Web-enhanced LFQA

张颖而

背景

- 什么是Web-enhanced Long-form Question Answering?

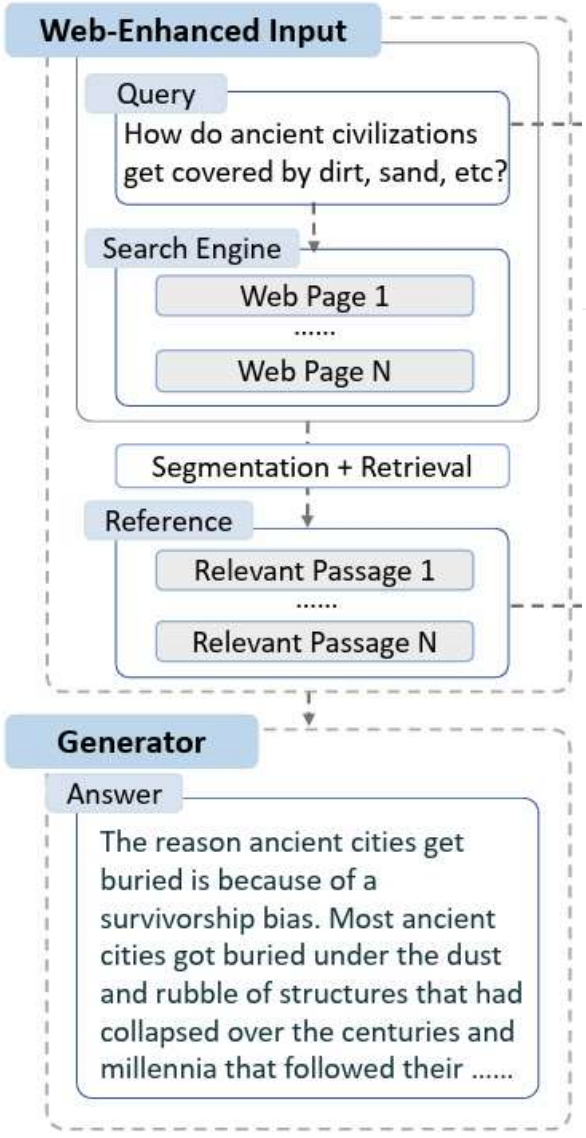
给定用户query, 先使用搜索引擎召回相关内容，再使用LLM 整合这些内容得到答案。

- 业务背景

需要在支付宝的智能助理场景上线，需要证明我们的技术是中文sota。

- Task

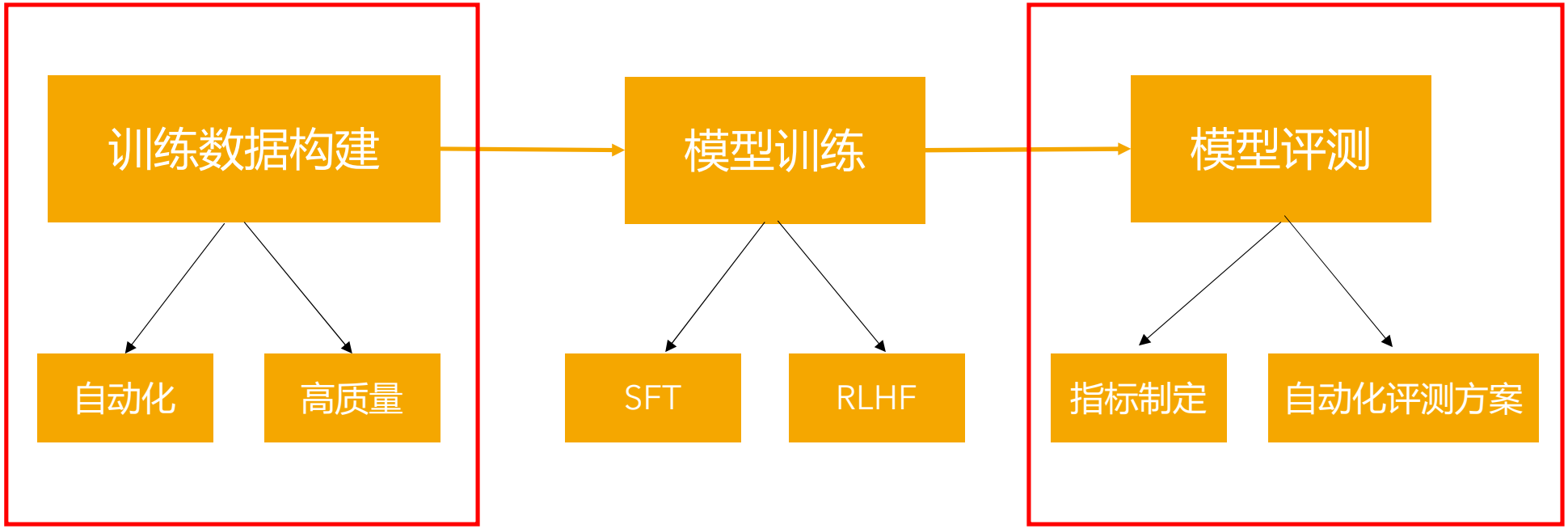
调研当前的sota，要对比的数据集，指标，技术方案的优化方向。



Sota 分析

	WebGPT	WebCPM	WebGLM
语言	英文	中文	英文
技术架构	interactive	Interactive \ pipeline	pipeline
训练数据集	ELI5, 未公开	5.5k, Human annotation，公开	ELI5, 公开
测试数据集	只公开272个测试集	公开webcpm数据集	272 个WebGPT测试集 TriviaQA
评测指标和方案	Human preference	Human preference	Human preference

Pipeline



数据集

当前检索增强领域还没有大规模的中文数据集，自动化创建10万-100万量级以上的高质量数据。

- **自动化流程**:

使用 Chatgpt/GPT-4 ， pipeline：优化prompt, 先删除无关文章，抽取，再融合。

SoT (Skeleton-of-thought) 方案提升回答质量: 回答问题的时候采用skeleton-of-thought，先列提纲再回答的方案。能够提升回答的质量。

- **高质量:**

Query筛选：知乎，ELI5, 搜狗问问，百度知道，DuReader, 政务，医疗等。爬虫+规则筛选。

Reference 清洗：Google search TOP 20 文章，爬虫团队清洗。

SoT方案：

指标和评估方案

- 指标

设计了统一的指标 coherence, helpfulness, factual consistency

- 自动化评估

设计GPT-4评估，确保与人工一致性：coherence 91.5%, helpfulness 83.0%

其中factual consistency 准确率很低，因为reference和 answer都很长，很难直接打分。
因此 factual consistency 的评估是一个待优化的方向。

- 训练一个新的模型拆分answer 为 subclaim，针对每一个subclaim 进行推理验证
- 训练一个NLI 模型，构造样本覆盖错误类型。
- 将factual consistency 与人工一致性提高到：sentence 粒度的95.5%

效果

Model	Answer Evaluation									
	WebCPM (zh)					WebGPT (en)				
	Cohr.	Help.	Fact/q.	Fact/s.	Avg. Len.	Cohr.	Help.	Fact/q.	Fact/s.	Avg. Len.
WebGPT 175b	-	-	-	-	-	0.6911	0.9154	0.8823	0.9752	209
WebGPT 13b	-	-	-	-	-	0.5478	0.7390	0.7977	0.9642	212
WebGLM 10B	-	-	-	-	-	0.5919	0.8566	0.8639	0.9688	169
WebCPM 10B	0.4899	0.6985	0.6784	0.8916	549	0.7316	0.8566	0.8125	0.9764	330
FoRAG-C 6B (Ours)	0.8618	0.7764	0.7739	0.9639	655	0.8603	0.8640	0.7610	0.9804	443
FoRAG-L 7B (Ours)	**0.9121**	**0.8668**	**0.8216**	**0.9727**	625	**0.9889**	**0.9595**	**0.8897**	**0.9894**	447