FoRAG: Factuality-optimized Retrieval Augmented Generation for Web-enhanced Long-form Question Answering

Tianchi Cai, Zhiwen Tan, Xierui Song, Tao Sun, Jiyan Jiang, Yunqi Xu, Yinger Zhang, Jinjie Gu

KDD 2024

Presenter: Yu-Hua Zeng

National Cheng Kung University

國立成功大學
National Cheng Kung University
1931

OUTLINE

- Introduction
- Related work
- Method
- Experiment
- Conclusion

Introduction - RAG

- LFQA(web-enhanced long-form question-answering task)
  - Access to search engine supplements massive and latest knowledge to LLMs
  - open domain dialogue [1] and question answering (QA) [2].
  - Bing Chat, Perplexity.ai
  - Recent researches have revealed the low factuality issue of these systems[3, 4]

[1] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kul shreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022).
[2] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. arXiv preprint arXiv:2208.03188 (2022).
[3] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. arXiv preprint arXiv:2304.09848 (2023).
[4] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large LanguageModelstoGenerateTextwithCitations. arXiv preprint arXiv:2305.14627 (2023).

Introduction - Challenge

1. Previous studies mostly rely on human evaluation [1, 2, 3], which is generally expensive to acquire.
   - Comparing the factual details of two lengthy texts.
2. Reinforcement Learning from Human Feedback (RLHF), conventionally adopts the holistic reward.
   - Reward provides a relatively sparse training signal, which undermines the reliability of RLHF.

   [question, candidate0, cadidate1, choice]

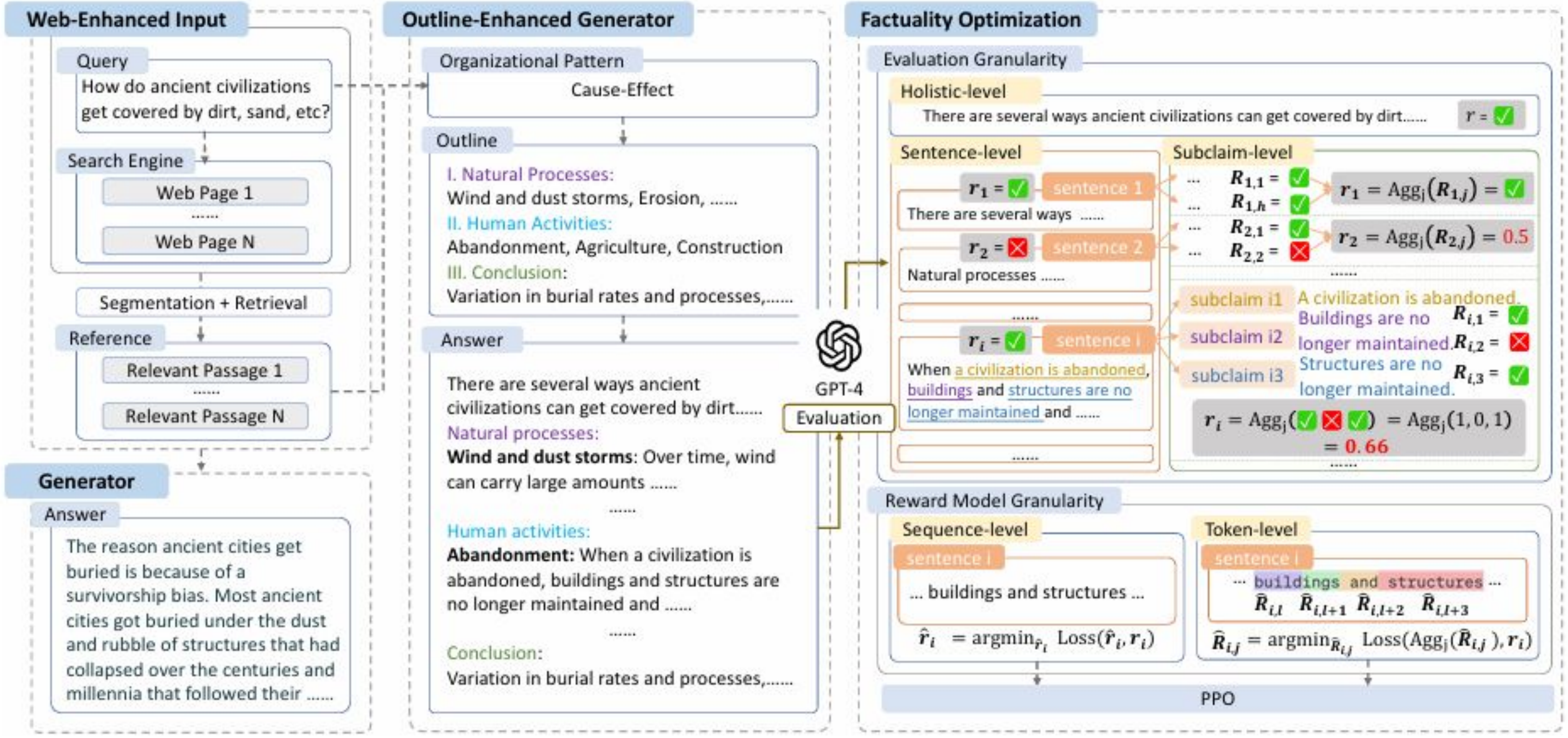
[1] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. arXiv preprint arXiv:2304.09848 (2023).
[2] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. arXiv preprint arXiv:2203.11147 (2022).
[3] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332 [cs.CL]

Related work

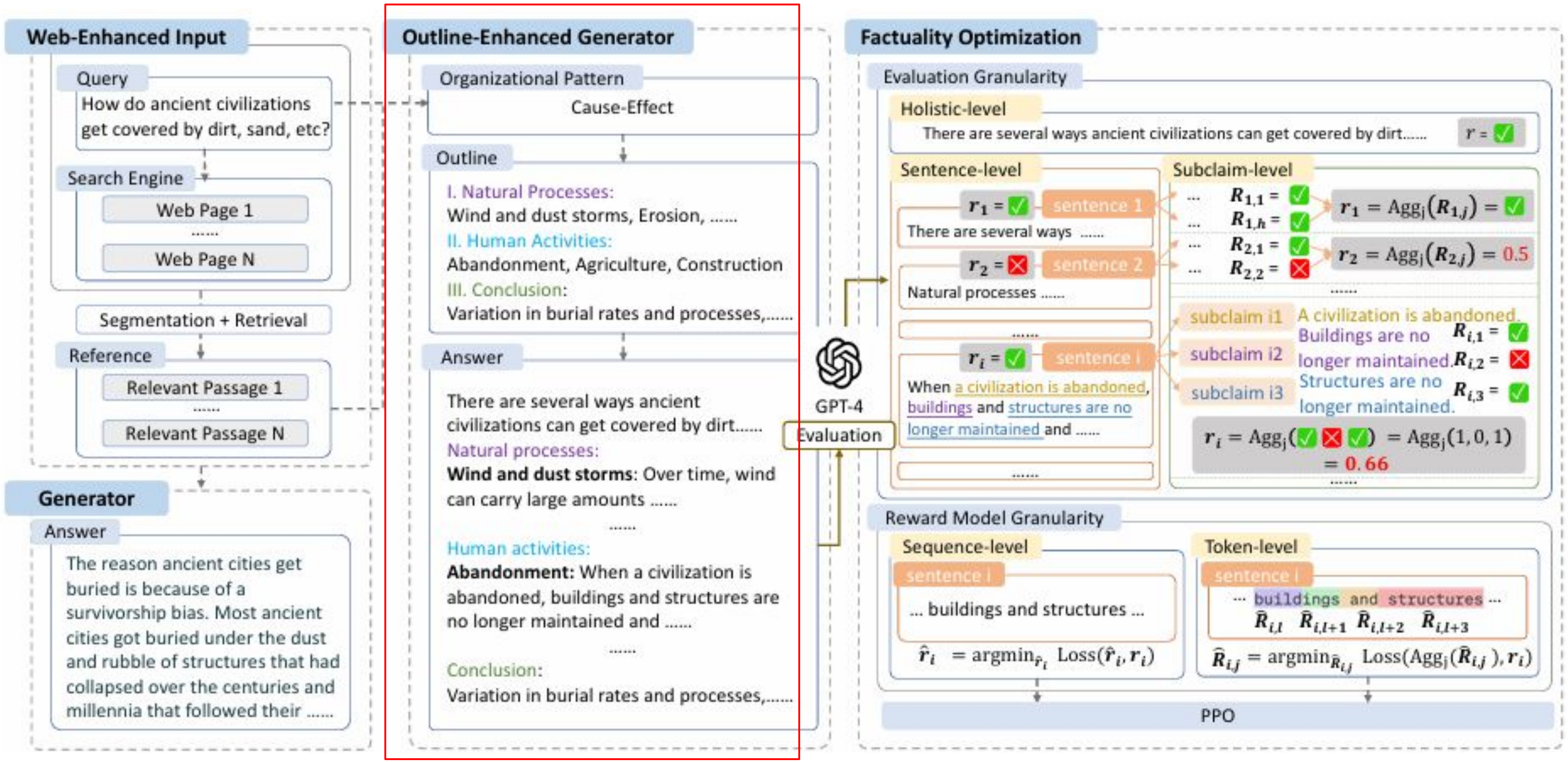
- WebGPT[1]
  - The dataset, and models are not accessible to the public.
- WebCPM[2] (ACL 2023)
  - Chinese Long-form Question Answering
- WebGLM[3] (KDD 2023)
  - Replacing the expert annotation with evaluations using LLMs
  - Utilizing a non-interactive way to use search engine.

[1] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332 [cs.CL]
[2] Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. WebCPM: Interactive Web Search for Chinese Long-form Question Answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 8968–8988.
[3] Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. WebGLM: Towards An Efficient Web-Enhanced Question Answering System with Human Preferences. In Proceedings of the 29th ACMSIGKDDConference on Knowledge Discovery and Data Mining (, Long Beach, CA, USA,) (KDD ’23). Association for Computing Machinery, New York, NY, USA, 4549–4560. https://doi.org/10.1145/3580305.3599931

Method - Framework



Method - Outline-Enhanced Generator



Method - Outline-Enhanced Generator

- Outline stage
    - consider which organizational pattern is best suitable to the current question.
    - Use the organizational pattern to output an outline.

###任务###
根据资料回答问题。

###要求###
第一步：根据问题和资料生成回答提纲。
1. 决定回答结构，从总分总、递进、对比、因果、并列、时序等结构中选择合适的来组织回答。
2. 根据回答结构，在提纲中要完整地列出答案中需要包括的要点。
3. 要点之间可以是并列、对照、递进等关系，不可以是重复或者包含关系。
4. 要点要保持精炼，至少有1点，不能多于5点，
5. 每个要点仅可参考1段资料，并在提纲中标注资料编号。

第二步：根据资料和提纲对问题进行回答。
1. 回答要以提纲为蓝本，对问题进行详细的回答。
2. 回答中可以采用编号或项目列表、小标题、latex公式等格式。
3. 回答中减少使用“首先”、“其次”、“再者”等简单的连接词。
4. 回答中不要生成重复内容。
5. 回答中不要标注资料来源。
6. 回答应当严格依据资料，不采用不在资料中的内容。

###格式###
【结构】：
<回答的组织结构>
【提纲】：
<分点介绍回答思路>
【回答】：
<根据资料和提纲回答问题>

下面是1个示例输入和2个满足要求的示例输出：

###示例输入###
###问题###
2023年西安房贷利率最新消息

###资料###
[1]一、西安商业贷款固定利率
1年以内（含）——4.35%
5年(含)以下——4.75%
5年以上——4.9%
贷款市场报价利率LPR：目前1年期LPR为3.45%，5年期LPR为4.2%
首套住房商业性个人住房贷款利率下限为不低于相应期限LPR减20个基点。
二套住房商业性个人住房贷款利率下限为不低于相应期限贷款市场报价LPR利率加20个基点。
二、西安公积金贷款利率
5年(含)以下——2.6%
5年以上——3.1%
[2]目前西安主流银行的首套房利率集中在4%左右，二套房利率差基本维持在4.9%。
[3]首先，虽然LPR在7月没有变动，但西安首套房贷款利率已经低至4%，并且低于2009年房贷利率打七折后的4.156%！这点也恰恰和楼市走访到的信息不谋而合，据了解，西安目前多数银行首套房贷利率主要集中于4%，二套房贷利率基本在4.9%左右。
[4]7月20日，中国人民银行授权全国银行间同业拆借中心公布了最新一期贷款市场报价利率（LPR）：1年期LPR为3.55%，5年期以上LPR为4.20%，均与上个月持平。但西安房贷利率较上月小幅下行，西安多家银行首套房贷利率从4.1%降至4%。
今年6月，LPR时隔10个月迎来下调，1年期和5年期以上LPR均跟随政策利率下调10个基点。记者了解到，从6月下旬开始，西安各大银行相继落实首套房贷利率政策动态调整机制，进行利率调整。目前，西安地区六大行及招商银行、宁夏银行、北京银行、西安银行等十余家银行首套房贷利率降至4%，较LPR下浮0.2%。二套房贷利率同步下行，跌破5%大关，维持在4.9%。

###Requirements###
Step One: Develop an answer outline based on the question and materials.
1. Choose a suitable organizational pattern for the answer structure, such as general-specific-general, progressive, comparative, cause-effect, parallel, chronological, among others.
2. Enumerate the essential points that need to be included in the outline, aligned with the the chosen structure.
3. The relationship between key points can be parallel, contrastive, progressive, etc., but should not be repetitive or inclusive.
4. Formulate a clear and concise outline that includes at least 1 but no more than 5 key points.
5. Each main point should reference only one specific part of the provided materials and must include the material's number within the outline.

Method - Outline-Enhanced Generator

- Expansion stage
  - Based on the outline generated at the former stage, the LLM expands each perspective to construct the final answer.

Step Two: Answer the question based on the materials and outline.
1. Utilize the outline as a blueprint to develop a comprehensive and informative answer.
2. Write the answer using formatting tools such as numbered lists, bullet points, subheadings, LaTeX formulas, etc., where appropriate.
3. Refrain from using basic sequential connectors like "firstly," "secondly," or "furthermore," in the answer.
4. Avoid redundancy and repetition of content within the answer.
5. Do not cite the number of the materials in the answer.
6. Adhere strictly to the information contained within the provided materials, without adding any information that is not included in the materials.

```
###任务###
根据资料回答问题。

###要求###
第一步：根据问题和资料生成回答提纲。
1. 决定回答结构，从总分总、递进、对比、因果、并列、时序等结构中选择合适的来组织回答。
2. 根据回答结构，在提纲中要完整地列出答案中需要包括的要点。
3. 要点之间可以是并列、对照、递进等关系，不可以是重复或者包含关系。
4. 要点要保持精炼，至少有1点，不能多于5点，
5. 每个要点仅可参考1段资料，并在提纲中标注资料编号。

第二步：根据资料和提纲对问题进行回答。
1. 回答要以提纲为蓝本，对问题进行详细的回答。
2. 回答中可以采用编号或项目列表、小标题、latex公式等格式。
3. 回答中减少使用"首先"、"其次"、"再者"等简单的连接词。
4. 回答中不要生成重复内容。
5. 回答中不要标注资料来源。
6. 回答应当严格依据资料，不采用不在资料中的内容。

###格式###
【结构】：
<回答的组织结构>
【提纲】：
<分点介绍回答思路>
【回答】：
<根据资料和提纲回答问题>

下面是1个示例输入和2个满足要求的示例输出：

###示例输入###
###问题###
2023年西安房贷利率最新消息

###资料###
[1 ]一、西安商业贷款固定利率
1年以内（含）——4.35%
5年(含)以下——4.75%
5年以上——4.9%
贷款市场报价利率LPR：目前1年期LPR为3.45%，5年期LPR为4.2%
首套住房商业性个人住房贷款利率下限为不低于相应期限LPR减20个基点。
二套住房商业性个人住房贷款利率下限为不低于相应期限贷款市场报价LPR利率加20个基点。
二、西安公积金贷款利率
5年(含)以下——2.6%
5年以上——3.1%
[2 ]目前西安主流银行的首套房利率集中在4%左右，二套房利率差基本维持在4.9%。
[3 ]首先，虽然LPR在7月没有变动，但西安首套房贷款利率已经低至4%，并且低于2009年房贷利率打七折后的4.156%！
这点也恰恰和楼市走访到的信息不谋而合，据了解，西安目前多数银行首套房贷利率主要集中于4%，二套房贷利率基本在4.9%左右。
[4 ]7月20日，中国人民银行授权全国银行间同业拆借中心公布了最新一期贷款市场报价利率（LPR）：1年期LPR为3.55%，5年期以上LPR为4.20%，均与上个月持平。但西安房贷利率较上月小幅下行，西安多家银行首套房贷利率从4.1%降至4%。
今年6月，LPR时隔10个月迎来下调，1年期和5年期以上LPR均跟随政策利率下调10个基点。记者了解到，从6月下旬开始，西安各大银行相继落实首套房贷利率政策动态调整机制，进行利率调整。目前，西安地区六大行及招商银行、宁夏银行、北京银行、西安银行等十余家银行首套房贷利率降至4%，较LPR下浮0.2%。二套房贷利率同步下行，跌破5%大关，维持在4.9%。
```

Method - Outline-Enhanced Generator

- Supervised Fine-tune
  - two open-sourced web-enhanced long-form QA datasets available for training web-enhanced RAG models.
  - WebGLM-QA[1]
  - WebCPM[2]

question	answer	references
string · *lengths* 123⇔153 6.1%	string · *lengths* 683⇔968 31.9%	sequence
in football whats the point of wasting the first two plays with a rush - up the middle - not regular rush plays i get those	The point of wasting the first two plays with a rush up the middle is to take advantage of the 40 second play clock and the two minute warning in professional and college football games. By running the ball directly up the middle, the offense hopes to be able to block everyone up and the running back can make a person miss and end up in the end zone[2]. Running up the middle is also the shortest path to the end zone, so it is often designed to move defensive players out of the way[3]. Additionally, running the ball up the middle enough times will cause linebackers and defensive backs to run toward the line of scrimmage, creating the perfect opportunity to throw the ball over their	["As you can see from these highlights, most of these plays are designed to be run directly up the middle. The coach hopes everyone can be blocked up; the running back can make a person miss and ultimately end up in the end zone.", "In both professional and college football, the offense has 40 seconds from the end of the previous play to run the next play. A team running out the clock will allow the play clock (which records the time remaining until a play must be run) to drain as much as possible before running its next play. In the NFL, this is particularly noteworthy due to the existence of the two-minute warning. If the trailing team has no timeouts

Question:
麦田怪圈是什么？它们是如何形成的?

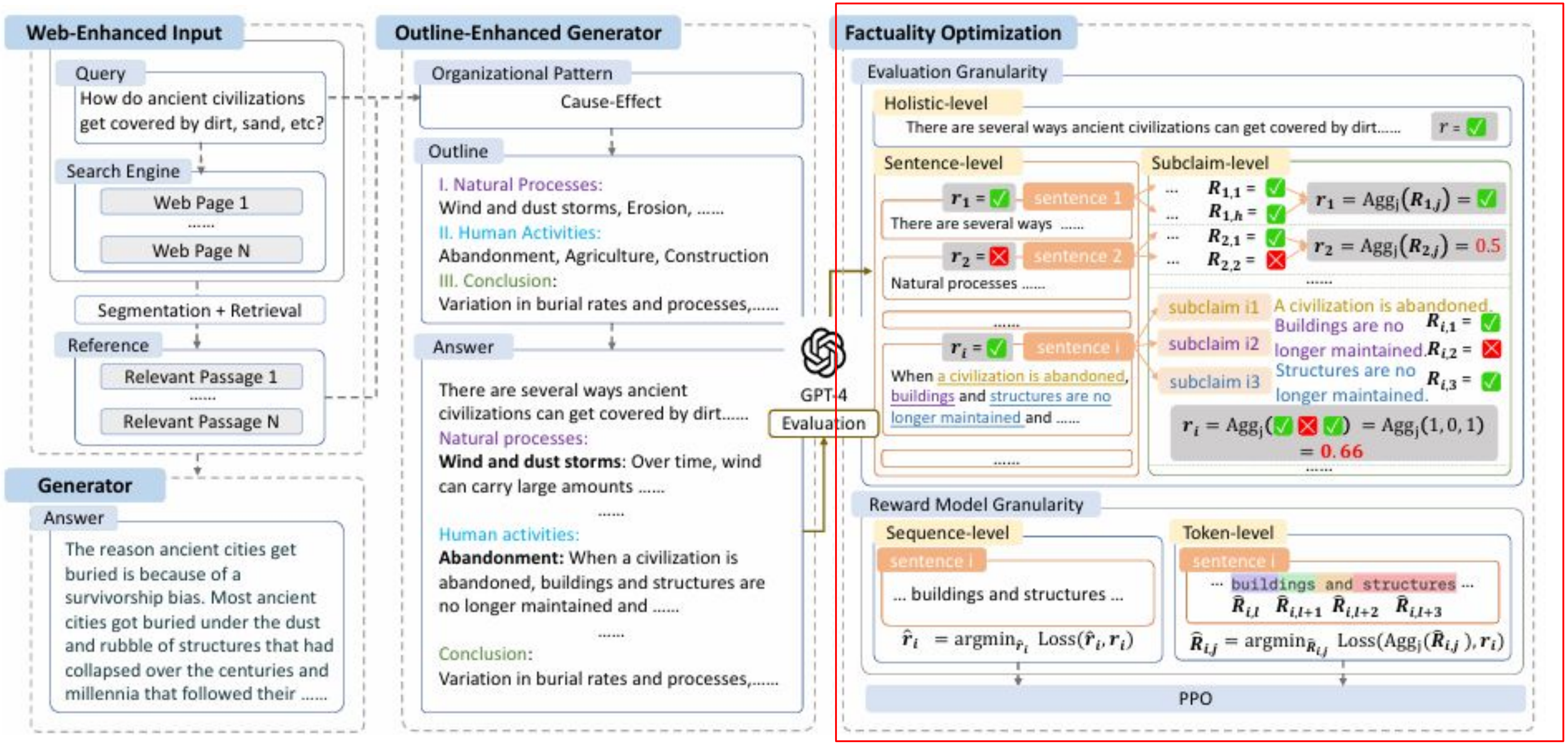
Translated Question:
What are crop circles? How are they made?

Human Action Sequence:
`Search→Load Page <1>→Quote→Scroll Down×5→Scroll Up→Scroll Down×11→Go Back→Search→Load Page <1>→Go Back→Load Page <3>→Scroll Down×4→Scroll Up×3→Quote→Scroll Down→Quote→Merge→Quote→Scroll Down→Quote→Finish`

Supporting Facts:
1. 麦田怪圈（Crop Circle），是指在麦田或其它田地上，通过某种未知力量（大多数怪圈是人类所为）把农作物压平而产生出来的几何图案。这个神秘现象有时被人们称之为“Crop Formation”。麦田怪圈的出现给了对支持外星人存在论的人们多种看法。
2. 人为说：人为说一般认为，麦田圈是用木板压成的。木板两头系上绳子形成圈套，在制作时，一脚踩在木板上拖动木板压倒麦子，并拉着细绳与圆心保持固定的距离，逐渐就可以形成一个圆圈。为了便于制造，主要形状所有圆圈的直径都可以被6除尽。以前曾经出现过制作麦田圈被当场抓获的事情，制作者使用的就是这种工具。
3. 自然形成说：也有人认为，麦田圈只是一种，成因还未被人类发现。就像雷电，古时候人类也是以为是雷神电母做的，对于麦田圈中经常出现人文信息的现象，他们认为这只是人们“先入为主”造成的错觉。
4. 磁场说：有专家认为，磁场中有一种神奇的移动力，可产生一股电流，使农作物“平躺”在地面上。美国专家杰弗里·威尔逊研究了130多个麦田怪圈，发现90%的怪圈附近都有连接高压电线的变压器，方圆270米内都有一个水池。由于接受灌溉，麦田底部的土壤释放出的离子会产生负电，与高压电线相连的变压器则产生正电，负电和正电碰撞后会产生电磁能，从而击倒小麦形成怪圈。

[1] Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. WebGLM: Towards An Efficient Web-Enhanced Question Answering System with Human Preferences. In Proceedings of the 29th ACMSIGKDDConference on Knowledge Discovery and Data Mining (, Long Beach, CA, USA,) (KDD '23). Association for Computing Machinery, New York, NY, USA, 4549–4560. https://doi.org/10.1145/3580305.3599931
[2] Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. WebCPM: Interactive Web Search for Chinese Long-form Question Answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 8968–8988.

Method - Factuality-Optimized RAG



Method - Doubly Fine-grained RLHF

- Holistic Level: standard granularity to evaluate the answers
- Sentence-level: segment the answer into sentences, then evaluate each sentence individually
- Subclaim-level: decompose each sentence into multiple subclaims via an LLM, each containing a single piece of factual information
- e.g. question: 什麼是黑洞？
- Holistic: 黑洞是一個質量極大的天體，能夠吸引一切物質，光也無法逃脫黑洞的引力。
- Sentence: ["黑洞是一個質量極大的天體，能夠吸引一切物質。"]
  ["光也無法逃脫黑洞的引力。"]
- Subcliam: ["光"],["無法逃脫"], ["黑洞的引力"]

[4]

Method - Reward & Loss

- 黑洞是一個質量極大的天體，能夠吸引一切物質，光也無法逃脫黑洞的引力。=> 1
  - Logloss
- ["光"],["無法逃脫"], ["黑洞的引力"] => [1, 0.7, 0.9]
  - MSE
- adopt PPO to optimize the generation model by maximizing the following reward

$$\hat{r}_t(s_t, a_t) = \sum_{j=1}^{L} \mathbf{1}(t = T_j)\hat{R}_\phi(\boldsymbol{a}|x, z)[j] - \beta \log \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}.$$

[4]

Experiment - Compare method

- WebGPT-13B、WebGPT-175B
- WebGLM-10B
- WebCPM-10B
- Our Method: fine-tuning on Llama2-7B-chat and ChatGLM2-6B
  - FoRAG-L7B(Ours)
  - FoRAG-C6B(Ours)

[4]

Experiment - Metrics

- coherence
- helpfulness
- factuality

你将获得针对某个问题编写的一个答案。
你的任务是根据一项指标对答案进行评分。
请你确保仔细阅读并理解这些说明。请在审阅时保持本文档处于打开状态，并根据需要进行参考。

评价标准：
连贯性(1-5) - 所有句子的集体质量。答案不应包含日期线、系统内部格式、大写错误或明显不合语法的句子（例如片段、缺少组件），以免文本难以阅读。答案中不应有不必要的重复。不必要的重复可能表现为重复整个句子或重复事实。答案应该结构良好、组织良好。答案不应该只是一堆相关信息，而应该逐句构建有关某个主题的连贯信息体。

你将获得针对某个问题编写的一个答案。
你的任务是根据一项指标对答案进行评分。
请你确保仔细阅读并理解这些说明。请在审阅时保持本文档处于打开状态，并根据需要进行参考。
评价标准：
帮助性(1-5) - 答案有效满足寻求信息的人的需求的程度。答案应该以易于理解的方式表达。答案应直接解决提出的问题，重点关注询问者寻求的具体信息或解决方案。所提供的信息应正确且基于事实、可验证的数据或公认的专业知识。答案包括完整回答问题的所有必要细节。它没有遗漏问题的所有关键方面。

我将向您展示一个问题、一系列文本片段和一份参考文档。所有的片段可以连起来形成对问题的完整回答。您的任务是在参考文档的帮助下评估每个文本片段是否包含事实错误。

请按照以下要求进行评估：
1. 如果文本片段仅包含类似“方法如下”、“根据资料可得”这样的通用开场白而没有传递具体信息，直接判定为“正确”。
2. 如果文本片段都能在参考文档或者问题中找到相应句子作为支持，或者可以从相应句子推理得到，直接判定为“正确”。关注关键词和细节的语义一致性。
3. 如果文本片段中有任何信息在参考文档或者问题中找不到明确的支撑，也不能根据相应句子推理得到，直接判定为“错误”。

[4]

Experiment - Results

Model	WebCPM (zh) Cohr.	WebCPM (zh) Help.	WebCPM (zh) Fact/q.	WebCPM (zh) Fact/s.	WebCPM (zh) Avg. Len.	WebGPT (en) Cohr.	WebGPT (en) Help.	WebGPT (en) Fact/q.	WebGPT (en) Fact/s.	WebGPT (en) 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

Model	Out. Enh.	Fac. Opt.	WebCPM (zh) Cohr.	WebCPM (zh) Help.	WebCPM (zh) Fact/q.	WebCPM (zh) Fact/s.	WebCPM (zh) Avg. Len.	WebGPT (en) Cohr.	WebGPT (en) Help.	WebGPT (en) Fact/q.	WebGPT (en) Fact/s.	WebGPT (en) Avg. Len.
FoRAG-C 6B	✗	✗	0.4598	0.6332	0.7613	0.9081	583	0.4081	0.7721	0.7868	0.9464	177
	✗	✓	0.4724	0.6407	0.8065	0.9395	585	0.5184	0.7868	0.8566	0.9763	181
	✓	✗	0.8643	0.7814	0.6055	0.9197	622	0.8566	0.8529	0.5993	0.9530	417
	✓	✓	0.8618	0.7764	0.7739	0.9639	655	0.8603	0.8640	0.7610	0.9804	443
FoRAG-L 7B	✗	✗	0.4296	0.6181	0.8090	0.8875	556	0.5221	0.8676	0.8750	0.9728	186
	✗	✓	0.4447	0.6256	**0.8618**	0.9394	570	0.5368	0.8860	**0.8970**	0.9818	189
	✓	✗	0.9095	**0.8668**	0.6583	0.9345	613	0.9816	0.9559	0.7978	0.9768	424
	✓	✓	**0.9121**	**0.8668**	0.8216	**0.9727**	625	**0.9889**	**0.9595**	0.8897	**0.9894**	447

Conclusion

- Outline-enhanced generator:

  An outline-enhanced generator is devised to ensure clear logic in long-form answers, and two corresponding datasets are constructed.

- Doubly fine-grained RLHF framework:

  A carefully designed doubly fine-grained RLHF framework is introduced to optimize the factuality of generated answers. This framework incorporates automatic evaluation and reward modeling at different levels of granularity

- FoRAG-L-7B model advantage:

  Applying FoRAG to Llama2-7B-chat, the resulting model, FoRAG-L-7B, outperforms WebGPT-175B while using only 1/24 of the parameters of WebGPT-175B.

Future work

- Doubly Fine-grained RLHF
- Small-size LM beats Large-size LM
- Outline-Enhanced Generator generate longer answer