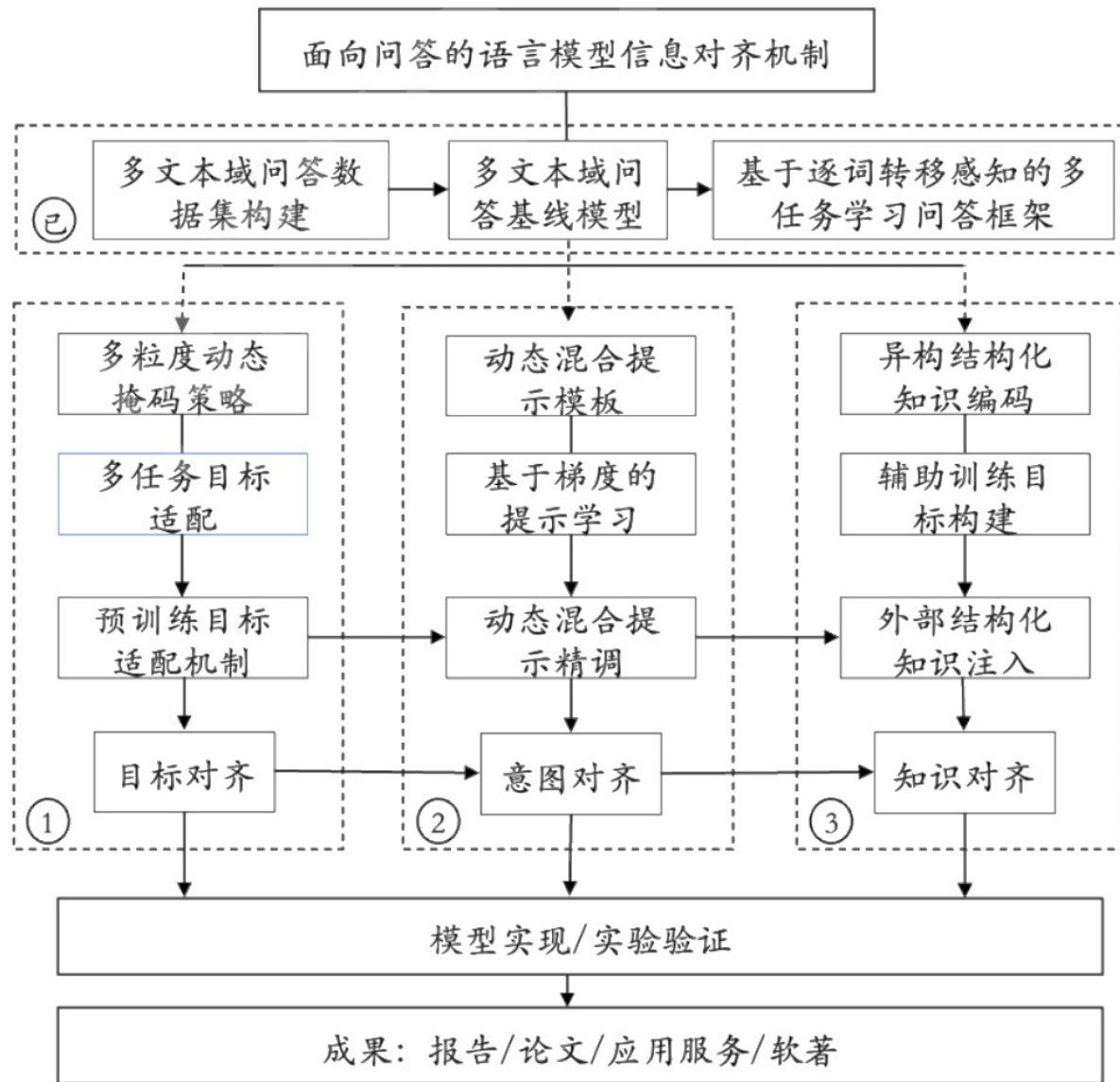
考察掩码内容粒度、掩码率和训练目标的动态变化对模型收敛速度和收敛结果的影响；在语言模型预训练阶段设计多粒度动态掩码策略以提高模型参数效率，并适配多任务目标以获取预训练目标适配机制。通过对比目标对齐前后模型在问答任务中的性能表现，分析该机制的有效性和稳定性。



本项目的研究脉络图

研究内容 ②

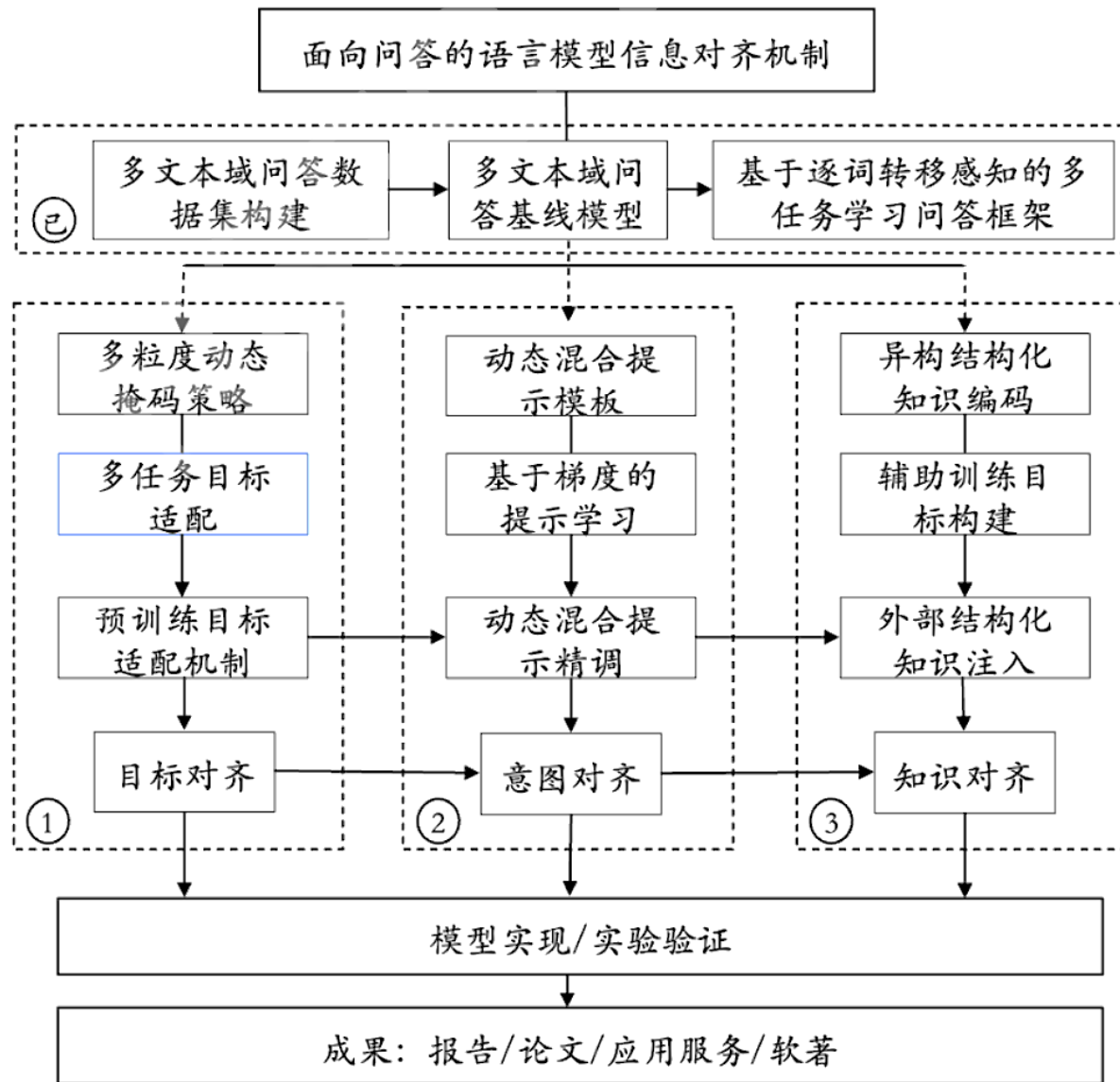
基于动态混合提示精调的意图对齐机制研究：考察预训练语言模型内在感知信息的模式，设计基于连续提示和离散提示的混合提示问答模板；针对不同问题和上下文输入动态生成提示内容的需求制定动态混合提示精调机制。通过用户意图的表达与语言模型的信息感知模式相互适配，获取意图对齐机制，并分析动态提示的有效性，以及模型优化算法的执行效率。



本项目的研究脉络图

研究内容 ③

基于结构化知识注入的知识对齐机制研究：
考察语言模型在训练过程中发生的知识遗忘现象，设计异构结构化知识编码与知识注入方法；通过构建基于结构化知识重建的辅助训练目标，获取基于外部结构化知识注入的知识对齐机制，对对齐后的模型进行生成回答的真实性和可靠性分析。



本项目的研究脉络图

特色与创新点

- 提出了基于预训练目标适配的目标对齐方案。基于掩码内容粒度、掩码率的动态变化策略与预训练目标的适配，提升模型的收敛速度和收敛效果。这种方案优化了模型的预训练目标，提高了模型参数效率。
- 提出了基于动态混合提示精调的意图对齐方案。基于连续提示和离散提示的混合提示问答模板可以为不同输入定制提示内容，从而使模型更好地理解问题背后的用户意图，以获取更准确、符合用户需求的回答。这是一种新颖的提示精调方案。
- 提出了基于结构化知识注入的知识对齐方案。在知识结构中引入混合提示得到异构结构化知识，丰富语言模型对事实知识的感知，确保回答的真实性。此方案还可应用在领域知识图谱中以实现模型的专业化，具有很好的可扩展性。

研究基础

- 申请者已经取得了和本项目密切相关的科研成果，主要包括：多领域问答系统构建方面和知识表示、感知与应用方面，相关研究成果已发表在EMNLP、KR、IP&M和SIGKDD等国际会议及期刊。
- 申请人以取得与本项目研究相关成果总结：本项目研究初期，申请人已在开放领域和特定领域收集了大规模语料数据，并通过规范的人工标注流程构建了一个高质量的问答数据集；设计并实现了一套基于联合学习的问答系统语言模型预训练范式；提出了多种知识感知、推理与知识注入的模型与方法；提出了不同领域的问答任务语言模型；提出了多种提示生成方法，并成功将其应用于多种场景的问答任务与语言模型中。这些相关研究成果的取得，为进一步深入研究本项目提供了良好的理论基础。

请各位专家批评指正