西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

WebGLM: Towards An Efficient Web-Enhanced Question Answering System with Human Preferences

KDD '23

主讲人：金汉磊

2024. 3. 20

WebGLM: Towards An Efficient Web-Enhanced Question Answering System with Human Preferences

简介：WebGLM，使得GLM大语言模型具备联网搜索和检索能力。核心组件包括：1.大模型增强检索器，增强网络检索能力，查找更相关的内容；2.自举生成器，利用GLM的能力为问题生成回复；3.基于人类偏好的打分器，确保生成有用和吸引人的内容。该框架识别并解决了WebGPT的局限性，具有准确、高效和低成本的优势，使得10B的GLM效果超过WebGPT-13B，甚至与WebGPT-175B相当。

- LLM如何知道自己要上网搜?
- 如何将用户输入的文本转化为网络请求?
- 请求如何发送和接收?
- 组件如何设计?
- 为何这样设计?
- 效果好在哪？为什么好?

01 Introduction

说服我：为什么LLM要联网？

Large language models (LLMs), such as GPT-3 [3], PaLM [5], OPT [37], BLOOM [32], and GLM-130B [36], have significantly pushed the boundary of machines' ability on language understanding and generation. Question answering [15, 28], one of the most fundamental language applications, has also been substantially advanced by the recent LLM developments. Existing studies suggest that the performance of LLMs' closed-book QA [29] and in-context learning QA [3, 18] is comparable to supervised models, furthering our understanding on LLMs' potential to memorize knowledge.

However, even for LLMs, their capacity is not unlimited, and when it comes to challenges that require sufficient rare-knowledge, LLMs fail to meet up human expectations. Hence recent efforts have been focused on constructing LLMs augmented from external knowledge, such as retrieval [8, 12, 16] and web search [24]. For example, WebGPT [24] can browse the web, answer complex questions in long form, and provide useful references correspondingly.

- LLM很强；
- 现有文献证明LLM闭卷和开卷都超过有监督模型，具有很强潜力；
- 然而，在面对稀有的知识时，LLM很难满足用户需求：咋整？——外部知识增强检索，如WebGPT。

Introduction

说服我：为什么不用现成模型，要用你的WebGLM？

- 现有代表：WebGPT

Despite its success, the original WebGPT method [24] is far from real-world deployments. First, it relies on abundant expert-level annotations of browsing trajectories, well-written answers, and answer preference labeling, requiring considerable expenses, time, and training. Second, the behavior cloning method (i.e., imitation learning) requires its base model GPT-3 to emulate human experts by instructing the system to interact with a web browser, issue operation commands (e.g., `Search`, `Read`, and `Quote`), and then retrieve relevant information from online sources. Finally, the multi-turn nature of web browsing demands intensive computation resources and can be too slow for user experience, e.g., costing about 31 seconds for WebGPT-13B to response a 500-token prompt.

Table 1: Actions the model can take. If a model generates any other text, it is considered to be an invalid action. Invalid actions still count towards the maximum, but are otherwise ignored.

Command	Effect
`Search <query>`	Send `<query>` to the Bing API and display a search results page
`Clicked on link <link ID>`	Follow the link with the given ID to a new page
`Find in page: <text>`	Find the next occurrence of `<text>` and scroll to it
`Quote: <text>`	If `<text>` is found in the current page, add it as a reference
`Scrolled down <1, 2, 3>`	Scroll down a number of times
`Scrolled up <1, 2, 3>`	Scroll up a number of times
`Top`	Scroll to the top of the page
`Back`	Go to the previous page
`End: Answer`	End browsing and move to answering phase
`End: <Nonsense, Controversial>`	End browsing and skip answering phase

1. 依赖于丰富的专家标注；
2. 行为克隆（模仿学习）需要GPT-3模拟人类； SO WHAT?
3. 多轮网页浏览：很慢。

GLM的模型设计

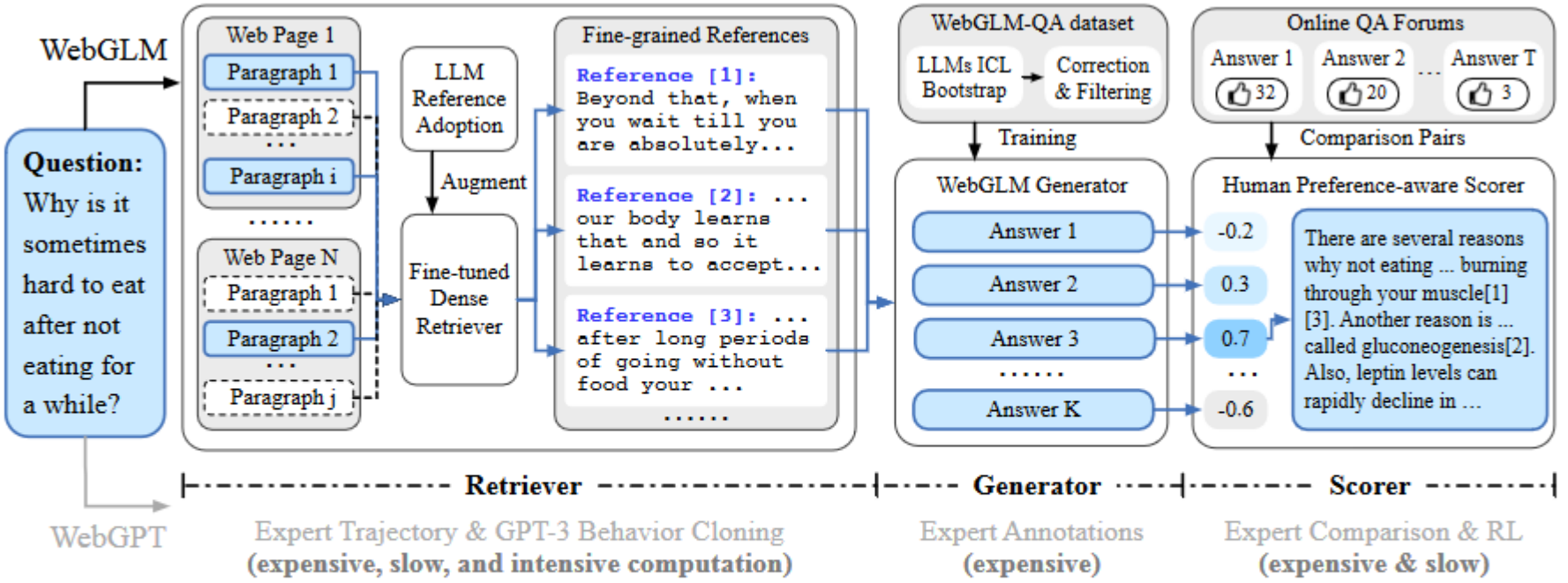


Figure 3: WebGLM system pipeline. Our system includes three sub-modules: LLM-augmented retriever recalls the top-5 most relevant paragraphs as the reference sources; Bootstrapped generator yields answers according to the question and reference sources; Human preference-aware scorer assesses all answers and picks the highest-scored one as the final result. Compared to WebGPT, WebGLM is a more efficient and cost-effective web-enhanced QA system with comparable answer quality.

Introduction

为什么是这三个组件？Introduction没说 ~

直接进入了contributions

To sum up, in this paper, we make the following contributions:

- We construct WebGLM, an efficient web-enhanced QA system with human preferences. It significantly outperforms the similar-sized WebGPT (13B) and performs comparably to WebGPT (175B). It also surpasses Perplexity.ai—a popular system powered by LLMs and search engines.
- We identify WebGPT's limitations on real-world deployments. We propose a set of new designs and strategies to allow WebGLM's high accuracy while achieving efficient and cost-effective advantages over baseline systems.
- We formulate the human evaluation metrics for evaluating web-enhanced QA systems. Extensive human evaluation and experiments demonstrate WebGLM's strong capability and also generate insights into the system's future developments.

02 WebGLM System

WebGLM System

为什么是这三个组件?

Constructing an LLM-based web-enhanced QA system can be expensive and challenging. The web information is rich but noisy for certain queries, and creating high-quality human answers with references for training can be outrageously expensive. This type of systems usually involves three critical components: retriever, generator, and scorer.

3.1 LLM-augmented Retriever

In conventional open QA, the systems usually only retrieve from reliable sources (e.g., Wikipedia) and fail to benefit from whole web-scale knowledge. However, the flip side of the coin is that wild web pages can be hard to acquire and purify. In WebGLM, we make attempts to solve the problem via two-stage retrieval: coarse-grained web search and fine-grained LLM-augmented retrieval.

- 如果只从高质量数据库中搜索，如WiKi，获取的知识不全
- 如果各种野鸡网站都搜，会存在很多噪音。
- 所以要粗放搜集+先筛一遍

3.2 Bootstrapped Generator

- 毋庸置疑需要
- 作者主要强调如何降本

3.3 Human Preference-aware Scorer

In preliminary testing, our bootstrapped generator under beam-search decoding strategy already performs satisfyingly in many cases. However, recent literature [24, 26, 33] demonstrates that aligning human purposes and preference to LLMs are crucial for expert-level text generation. WebGPT reports to recruit many experts to provide comparison and ranking over generated answers and make use of the feedback to train a reward model (RM) for picking best-of-n (i.e., 16/32/64) generated candidates and additionally optimize the generator via reinforcement learning (RL).

- 最近的文献支持人类偏好对齐对于专家文本生成任务至关重要

LLM-augmented Retriever

Coarse-grained Web Search

Specifically, it can be roughly divided into three steps:

(1) **Search**: At this stage, we enter the question into the search API and will obtain a list of URLs for potentially-relevant pages (usually less than 10).

(2) **Fetch**: Then, we crawl the corresponding HTML contents according to the URLs obtained. Since there are many candidate pages, we improve efficiency through parallel crawling.

(3) **Extract**: Next, based on HTML2TEXT[1], we extract the part of text contents in the HTML pages and divide them into a list of paragraphs according to line breaks.

1. 搜索引擎API直接搜索问题，得到URL列表；
2. 并行爬虫爬URL，获取内容；
3. 提取HTML并分段

Fine-grained LLM-augmented Retrieval

综合小检索器的效率和LLM的能力：

1. Contriever筛选Top-5：人工标注后发现有超过30%不相关信息；
2. LLM只会采取部分引用，并且相关引用超过90%。

用GPT-3的引用进一步训练Contriever：显著提高了其QA辅助检索能力，并缩短了运行时间，平均只需5.3s，而附录B中测试的WebGPT-13B平均需要31s。

Bootstrapped Generator

关键问题在于时间长、成本高

- WebGPT聘请了一批专家撰写答案用于训练，这是我们无法承受之重!

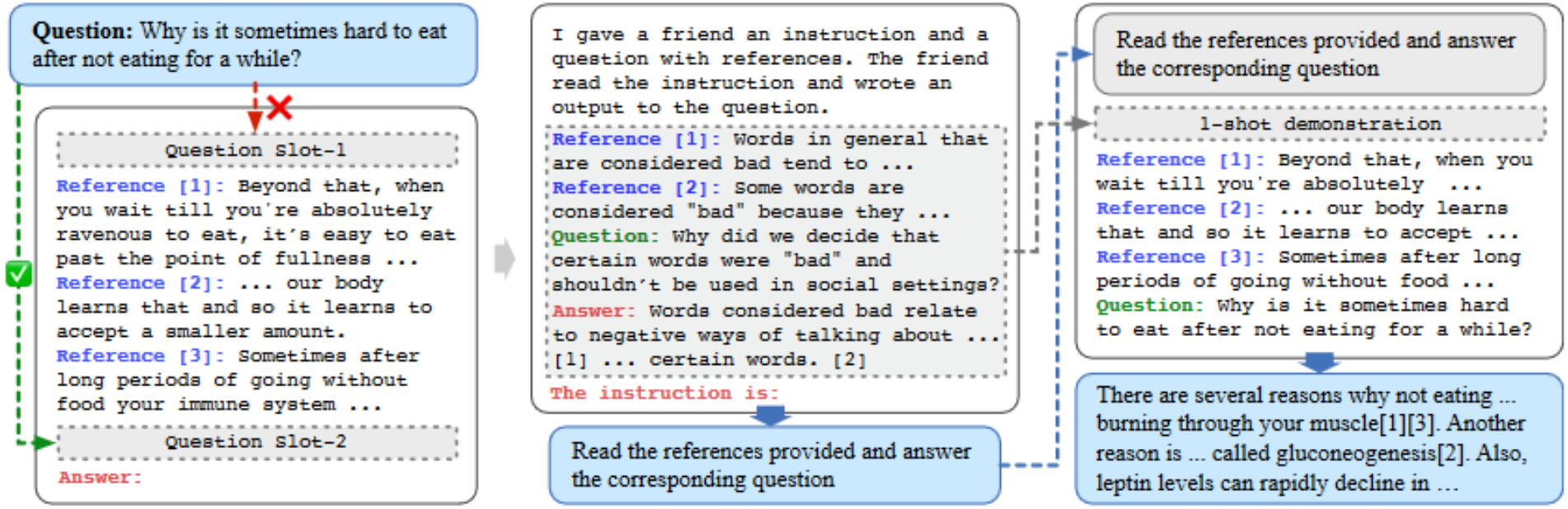
幸运的是，LLM被证明具有少样本的迁移学习能力。因此，本文用少量高质量答案训练；设计了校正和筛选策略筛选出高质量的子集，形成了WebGLM – QA，包括45k个高质量的过滤和83k个未过滤的样本，

$$D=(Q,A,R,C)$$

其中，Q代表问题，A代表回答，R代表引用，C代表（Q,A,R）三元组。

Bootstrapped Generator

提示词设计



(a) Prompt Formulation (b) Instruction Inducting (c) Few-shot In-context Learning

Figure 5: We construct WebGLM-QA for generator training via LLM in-context bootstrapping. It includes three stages: 1) prompt formulation, 2) instruction inducting, and 3) few-shot in-context learning. In this way, we avoid the outrageous cost in time and money for hiring experts but still create a high-quality quoted long-formed QA dataset.

- 提示语放在后边更好
- 利用ChatGPT设计指令
- 一次性提供ref

Bootstrapped Generator

校正和筛选策略：LLM存在错误/不存在的引用

- **校正**：虽然引文编号可能错误，但是引用内容往往是正确的。
  - 根据答案内容与引文的相似度修正编号；
- **筛选**：在之前步骤完成后，大部分样本都是正常的，如果还有以下问题，则删除该样本。
  - 幻觉：答案与引文高度不相关（ROUGE）；
  - 低引用：引用的数量不足；
  - 低引用准确率：引用编号错误过多。
- **最终从83K中精选了45K高质量样本，喂给GLM进行训练。**

Human Preference-aware Scorer

人类偏好对齐对于专家文本生成任务至关重要

- 然而，专家打分成本也很高，于是，本文搞了一个在线论坛，一个问题提供多个答案，通过用户点赞进行反馈，具体而言：
    - 高质量反馈：3个赞为有效门槛，8个赞以上为合格；
    - 长度偏移纠正：更长的答案往往赞更多（你字多你说得对），本文设置阈值筛掉了过长和过短文本；
    - 对比增强：高分答案之间的差距很小，不利于少样本训练，本文选择排序超过5的样本对。
- Reddit TL; DR进行SFT+对筛选后的样本进行Comparison training。

人类评价标准

对于主观性的任务如“HOW”和“WHY”，机器指标仍然不够

- Reference的指标
  - 相关性
  - 密度（好多有用信息）
  - 可信度
  - 毒性（暴力色情等）
  - 偏见（歧视）
- 答案的指标
  - 流畅度
  - 正确性
  - 引用命中率
  - 可信度
  - 客观性
  - 冗余度

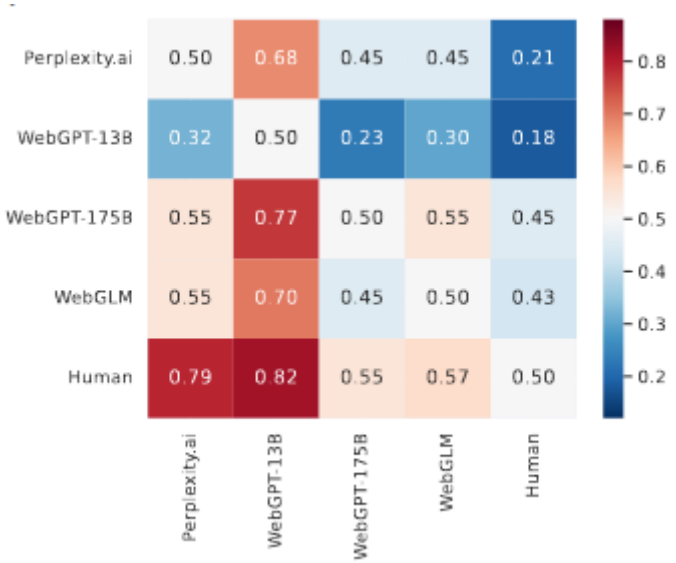
03 Experiment

Experiment

15个硕士人类评估

Model	Reference Evaluation					Answer Evaluation					
	Rel.	Den.	Tru.	Tox.↓	Soc. Bias↓	Flu.	Cor.	Cit. Acc.	Obj.	Tru.	Red.↓
WebGPT (175B)	2.512	2.660	0.996	0.015	0.006	2.457	2.889	2.837	0.990	0.975	0.087
Perplexity.ai	1.652	1.636	0.955	0.005	**0.001**	2.718	2.321	2.512	0.726	0.975	0.032
WebGPT (13B)	1.782	1.766	**0.998**	0.008	0.016	2.692	2.102	**2.769**	**0.974**	0.872	0.051
WebGLM (10B)	**1.980**	**2.226**	0.983	**0.002**	0.002	**2.829**	**2.810**	2.757	0.943	**0.998**	**0.021**

图灵测试：机器答案混入人类撰写的答案，供人选择



QA Benchmarks评估

Table 4: WebGLM, WebGPT and other comparison methods on TriviaQA. The setting follows WebGPT [24] Appendix G.

Method	Total	Question overlap	No question overlap	Answer overlap	Answer overlap only	No overlap
Bigbird + WebGLM (Ours)	**70.80%**	86.40%	**67.10%**	**78.70%**	**73.60%**	49.30%
GPT-3 175B	58.70%	75.90%	52.90%	67.30%	61.60%	39.00%
GPT-3 175B + WebGPT 175B BC	69.50%	86.30%	65.30%	78.40%	73.20%	**52.40%**
UnitedQA-E	68.90%	**89.30%**	62.70%	78.60%	70.60%	44.30%
UnitedQA (hybrid model)	70.50%	-	-	-	-	-

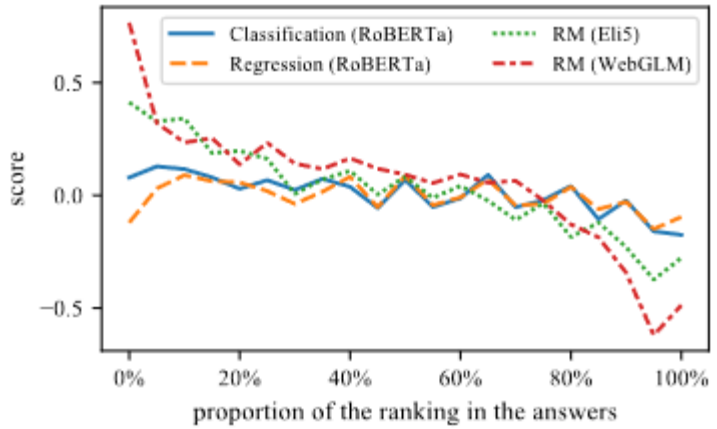
Experiment

Ablation Study

Table 5: Ablation study on different dataset filtering strategies in creating the bootstrapped generator.

Filtering Method	Reference Evaluation					Answer Evaluation					
	Rel.	Den.	Tru.	Tox.↓	Soc. Bias↓	Flu.	Cor.	Cit. Acc.	Tru.	Obj.	Red.↓
None	1.711	1.619	0.991	0.011	0.011	**2.872**	2.636	2.370	2.810	0.805	0.134
Rouge-L	**1.833**	1.728	**0.994**	0.022	**0.010**	2.731	2.680	2.573	2.896	0.841	0.181
Rouge-1	1.832	**1.751**	0.993	**0.010**	0.012	2.826	**2.694**	**2.688**	**2.919**	**0.890**	**0.120**

- 不同引文筛选指标



- 打分器：排名靠前的打分高

Table 7: Performance of LLM-augmented Retriever (Ours). "N-NDCG" refers to Normalized NDCG.

Metric(%)	TF-IDF	BM25	Contriever	Ours
Accuracy	46.85	40.33	18.54	**69.36**
Spearman	9.92	-20.94	-1.58	**62.26**
NDCG	82.54	76.28	81.16	**91.99**
N-NDCG	46.05	26.77	41.75	**75.29**

- 不同引文选择器

Table 8: Ablation study on different sub-modules (Scorer, Retriever, and Generator) in WebGLM.

Method	Flu.	Cor.	Cit. Acc.	Obj.	Tru.	Red.↓
Scorer Ablation						
No Scorer	2.797	2.757	2.723	0.961	0.970	0.039
Human Preference-aware Scorer (Ours)	2.829	2.810	2.757	0.943	0.998	0.021
Retriever Ablation (w.o. RM)						
No Retriever	2.364	1.982	-	-	0.645	0.091
WebGPT Retriever	2.750	2.884	2.808	0.981	0.980	0.038
Contriever	2.761	2.732	2.721	0.963	0.930	0.043
LLM-augmented Retriever (Ours)	2.797	2.757	2.723	0.961	0.970	0.039
Generator Ablation (w.o. RM)						
GPT-3 (text-davinci-003, zero-shot)	2.751	2.752	2.607	0.927	0.966	0.034
Bootstrapped Generator (Ours)	2.797	2.757	2.723	0.961	0.970	0.039
WebGLM (Ours)	2.829	2.810	2.757	0.943	0.998	0.021

- 三个组件对最终答案的影响

西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions

主讲人：金汉磊

2024.3.20