

Web-enhanced LFQA

张颖而

背景

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

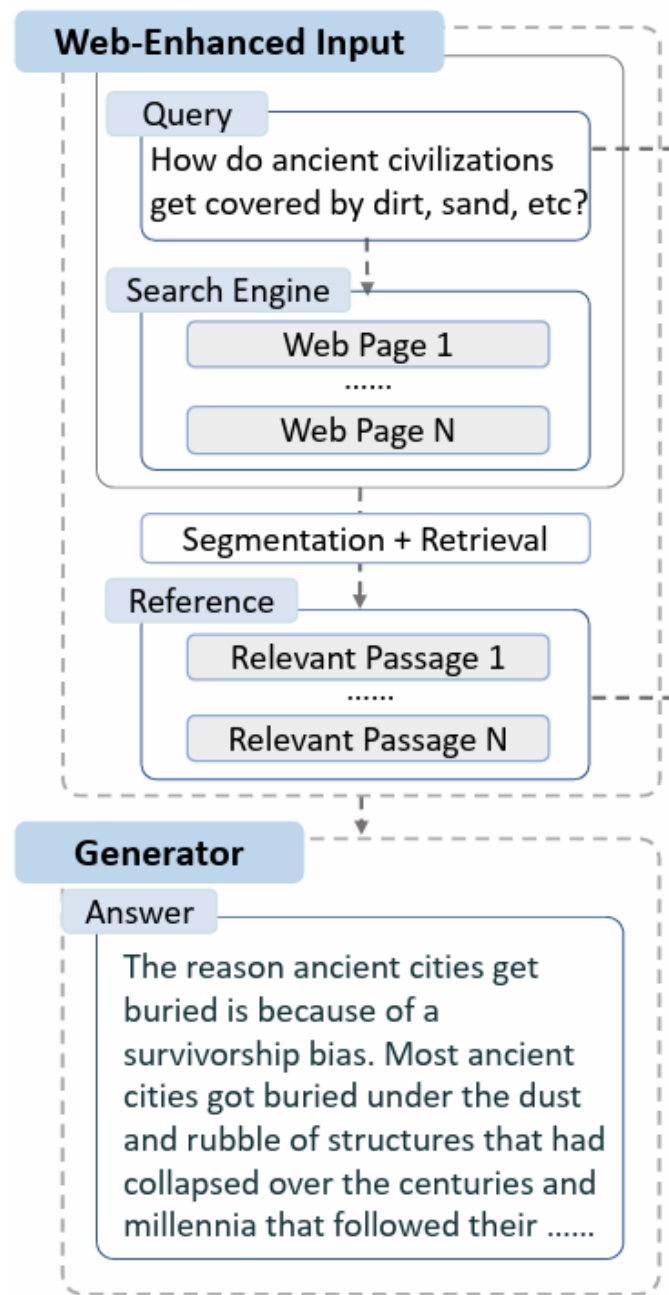
针对用户query, 先使用**搜索引擎**召回相关内容, 再使用**LLM** 整合这些内容得到答案。

- 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sota。

- Task

调研当前的sota, 要对比的数据集, 指标, 技术方案的更迭, 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	问题 ELI5翻译, Bing Search, 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

数据集

- 测试集

Webcpm 测试集, WebGPT 测试集

- 训练集

最重要的是要创建10万-100万量级以上的高质量数据

Query 来源: 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。

自动化流程: GPT-4, 先删除无关文章, 抽取, 再融合。

指标和评估方案

- 指标

coherence, helpfulness, factual consistency

- 自动化评估

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
- 将factual consistency 与人工一致性提高到: sentence 粒度的95.5%

技术方案

- 采用pipeline 架构

先从每一篇文章中抽取内容，再合并抽取的内容。

- 优化方案

1. 回答问题是时候采用skeleton-of-thought，先列提纲再回答的方案。能够提升回答。
2. Factual consistency 评估的优化.
3. Factuality optimization approach based on doubly fine-grained RLHF framework

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	<u>0.9154</u>	<u>0.8823</u>	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)	<u>0.8618</u>	<u>0.7764</u>	<u>0.7739</u>	<u>0.9639</u>	655	<u>0.8603</u>	0.8640	0.7610	<u>0.9804</u>	443
FoRAG-L 7B (Ours)	0.9121	0.8668	0.8216	0.9727	625	0.9889	0.9595	0.8897	0.9894	447