

# Interactive Web Search for Chinese Long-form Question Answering

Anonymous ACL submission

## Abstract

Long-form question answering (LFQA) aims at answering complex, open-ended questions with detailed, paragraph-length responses. The de facto paradigm of LFQA necessitates two procedures: information retrieval, which searches for relevant supporting facts, and information synthesis, which integrates these facts into a coherent answer. In this paper, we introduce IWSQA, the first Chinese LFQA dataset. One unique feature of IWSQA is that its information retrieval is based on interactive web search, which engages with a search engine in real time. Following WebGPT (Nakano et al., 2021), we develop a web search interface. We recruit annotators to search for relevant information using our interface and then answer questions. Meanwhile, the web search behaviors of our annotators would be recorded. In total, we collect 5,500 high-quality question-answer pairs, together with 14,315 supporting facts and 121,330 web search actions. We fine-tune pre-trained language models to imitate human behaviors for web search and to generate answers based on the collected facts. Our LFQA pipeline, built on these fine-tuned models, generates answers that are no worse than human-written ones in 32.5% and 47.5% of the cases on our dataset and DuReader (He et al., 2018), respectively. The interface, dataset, and codes would be released publicly.

## 1 Introduction

Long-form question answering (LFQA) (Fan et al., 2019) targets answering complex, open-ended questions with detailed, paragraph-length responses. Current LFQA solutions generally follow the *retrieve-then-synthesize* paradigm, which comprises two core ingredients: information retrieval and information synthesis. The former searches external knowledge sources (e.g., the web) for diverse relevant supporting facts, and the latter integrates the collected facts into a coherent answer.

One defect of the conventional LFQA paradigm is that it often resorts to *non-interactive* retrieval methods, which use the original question as the query to retrieve a pile of uncurated information. On the contrary, humans are able to perform *interactive web search* by engaging with a search engine in real time. For a complex question, humans tend to decompose it into multiple sub-questions and ask them in sequence. By identifying and browsing relevant information, humans can improve their understanding of the topic and refine their searches by asking follow-up questions or related terms. This iterative process enables expanding the scope of their searches and improving the results they receive. Overall, interactive web search not only provides access to diverse information sources, but also reflects the cognitive process of how humans solve questions, which allows for better interpretability.

WebGPT (Nakano et al., 2021) is one pioneering work that supports interactive web search for LFQA. The authors first build a web search interface backed up by Microsoft Bing, then recruit annotators to collect information using the interface to answer questions. After that, they fine-tune GPT-3 (Brown et al., 2020) to imitate human behaviors for web search and to organize the collected information into answers. In the experiments, WebGPT shows exceptional ability in LFQA, even surpassing human experts. Despite its impressive performance, WebGPT still remains mysterious to the community. This is because WebGPT’s interface, dataset, and trained models are not publicly available, and the inner workings of its core design elements remain opaque. These factors make it hard for the community to understand the challenges of interactive web search for LFQA and to continue exploring this line of study.

In view of this, we deem it urgent to provide an accessible platform and public benchmark for this area. To this end, we first construct an **open-source** interface (Figure 1) to record web search behaviors

Resource	IWSQA (this work)	DuReader (He et al., 2018)	CMRC (Cui et al., 2019)	C <sup>3</sup> (Sun et al., 2020)	WebGPT (Nakano et al., 2021)	GopherCite (Menick et al., 2022)
Language?	ZH	ZH	ZH	ZH	EN	EN
Is Public?	✓	✓	✓	✓	✗	✗
Targets long-form QA?	✓	✗	✗	✗	✓	✓
Has free-form answer?	✓	✓	✗	✗	✓	✓
Has web search behavior?	✓	✗	✗	✗	✓	✗
Avg. question length	29.0	9.6	16.3	12.2	—	—
Avg. supporting fact length	512.5	187.3	495.5	116.9	—	—
Avg. answer length	244.4	104.9	17.0	5.5	—	—

Table 1: A comparison of our IWSQA to relevant datasets. “—” means the information is unknown. For the length statistics, we record the number of Chinese characters. **The data scale of IWSQA is comparable to WebGPT.**

when humans gather relevant information for long-form questions. In the interface, users can execute pre-defined actions to perform multiple rounds of searching and browsing. When finding relevant information on a web page, they can record it as a supporting fact. Meanwhile, their web-browsing behaviors will be recorded. After collecting enough information, users can finish the web search and answer the questions based on their collected facts.

Based on the interface, we choose Chinese as the testbed and construct IWSQA, **the first public** QA dataset that involves interactive web search, and also **the first** Chinese LFQA dataset. IWSQA contains 5,500 high-quality question-answer pairs, together with 14,315 supporting facts and 121,330 web search actions. Table 1 summarizes the difference between IWSQA and relevant QA datasets. Among existing Chinese QA datasets, IWSQA possesses the longest question, supporting fact, and answer, which shows the complexity of the questions and the richness of the annotated answers.

Then we propose a general framework consisting of (1) a *search model*, which imitates human web search behaviors for information retrieval. Specifically, the search model comprises three modules to execute a series of pre-defined actions on our interface: an action prediction module, a search query generation module, and a supporting fact extraction module; (2) a *synthesis model*, which generates a coherent answer conditioned on the collected facts.

In the experiments, we choose 8 representative pre-trained language models (PLMs) with up to 10B parameter size, and evaluate their ability of interactive web search and information synthesis. We find that scaling model sizes is critical to achieving better performance. By selecting the best-performing backbone PLM for the search and synthesis model, we combine them into a holistic LFQA pipeline and compare its capability with humans. Human evaluation reveals that our pipeline generates answers that are no worse than humans

32.5% of the time on our test set. When applied to questions whose annotated answers are longer than 400 Chinese characters from DuReader (He et al., 2018), our pipeline generates answers that are better than golden annotated ones 47.5% of the cases. We also show that our search model surpasses the conventional non-interactive retrieval method. Finally, we analyze the contribution of core design elements of our framework and the human-like behaviors our models acquire. We would open-source our interface, dataset, codes, and trained models. We also envision these resources to serve as the testbed for other research topics, such as behavior cloning (Bain and Sammut, 1995) and planning in interactive environments (Huang et al., 2022a).

## 2 Related Work

**Retrieval and Synthesis in LFQA.** For information retrieval, prior works generally resort to local repositories (e.g., Wikipedia). Recently there is a surge of interest in leveraging the whole web as the knowledge source (Nakano et al., 2021; Lazaridou et al., 2022; Menick et al., 2022; Thoppilan et al., 2022), which not only widens the scope of information sources but enables real-time coverage of up-to-date knowledge. On the other hand, how to structure the retrieved facts into a plausible and nuanced answer for LFQA is still under-explored. Some investigated how humans craft complicated answers, either by studying the functional structures of long-form answers (Xu et al., 2022) or exploring how to compose exemplification in answers (Wang et al., 2022); others revisit existing evaluation metrics of LFQA (Krishna et al., 2021).

**Comparison with WebGPT.** Our main contribution is to provide an open-source interface, benchmark, and model implementation to facilitate relevant research. When constructing the benchmark, we largely follow WebGPT and also propose improved design elements (with details elaborated

Action Name	Functionality
Q Search <query>	Call Bing search with <query>
← Go Back	Return to the previous window
Load Page <1>	Load the details of page <1>
Load Page <2>	Load the details of page <2>
Load Page <3>	Load the details of page <3>
↑ Scroll Up	Scroll up for a pre-set stride
↓ Scroll Down	Scroll down for a pre-set stride
„ Quote <content>	Extract <content> from the current page as a supporting fact
↑ Merge	Merge two facts into a single fact
Finish	End the search process

Figure 1: Left: an example screenshot of our interface in the *search mode*. Right: the actions our interface supports.

in appendix E), including (1) *interface*: we modify the actions defined by WebGPT to make them easier for model learning and more user-friendly; (2) *framework*: we decompose web search into 3 sub-tasks and implement a modular search model, which demonstrates better performance than WebGPT’s multi-task training (appendix F.3). We additionally explore how to teach the synthesis model to ignore irrelevant facts (§ 6.3) and generate novel contents (appendix F.1); (3) *evaluation and analysis*: besides evaluating **the whole** pipeline following WebGPT (§ 6.2), we also evaluate each **individual** module (§ 6.1 and § 6.3). This fine-grained evaluation helps us better understand the contribution of core design elements of our framework and the human behaviors learned by our model.

**Interactive Decision Making with PLMs.** Besides traditional linguistic NLP tasks, recent research demonstrates PLMs with promising capabilities of decision making in complex interactive environments, such as planning in robotic tasks (Huang et al., 2022a; Ahn et al., 2022; Huang et al., 2022b), manipulating search engines (Nakano et al., 2021), shopping on e-commerce websites (Yao et al., 2022), etc. By harnessing the rich world knowledge learned during pre-training, PLMs can perform grounded actions to interact with the real world. We envision our benchmark to serve as the testbed for future explorations in this area.

### 3 Web Search Environment

Following WebGPT, we construct a text-only interface to record web search behaviors when humans gather relevant information for long-form questions. Our interface, backed up by **Bing Search**

**API**, supports 10 mainstream web search actions as shown in Figure 1. When an action is executed, our interface responds with changes in the window.

When the action **Search** is performed, the interface enters *search mode* (Figure 1), which displays the links recommended by Bing for a specific query *<query>*. Each link comprises a title and a brief snapshot of the specific web page. Each window displays three links one time, and more links can be accessed by executing the **Scroll Down** action.

When finding the *i*-th link in the current window to be relevant, users could execute the **Load Page <i>** action ( $i \in \{1, 2, 3\}$ ). The interface would enter the *browsing mode* (Figure 5 in the appendix) and render the texts cleaned from the HTML of the *<i>*-th web page. The content users could view at a time in the window is restricted up to 500 Chinese characters, and more content can be accessed with the **Scroll** action. Users can utilize the **Quote** action to extract consecutive sentences in the current window as a supporting fact. To enable extracting texts that stretch across two windows, the **Merge** action is designed to merge the last two facts into a single fact (see appendix A.2 for more details). We also display all the existing extracted supporting facts for users.

After browsing the *i*-th page, users can return to the previous *search mode* using the **Go Back** action to access other links. Meanwhile, a refined query can be sent at any time. In general, users can freely interact with our interface multiple times until executing the **Finish** action or triggering the maximum number of actions (100 in our case). The interface would automatically record meaningful actions and observations during web search. Owing

232 to the multilingual nature of Bing system, although  
233 this work focuses on Chinese, **our interface can be**  
234 **flexibly adapted to other languages as well.** For  
235 more technical details, please refer to appendix A.

## 236 4 Data Collection

237 We employ 23 **annotators** from different walks of  
238 life, who are experienced in search engine operation  
239 and at least have an undergraduate degree. We  
240 ask them to answer long-form questions by first  
241 searching for relevant information using our inter-  
242 face, then writing a nuanced answer. For quality  
243 control, we recruit 8 experts familiar with QA re-  
244 search as **quality inspectors**. Next, we introduce  
245 the construction process of our dataset, with de-  
246 tailed annotation guides left in appendix B.

247 **Question Creation.** Creating new long-form  
248 questions from scratch without any reference is  
249 counterproductive, thus we turn to public QA fo-  
250 rumns as the question source. However, for most  
251 of the long-form questions on Chinese QA forums,  
252 such as [Zhihu](#) and [Baidu Zhidao](#), users can eas-  
253 ily access a variety of well-organized answers by  
254 directly querying the original question. In view  
255 of this, we instead engage annotators to refer to  
256 the questions on an English QA forum [Reddit](#), and  
257 then create new questions written in Chinese. The  
258 details of this creation process (including quality  
259 control) are elaborated in appendix C. We find em-  
260 pirically that questions created in this way often  
261 necessitate multiple rounds of searching and brows-  
262 ing to collect sufficient information.

263 **Interactive Web Search.** Given a question, we  
264 ask annotators to search for accurate and relevant  
265 information from trusted sources using our inter-  
266 face. This process may involve sending refined  
267 queries to Bing multiple times, as well as explor-  
268 ing various web pages they deem to be relevant.  
269 We require annotators to carefully judge the fac-  
270 tual accuracy of the information before extracting  
271 it as a supporting fact. The search process would  
272 be finished until sufficient supporting facts are col-  
273 lected. Among our created questions, 26.2% are  
274 unanswerable and finally discarded because anno-  
275 tators cannot find sufficient useful information.

276 **Answer Annotation.** After gathering enough  
277 supporting facts, the annotators would write self-  
278 contained answers based on their collected infor-  
279 mation. We give them instructions for answer anno-  
280 tation, including writing answers that are relevant

281 to the question and have rich content, main-  
282 taining logical consistency, clarity, and coherence, and  
283 providing viewpoints in an unbiased manner.

284 **Quality Control.** Each annotated instance is at  
285 least checked and approved by 2 quality inspectors  
286 before being selected into the final dataset. First,  
287 inspectors would manually inspect the action se-  
288 quences recorded on the interface and discard those  
289 with low quality (e.g., those with evident clerical  
290 errors in the issued queries). Second, they would  
291 carefully check the collected supporting facts. If  
292 these facts are apparently insufficient to answer the  
293 question, irrelevant to the question, or factually in-  
294 correct, the corresponding action sequence would  
295 be abandoned. The above procedures remove 25%  
296 collected instances. For the remaining instances,  
297 inspectors would carefully examine their annotated  
298 answers. If an answer contradicts our instructions  
299 mentioned above, inspectors would return it to an-  
300 notators and point out which requirement is not  
301 satisfied. Annotators would revise their answers  
302 possibly for multiple rounds until the revised an-  
303 swer is up to standard. We also sample 400 in-  
304 stances from the final dataset for a spot-checking  
305 and find that 98% entirely meets our criteria.

306 **Dataset Statistics.** Ultimately, we collect 5,500  
307 instances, each formatted in a tuple of (*question*,  
308 *web search behavior*, *supporting fact*, *answer*), and  
309 also record the observations at each action execu-  
310 tion. **We display a few examples in appendix G**  
311 **for reference.** We partition the dataset into {4, 700,  
312 400, 400} as the training, development, and test  
313 set. On average, each question involves performing  
314 22.1 actions, sending 2.6 queries, and loading 3.2  
315 web pages. The detailed proportion of each action  
316 is visualized in Figure 6 in the appendix.

## 317 5 Framework

318 In this section, we introduce how to teach PLMs  
319 for (1) interactive web search using our interface  
320 (§ 5.1) and (2) information synthesis (§ 5.2). The  
321 overall framework is illustrated in Figure 2.

### 322 5.1 Search Model

323 **Overview.** We partition web search into 3 sub-  
324 tasks: action prediction, search query generation,  
325 and supporting fact extraction. Each task is cast  
326 as a text-to-text format and we train 3 separate  
327 modules using a generative PLM. By combining  
328 the 3 modules, we build the search model, which

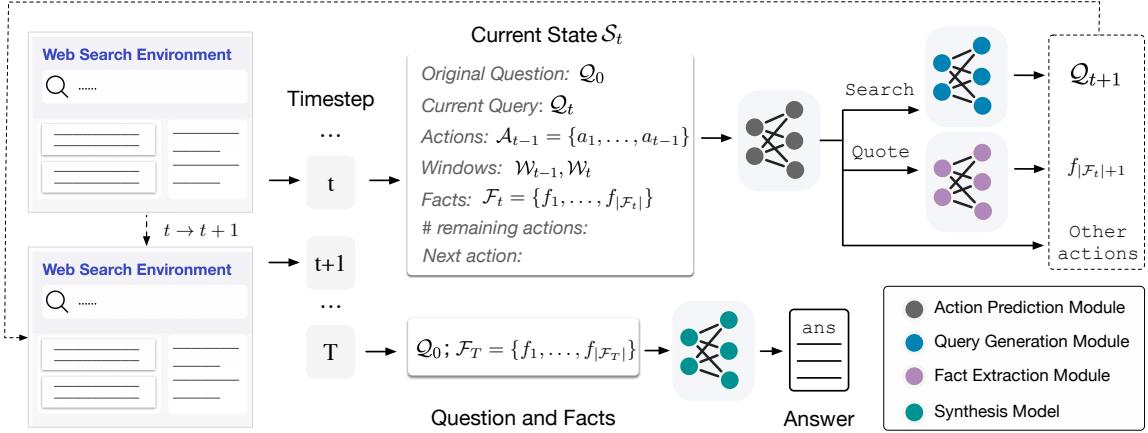


Figure 2: Illustration of our LFQA framework. For an action sequence of  $T$  steps, the search model (consisting of 3 modules) executes actions to collect supporting facts, which are sent to the synthesis model for answer generation.

executes a series of actions to gather relevant information. The action prediction module decides which action to perform at each step. If the module predicts `Search` or `Quote` as the current action, then it calls the other two modules to generate the contents of the query or the supporting fact.

Each module performs inference conditioned on the current state  $S_t$  of the interface at time step  $t$ .  $S_t$  comprises the original question  $Q_0$ , the query currently searching  $Q_t$ , the past action sequence  $A_{t-1} = \{a_1, \dots, a_{t-1}\}$ , the last and the current content displayed in the window  $W_{t-1}$  and  $W_t$ , current supporting facts  $F_t = \{f_1, \dots, f_{|F_t|}\}$ , and the number of remaining actions. If an action is executed, the components of  $S_t$  would be updated.  $W$  can be either the three links in the *search mode* or the specific page content in the *browsing mode*. We only maintain the recent two observations ( $W_{t-1}$  and  $W_t$ ) displayed in the window instead of concatenating all the past observations because the latter may exceed the input length limit of the PLM. Next, we introduce the three modules in detail.

**Action Prediction.** This module predicts which action to perform next. Since there are 10 possible actions in total, action prediction can be viewed as a 10-category classification task. Take the action `Search` as an example, denote  $\{x_1, \dots, x_N\}$  as the tokenized sequence for the action name `Search`, where  $x_*$  denotes a specific token. The probability of `Search` can be factorized as follows:

$$\mathcal{P}(\text{Search}|S_t) = \mathcal{P}(x_1|S_t) \times \prod_{i=2}^N \mathcal{P}(x_i|x_1, \dots, x_{i-1}).$$

During inference, we select the action with the highest probability to perform on the interface.

**Search Query Generation.** This module generates a query  $Q_{t+1} = \{q_1, \dots, q_{|Q_{t+1}|}\}$  to search Bing, which is also formulated as text generation:

$$\mathcal{P}(Q_{t+1}|S_t) = \mathcal{P}(q_1|S_t) \times \prod_{i=2}^{|Q_{t+1}|} \mathcal{P}(q_i|S_t, q_1, \dots, q_{i-1}).$$

**Supporting Fact Extraction.** Assume in the *browsing mode*, the current content of the window is  $W_t = \{w_1, \dots, w_{|W_t|}\}$ . We aim to extract a supporting fact  $f = \{w_i, \dots, w_j\}$  from  $W_t$ , where  $1 \leq i \leq j \leq |W_t|$ . While a naive solution is to directly generate all the tokens of  $f$  auto-regressively, this solution suffers from low inference speed in practice. As an alternative, we only generate the first and last few ( $N_f$ ) tokens of  $f$  given  $S_t$ . We compare the above two output formats in appendix F.2. Formally, we maximize  $\mathcal{P}([\text{s}], w_i, \dots, w_{i+N_f-1}, [\text{e}], w_{j-N_f+1}, \dots, w_j|S_t)$ , where  $[\text{s}]$  and  $[\text{e}]$  denote the special tokens that indicate the start and end of the fact  $f$ . During inference, after decoding the start and end tokens, we can locate the desired sequence in  $W_t$  by text matching. If the start / end tokens occur in multiple locations of  $W_t$ , we always extract the longest sequence from  $W_t$ , and a large  $N_f$  could lower the frequency of this multi-location issue. Note disjoint spans in  $W_t$  can be extracted by executing multiple `Quote` actions consecutively.

## 5.2 Synthesis Model

The **information synthesis** task learns to organize a series of supporting facts into a coherent answer. However, not as perfect as humans, the trained search model occasionally gathers irrelevant noises, which would influence the quality of the generated

answer. To remedy this, we corrupt the collected facts in the training data of the synthesis model by introducing noises. Specifically, given a series of human-extracted facts  $\{f_1, \dots, f_N\}$ , we randomly select a few unrelated facts  $\{f'_1, \dots, f'_{N'}\}$  from other training instances. After randomly shuffling all the facts, we concatenate them as the final input. During training, the model is optimized to generate the human-annotated answer conditioned on the corrupted supporting facts, i.e., maximizing  $\mathcal{P}(\text{Answer}|\mathcal{Q}_0, f_1, \dots, f_N, f'_1, \dots, f'_{N'})$ . Since the annotated answer does not contain the information of  $f'_*$ , the model learns to ignore irrelevant facts and only focus on important ones for generation.

## 6 Experiments and Analyses

Our problem consists of 4 sub-tasks: action prediction, search query generation, supporting fact extraction, and information synthesis. Correspondingly, we first train 4 modules and evaluate each sub-task independently by feeding the ground truth input to each module (§ 6.1). Then we combine all modules into a unitary pipeline and only feed the question to the pipeline for a holistic evaluation (§ 6.2). Finally, we conduct in-depth analyses for each module to understand their behaviors (§ 6.3).

### 6.1 Individual Sub-task Evaluation

**Settings.** We evaluate 8 typical generative PLMs that support Chinese, covering 3 architectures:

- T5 architecture (Raffel et al., 2019): **mT5<sub>BASE</sub>** (Xue et al., 2021), a 580M model pre-trained on mC4; **mT0<sub>BASE</sub>** (Muennighoff et al., 2022), which fine-tunes **mT5<sub>BASE</sub>** on diverse downstream tasks; **Mengzi-T5<sub>BASE</sub>** (Zhang et al., 2021b), a 220M model pre-trained on 300G internet corpora.
- BART architecture (Lewis et al., 2020): **mBART<sub>LARGE</sub>** (Liu et al., 2020), a 680M model pre-trained on monolingual corpora of multiple languages; **C-BART<sub>LARGE</sub>** (Shao et al., 2021), a 406M model pre-trained on 200G web texts.
- CPM architecture (Zhang et al., 2021a): **CPM<sub>2.6B</sub>**, **CPM<sub>7B</sub>**, and **CPM<sub>10B</sub>**, which contain 2.6B, 7B, and 10B parameters, respectively, and are pre-trained with increasing sizes of data.

Among these PLMs, **mT5<sub>BASE</sub>**, **mT0<sub>BASE</sub>**, and **mBART<sub>LARGE</sub>** are multilingual and the others are Chinese-only PLMs. We elaborate on details of the above PLMs in appendix D. We adopt recommended fine-tuning configurations of the original

Task Metric	Action	Query	Fact	Synth.
	<i>Mi.</i>	<i>Ma.</i>	<i>R-L</i>	<i>R-L</i>
<b>mT5<sub>BASE</sub></b>	53.8	44.0	62.4	56.7
<b>mT0<sub>BASE</sub></b>	58.2	52.1	64.6	60.0
<b>Mengzi-T5<sub>BASE</sub></b>	58.1	51.2	62.6	61.9
<b>mBART<sub>LARGE</sub></b>	53.6	41.1	50.4	56.5
<b>C-BART<sub>LARGE</sub></b>	43.8	31.3	56.1	49.3
<b>CPM<sub>2.6B</sub></b>	55.6	49.8	61.6	52.6
<b>CPM<sub>7B</sub></b>	58.9	50.5	67.8	59.8
<b>CPM<sub>10B</sub></b>	<b>60.4</b>	<b>54.5</b>	<b>70.0</b>	<b>62.4</b>
				<b>61.2</b>

Table 2: Sub-task evaluation (test performance) using 8 PLMs. We report *Micro-F1* (*Mi.*), *Macro-F1* (*Ma.*) for action prediction, and *Rouge-L* (*R-L*) for query generation, fact extraction, and information synthesis.

papers for all PLMs. For evaluation metrics, we treat action prediction as a 10-category classification task and choose *Micro-F1* and *Macro-F1* as the metric. We treat the other three tasks as text generation and calculate *Rouge-L* of the generated sequence and the ground truth.

**Results.** The results are listed in Table 2, from which we conclude that: (1) **mT0<sub>BASE</sub>** outperforms **mT5<sub>BASE</sub>** in action prediction, query generation, and supporting fact extraction, but performs poorer in information synthesis. We conjecture this is because **mT0<sub>BASE</sub>** enhances language skills more related to the first three tasks during its multi-task fine-tuning. Rather, the information synthesis ability might have been weakened. Besides, **Mengzi-T5<sub>BASE</sub>** performs generally well on all tasks despite owning much fewer parameters; (2) in general, **mBART<sub>LARGE</sub>** and **C-BART<sub>LARGE</sub>** show inferior performance than all other PLMs, except that **mBART<sub>LARGE</sub>** exhibits excellent performance in information synthesis; (3) comparing the results of **CPM<sub>2.6B</sub>**, **CPM<sub>7B</sub>**, and **CPM<sub>10B</sub>**, we find that **the performance generally gets improved as the model size increases**. Blessed by the scaling law (Kaplan et al., 2020), larger PLMs own stronger understanding and generation abilities and could achieve better downstream performance.

### 6.2 Holistic Pipeline Evaluation

We choose the modules trained by **CPM<sub>10B</sub>**, which performs the best among all the PLMs in § 6.1, and combine them into the overall pipeline. Then we evaluate its performance compared with humans.

**Compared Answer Pairs.** For each test question of IWSQA, we compare the annotated answer with 3 types of answers generated by our synthesis model. Specifically, the 3 types of answers differ

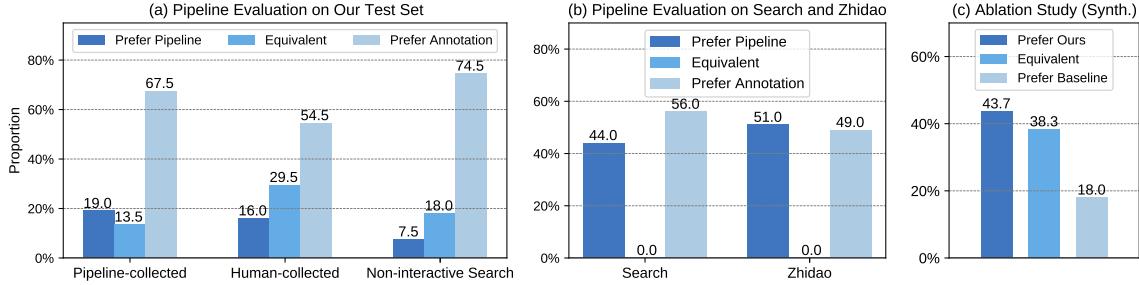


Figure 3: Results of human evaluation. (a) Comparison of human annotation and our pipeline-generated answers with different sources of supporting facts. (b) The experiments on two DuReader datasets: Search and Zhidao. We compare our pipeline with the golden annotation. (c) Ablation study for our synthesis model.

in the source of supporting facts, including (1) the facts collected by our search model, (2) ground-truth human-collected facts, and (3) the facts collected using a commonly adopted non-interactive web search method. For (3), we directly input the original question into Bing, extract the paragraphs from all the retrieved links, and rank them using TF-IDF. Then we concatenate the top- $k$  paragraphs as the input until it exceeds **CPM<sub>10B</sub>**'s maximum sequence length (3072 tokens).

**Evaluation Protocol.** We engage 8 annotators to manually compare different answers based on human preference. Given a question and a pair of answers, we ask them to perform an overall assessment and decide which answer they would prefer based on multiple factors, including the overall usefulness, coherence, and relevance to the question. Since all three retrieval methods use the same search engine, their collected facts sometimes have a high overlap, which leads to similar answers. Thus we allow annotators to mark two answers as *equivalent* if both are of comparable quality.

**Results.** We derive from the results in Figure 3 (a) that: (1) the answers obtained purely by our pipeline are preferred or comparable to human-written answers  $19.0\% + 13.5\% = 32.5\%$  of the time. This result implies ample opportunity for advancement of our pipeline in future endeavors, which is discussed in appendix I. (2) When applying our synthesis model to the human-collected facts, the performance grows to  $16.0\% + 29.5\% = 45.5\%$  preference or equivalence, which is due to the improved quality of the collected facts. (3) The facts gathered by non-interactive search lead to slightly worse performance ( $7.5\% + 18\% = 25.5\%$ ) than our search model. The **superiority of our search model over non-interactive search** may be because our model

(a) sends diverse queries to Bing multiple times so that more abundant information can be retrieved, and (b) it critically decides whether a web page contains important information, which performs better than TF-IDF. In appendix H, we display several answers generated by our pipeline.

**Experiments on DuReader.** Next, we apply our pipeline (search model and synthesis model) to 2 Chinese QA datasets from DuReader, i.e., Zhidao and Search. Although not specially designed for LFQA, DuReader contains a variety of question types, and we randomly sample 400 test questions whose annotated answers are longer than 400 Chinese characters. For these questions, we engage annotators to compare our pipeline-generated answers with the golden annotations of DuReader. From the results in Figure 3 (b), we find that our pipeline generates answers better than the annotated ones 44.0% and 51.0% of the time on Search and Zhidao (47.5% on average). The fact that the same pipeline surpasses fewer human-written answers on our dataset than DuReader also reflects the **high quality of our annotated answers**. Note the *equivalent* ratio is 0% because both answers are based on totally different supporting facts, and it is easy to determine which one is better.

### 6.3 Further Analysis

Next, we conduct in-depth analyses to gain a deeper understanding of each module. Without loss of generality, we evaluate **CPM<sub>7B</sub>** in this section.

**Ablation Study for the Synthesis Model.** We evaluate whether corrupting the synthesis model's training data by introducing irrelevant facts improves its ability to ignore noisy facts. We train a baseline model without corrupting the training data and keep other settings the same as our model. For

479  
480  
481  
482  
483  
484  
485  
486  
487  
488

489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500

501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515

516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551

Task Metric	Action Mi.	Fact Ma.	Task R-L	Query R-L
$\mathcal{S}_t$	58.9	50.5	59.8	$\mathcal{S}_t$
- $\mathcal{F}_t$	55.5	49.3	54.7	- $\mathcal{F}_t$
- $\mathcal{W}_{t-1}$	57.7	52.0	59.3	- past queries $\in \mathcal{A}_{t-1}$
- $\mathcal{A}_{t-1}$	53.4	44.1	60.3	- seen titles $\in \mathcal{A}_{t-1}$

Table 3: Ablation study of the search model when different components are removed from  $\mathcal{S}_t$ , respectively.

each test question, we feed the supporting facts collected by our search model to both synthesis models and generate two answers. Annotators would evaluate which answer is more relevant to the original question (the *equivalent* option is allowed).

According to Figure 3 (c), by corrupting the training data, our model performs better than the baseline 43.7% of the time and is worse 18.0% of the cases. This demonstrates that **our method indeed enhances the model’s ability to ignore noisy information**, which makes the generated answer more relevant to the original question. In appendix F.1, we further explore the use of another corruption method that flexibly balances generating novel contents and copying supporting facts.

**Effects of Components in  $\mathcal{S}_t$ .** We conduct ablation studies for several components of  $\mathcal{S}_t$  to examine how they contribute to each module of the search model. This is achieved by modifying both the training and evaluation data of each module. For action prediction and supporting fact extraction, we remove one of the following: the existing collected facts  $\mathcal{F}_t$ , the contents displayed in the last window  $\mathcal{W}_{t-1}$ , or the past actions  $\mathcal{A}_{t-1}$ . For query generation, the following items are removed from  $\mathcal{S}_t$ : the existing collected facts  $\mathcal{F}_t$ , the already searched queries, or the titles of the links browsed before. The information of the latter two items is included in  $\mathcal{A}_{t-1}$ . Specifically, for the past action Search / Load Page,  $\mathcal{A}_{t-1}$  not only includes the action name, but also records the specific searched query / the title of the loaded page.

The results are listed in Table 3, from which we observe that: (1) for action prediction, the removal of either  $\mathcal{F}_t$  or  $\mathcal{W}_{t-1}$  only leads to minimal performance changes, while removing  $\mathcal{A}_{t-1}$  leads to a significant performance drop. This shows that **the past actions are critical factors for action prediction**; (2) for supporting fact extraction, only removing  $\mathcal{F}_t$  impairs the performance significantly ( $-5.1$ ). This indicates that aligned with humans, **the module considers what has been extracted to**

Question and Generated Query	
<b>计算机科学与电气工程有什么区别？优缺点是什么？</b> What’s the difference between computer science and electrical engineering? What are their advantages and disadvantages?	
[1] <b>计算机科学与工程</b> Computer science and engineering [2] <b>电气工程是什么</b> What is electrical engineering [3] <b>计算机科学的优缺点</b> Advantages and disadvantages of computer science	
<b>为什么在你累的时候揉眼睛感觉这么好？</b> Why does it feel good to rub your eyes when you’re tired?	
[1] <b>为什么在你累的时候揉眼睛感觉这么好？</b> Why does it feel good to rub your eyes when you’re tired? [2] <b>累的时候揉眼睛</b> Rubbing your eyes when you’re tired [3] <b>眼睛酸了累了手揉一揉就缓解</b> Rubbing your eyes relieves fatigue when your eyes are tired and sore [4] <b>眼保健操缓解眼睛疲劳</b> Eyecare exercise relieves eye fatigue	

Figure 4: Case study for query generation. We show the generated queries for two test questions.

**decide which information to extract next;** (3) for query generation, removing either searched queries or accessed link titles in  $\mathcal{A}_{t-1}$  causes a great negative impact ( $-2.5$ ), which means **the module might have learned to generate queries based on what has been searched and newly observed information during web search**. This feature is humanoid in that humans also consider both information to avoid sending repetitive queries and to ask follow-up questions about an accessed link.

**Case Study for Query Generation.** To fathom the human behaviors learned by our query module, we conduct a case study by sampling the generated queries for different questions in the test set. We illustrate two representative results in Figure 4 (more results in appendix F.4) to showcase the typical strategies learned by our query module, including copying the original question, decomposing the question into multiple sub-questions, rephrasing questions with related terms, etc. These strategies make the queries more diverse, which helps gather more abundant information from various sources.

## 7 Conclusion

In this paper, we construct a benchmark of interactive web search for Chinese long-form QA, together with an open-source interface. We decompose the task into 4 sub-tasks and design a modular pipeline. By fine-tuning representative PLMs, we conduct both an individual evaluation for each module and a holistic evaluation for the pipeline. In-depth analyses are carried out to understand the core design elements of our framework. We expect our interface, dataset, framework, and analyses to facilitate more future explorations in this area.

## 628 Limitations

629 There are some limitations not well addressed in  
630 this paper:

- 631 • Although we have improved several WebGPT’s  
632 design elements (with differences detailed in  
633 appendix E), our interface and framework still  
634 share similarities with WebGPT. However, we  
635 contend that our main contribution is to provide  
636 an open-source interface and a public bench-  
637 mark, which allow the community to study in-  
638 teractive web search for LFQA. The environ-  
639 ment we provide can also be flexibly adapted  
640 to building benchmarks and facilitating corre-  
641 sponding research in other languages. We deem  
642 the openness of our resources as a distinct and  
643 meaningful contribution to this field.
- 644 • The human evaluation shows that our pipeline  
645 performs worse than humans in the process of  
646 information retrieval and synthesis 67.5% of the  
647 time, which still leaves room for improvement  
648 (see appendix I for future works). However, the  
649 focus of this paper is not to present a state-of-the-  
650 art LFQA model, and we hope our benchmark  
651 could facilitate future works to design more so-  
652 phisticated algorithms. The fact that representa-  
653 tive Chinese PLMs with up to 10B parameters  
654 still exhibit inferior performance also reflects  
655 the difficulty of our dataset.
- 656 • Another promising direction pointed by We-  
657 bGPT is reinforcement learning from human  
658 feedback (RLHF) (Stiennon et al., 2020), which  
659 is not explored in this work. We contend that  
660 our goal is not to perform a thorough replica-  
661 tion study for WebGPT, but to build a public  
662 benchmark for the task of interactive web search.  
663 Our provided environment can be utilized by re-  
664 searchers to study RLHF in the future.

## 665 Ethical Statement

666 In this research, we adhere to the highest ethical  
667 standards and commit to making every effort to  
668 minimize any potential harm. Specifically:

- 669 • When creating our dataset, we have ensured that  
670 all data collected is obtained through legitimate  
671 and legal means. In addition, we have obtained  
672 the appropriate permissions and consent from  
673 all necessary parties.
- 674 • We have also taken steps to protect the privacy of  
675 individuals whose data is included in our dataset  
676 through de-identification during annotation.

- We are committed to eliminating bias, discrim-  
677 ination, or stereotypes during annotation by re-  
678 moving any suspect examples.
- We take the responsibility of open-sourcing the  
679 interface, dataset, codes, and trained models to  
680 the public. However, there are cases that these re-  
681 sources are maliciously used. For instance, our  
682 models may be utilized to generate responses  
683 without proper attribution of the information  
684 source, causing severe consequences. We would  
685 strive to ensure that they are used ethically and  
686 not for any malicious or harm-causing intent.

## 689 References

- Michael Ahn, Anthony Brohan, Noah Brown, Yev-  
gen Chebotar, Omar Cortes, Byron David, Chelsea  
Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol  
Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu,  
Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang,  
Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jes-  
month, Nikhil Joshi, Ryan Julian, Dmitry Kalash-  
nikov, Yuheng Kuang, Kuang-Huei Lee, Sergey  
Levine, Yao Lu, Linda Luu, Carolina Parada, Pe-  
ter Pastor, Jornell Quiambao, Kanishka Rao, Jarek  
Rettinghouse, Diego Reyes, Pierre Sermanet, Nico-  
las Sievers, Clayton Tan, Alexander Toshev, Vincent  
Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu,  
Mengyuan Yan, and Andy Zeng. 2022. [Do as i can](#)  
and [not as i say: Grounding language in robotic af-  
fordances](#). In *arXiv preprint arXiv:2204.01691*.

- Michael Bain and Claude Sammut. 1995. [A framework  
for behavioural cloning](#). In *Machine Intelligence 15*,  
pages 103–129.

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie  
Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind  
Neelakantan, Pranav Shyam, Girish Sastry, Amanda  
Askell, Sandhini Agarwal, Ariel Herbert-Voss,  
Gretchen Krueger, Tom Henighan, Rewon Child,  
Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu,  
Clemens Winter, Christopher Hesse, Mark Chen, Eric  
Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,  
Jack Clark, Christopher Berner, Sam McCandlish,  
Alec Radford, Ilya Sutskever, and Dario Amodei.  
2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

- Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng  
Chen, Wentao Ma, Shijin Wang, and Guoping Hu.  
2019. [A span-extraction dataset for Chinese ma-  
chine reading comprehension](#). In *Proceedings of the  
2019 Conference on Empirical Methods in Natu-  
ral Language Processing and the 9th International  
Joint Conference on Natural Language Processing  
(EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong,  
China. Association for Computational Linguistics.

733	Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. <b>ELI5: Long form question answering</b> . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 3558–3567, Florence, Italy. Association for Computational Linguistics.	788
734		789
735		790
736		791
737		792
738		793
739		
740	Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. <b>DuReader: a Chinese machine reading comprehension dataset from real-world applications</b> . In <i>Proceedings of the Workshop on Machine Reading for Question Answering</i> , pages 37–46, Melbourne, Australia. Association for Computational Linguistics.	
741		
742		
743		
744		
745		
746		
747		
748	Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. <b>Language models as zero-shot planners: Extracting actionable knowledge for embodied agents</b> . <i>arXiv preprint arXiv:2201.07207</i> .	
749		
750		
751		
752	Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022b. <b>Inner monologue: Embodied reasoning through planning with language models</b> . <i>arXiv preprint arXiv:2207.05608</i> .	
753		
754		
755		
756		
757		
758	Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. <b>Scaling laws for neural language models</b> . <i>arXiv preprint arXiv:2001.08361</i> .	
759		
760		
761		
762		
763	Diederik P. Kingma and Jimmy Ba. 2015. <b>Adam: A method for stochastic optimization</b> . In <i>3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings</i> .	
764		
765		
766		
767		
768	Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. <b>Hurdles to progress in long-form question answering</b> . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 4940–4957, Online. Association for Computational Linguistics.	
769		
770		
771		
772		
773		
774		
775	Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. <b>Internet-augmented language models through few-shot prompting for open-domain question answering</b> . <i>arXiv preprint arXiv:2203.05115</i> .	
776		
777		
778		
779		
780	Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. <b>BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension</b> . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 7871–7880, Online. Association for Computational Linguistics.	
781		
782		
783		
784		
785		
786		
787		
788	Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. <b>Multilingual denoising pre-training for neural machine translation</b> . <i>Transactions of the Association for Computational Linguistics</i> , 8:726–742.	
789		
790		
791		
792		
793		
794	Jacob Menick, Maja Trebach, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. <b>Teaching language models to support answers with verified quotes</b> . <i>arXiv preprint arXiv:2203.11147</i> .	
795		
796		
797		
798		
799		
800	Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. <b>Crosslingual generalization through multitask finetuning</b> .	
801		
802		
803		
804		
805		
806		
807		
808	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. <b>Webgpt: Browser-assisted question-answering with human feedback</b> . <i>arXiv preprint arXiv:2112.09332</i> .	
809		
810		
811		
812		
813		
814	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. <b>Exploring the limits of transfer learning with a unified text-to-text transformer</b> . <i>ArXiv preprint</i> , abs/1910.10683.	
815		
816		
817		
818		
819	Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. <b>Bloom: A 176b-parameter open-access multilingual language model</b> . <i>arXiv preprint arXiv:2211.05100</i> .	
820		
821		
822		
823		
824		
825	Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. <b>Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation</b> . <i>arXiv preprint arXiv:2109.05729</i> .	
826		
827		
828		
829		
830	Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. <b>Learning to summarize with human feedback</b> . <i>Advances in Neural Information Processing Systems</i> , 33:3008–3021.	
831		
832		
833		
834		
835		
836	Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. <b>Investigating prior knowledge for challenging Chinese machine reading comprehension</b> . <i>Transactions of the Association for Computational Linguistics</i> , 8:141–155.	
837		
838		
839		
840		
841	Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al.	
842		
843		

844        2022. **Lambda: Language models for dialog applica-**  
845        **tions.** *arXiv preprint arXiv:2201.08239*.

846        Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol  
847        Choi, and Mohit Iyyer. 2022. **Modeling exemplifi-**  
848        **cation in long-form question answering via retrieval.**  
849        In *Proceedings of the 2022 Conference of the North*  
850        *American Chapter of the Association for Compu-*  
851        *tational Linguistics: Human Language Technologies,*  
852        *pages 2079–2092, Seattle, United States. Association*  
853        *for Computational Linguistics.*

854        Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Di-  
855        nan, Kyunghyun Cho, and Jason Weston. 2019. **Neu-**  
856        **ral text generation with unlikelihood training.** *arXiv*  
857        *preprint arXiv:1908.04319*.

858        Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022.  
859        **How do we answer complex questions: Discourse**  
860        **structure of long-form answers.** In *Proceedings of the*  
861        *60th Annual Meeting of the Association for Compu-*  
862        *tational Linguistics (Volume 1: Long Papers), pages*  
863        *3556–3572, Dublin, Ireland. Association for Compu-*  
864        *tational Linguistics.*

865        Linting Xue, Noah Constant, Adam Roberts, Mihir Kale,  
866        Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and  
867        Colin Raffel. 2021. **mT5: A massively multilingual**  
868        **pre-trained text-to-text transformer.** In *Proceedings*  
869        *of the 2021 Conference of the North American Chap-*  
870        *ter of the Association for Computational Linguistics:*  
871        *Human Language Technologies, pages 483–498, On-*  
872        *line. Association for Computational Linguistics.*

873        Shunyu Yao, Howard Chen, John Yang, and Karthik  
874        Narasimhan. 2022. **Webshop: Towards scalable real-**  
875        **world web interaction with grounded language agents.**  
876        *arXiv preprint arXiv:2207.01206*.

877        Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang,  
878        Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu,  
879        Wendi Zheng, Xiao Xia, et al. 2022. **Glm-130b:**  
880        **An open bilingual pre-trained model.** *arXiv preprint*  
881        *arXiv:2210.02414*.

882        Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian  
883        Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe  
884        Ji, Jian Guan, et al. 2021a. **Cpm: A large-scale**  
885        **generative chinese pre-trained language model.** *AI*  
886        *Open*, 2:93–99.

887        Zhuosheng Zhang, Hanqing Zhang, Keming Chen,  
888        Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming  
889        Zhou. 2021b. **Mengzi: Towards lightweight yet inge-**  
890        **nious pre-trained models for chinese.** *arXiv preprint*  
891        *arXiv:2110.06696*.

## 892 Appendices

### 893 A Implementation Details of the Interface

894 Our interface includes two components: an API  
895 back end and a website front end.

#### 896 A.1 API Back End

897 The API backend implements three APIs with dif-  
898 ferent functions: (1) *search*, which receives queries  
899 from users and returns search results recommended  
900 by Bing; (2) *extract*, which receives a URL and  
901 returns the text-only contents of the corresponding  
902 web page; (3) *record*, which receives the actions  
903 conducted by agents and stores them in a database.

904 **Search API.** The search API is based on Bing  
905 API. When it receives keywords from users, it calls  
906 Bing API to search for relevant results and con-  
907 verts them into the format we specify. Each result  
908 consists of a title, the link to the page, and a brief  
909 summary of the page contents. To ensure the origi-  
910 nality of the answers generated during annotation,  
911 we have implemented a filter in the search API to  
912 exclude results from certain websites (e.g., Red-  
913 dit forums). This is necessary because some of  
914 the questions are sourced from websites that may  
915 appear in search results.

916 **Extract API.** The contents of web pages often  
917 include huge quantities of layout information and  
918 multimedia that is inappropriate to display directly  
919 to agents and is meaningless for our task. There-  
920 fore, we use a third-party tool<sup>1</sup> to extract the sim-  
921 plified text-only contents of web pages. This ensures  
922 that only clean and meaningful text will be pre-  
923 sented to the users.

924 **Record API.** Actions conducted by users are  
925 recorded in the website front end, when users finish  
926 the annotation process of a question, the front end  
927 will call this Record API, and the detailed action in-  
928 formation and meaningful observations during web  
929 search will be uploaded and stored in our database.

#### 930 A.2 Website Front End

931 The website front end is designed as a graphic user  
932 interface for human annotators, which supports  
933 two modes: the *search mode* and the *browsing*  
934 mode. Each time an action is performed, it will  
935 be recorded and the corresponding changes will be  
936 rendered in our website and displayed to the users.

<sup>1</sup><https://github.com/mozilla/readability>

937 **Window.** In the *search mode*, the window dis-  
938 plays the searched results returned by our API back  
939 end. We present at most three links at a time in  
940 each window, and the *Scroll* action can be used  
941 to access other links. In the *browsing mode*, when  
942 clicking a specific link, *Load Page* action is trig-  
943 gered and the front end will call the extract API  
944 and display the text-only contents of the web page.  
945 The length of content in each window is limited up  
946 to 500 Chinese characters, and the *Scroll* action  
947 can be used to access more content. In the main  
948 paper, we illustrate an example for the *search mode*  
949 of our interface, here we present the example for  
950 the *browsing mode* in Figure 5. In addition, we  
951 also display the existing supporting facts and the  
952 remaining number of actions for ease of human  
953 annotation.

954 **Actions.** Once an action is performed, we record  
955 the current state of the interface, which includes the  
956 content displayed in the window, the current query  
957 issued, the existing collected supporting facts, the  
958 remaining number of actions, etc. We also record  
959 the specific information about the current action,  
960 for instance, *Search <query>* includes the con-  
961 tent of the query, *Load Page <idx>* includes all  
962 the detailed information about a web page, and  
963 *Quote <content>* includes the consecutive  
964 sentences selected by the user.

965 It should be noted that the action *Merge* is  
966 specially designed for extracting a supporting  
967 fact that crosses the boundary of two windows  
968 in the *browsing mode*. For instance, the user  
969 can perform *Quote <content1>*, *Scroll*  
970 *Down*, *Quote <content2>*, and *Merge* to get  
971 one supporting fact, which is concatenated by both  
972 *content1* and *content2*. We do not support  
973 merging multiple supporting facts because such  
974 behavior is hard for model learning and less fre-  
975 quently used by our annotators in the pilot study.  
976 Instead, merging the last two supporting facts is  
977 easier to imitate by PLMs and is used more often.

978 Besides, we also implement (1) the *Undo* action,  
979 which supports revoking the last action performed,  
980 and (2) the *Reset* action, which terminates the  
981 current annotation and starts a new one. Both ac-  
982 tions will not be recorded since they do not belong  
983 to meaningful web search behaviors.

## 984 B Annotation Principle

985 Below we present the annotation principles for web  
986 search, supporting fact extraction, and question an-

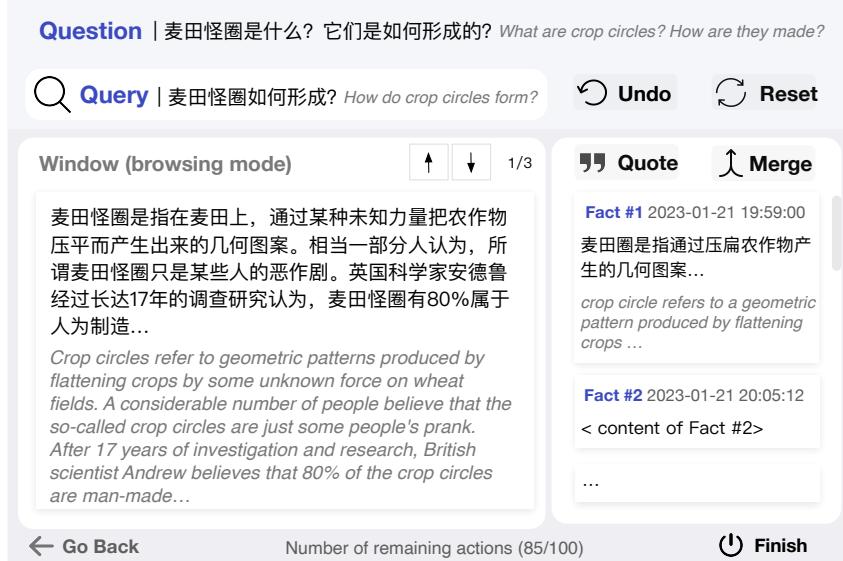


Figure 5: An example of *browsing mode* of our interface.

swering. These principles are part of our annotation guides, which are sent to our contractors before annotation. The original version of the following is written in Chinese, and we have translated it into English.

### B.1 Web Search Principle

**Look for Relevant Information.** In the search process, it is important to ensure that the content being searched is closely related to the question at hand. During the labeling process, users may encounter various concepts that are related to the question but may not be central to the main idea. These peripheral concepts should be ignored in the search. For instance, when searching for information about “the principle of the constant speed of light”, it is possible to come across the concept of “Lorentz transformation”, which is related to the topic but only tangentially. As such, it is not necessary to include a detailed explanation of “Lorentz transformation”.

**Send Simple Queries.** Search engines are often less effective when the question being asked is long and complex. In such cases, it is advisable to simplify and refine the main question or keywords to improve the chances of finding relevant information and reduce the number of unnecessary search actions. For example, instead of searching for the question “I have a question that bothers me a lot, why do most crustaceans / seafood turn from light gray to red / orange when heated?”, it would be more effective to simplify it to “why does seafood

change color when heated?”. This ensures the simplicity of the queries, making it more likely to find relevant information.

**Avoid Unnecessary Search.** Search engines typically rank web pages based on their relevance to the query, with higher-ranked results being more relevant. If the top-ranked results for a particular search do not align with the user’s needs, it may not be productive to continue scrolling through the results to find relevant information. Instead, it is more efficient to issue a new query to reduce the number of unnecessary search actions.

### B.2 Supporting Fact Extraction Principle

**Find Diverse Relevant Facts.** The supporting facts should contain information that is relevant to the original question. When possible, it is generally more effective to extract supporting facts from diverse sources, while ensuring that the content remains highly relevant to the original question. It is important to avoid duplicating summaries of the same content from different sources, as this does not contribute to answering the question.

**Avoid Recording Fragmentary Facts.** The extracted supporting fact should contain complete and coherent information. It is important to avoid intercepting sentences with incomplete semantics or taking them out of context, as this can alter the meaning of the supporting fact. In addition, please ensure the integrity of the supporting fact by including all relevant information and expressing it

987  
988  
989  
990  
991

992

993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018  
1019  
1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1048 in a coherent manner.

1049 **Ensure the Factual Accuracy.** It is important to  
1050 summarize information from trusted sources when-  
1051 ever possible. This helps ensure the reliability of  
1052 the information being used. You can also judge the  
1053 factual accuracy of a supporting fact by comparing  
1054 it with other searched results.

### 1055 **B.3 Answer Principle**

1056 A good long-form answer is typically well-  
1057 researched, well-written, and provides a thorough  
1058 and detailed response. It should be well-organized  
1059 and easy to read, with clear and concise language  
1060 that is appropriate for the intended audience. Ad-  
1061 ditionally, a good answer should be objective and  
1062 unbiased, presenting multiple viewpoints on the  
1063 topic if applicable.

1064 **Coherence and Relevance.** Coherence refers to  
1065 the overall logical consistency and clarity of the  
1066 answer. The desired answer should have a clear  
1067 structure, with each paragraph building upon the  
1068 previous one and contributing to the overall argu-  
1069 ment. The ideas presented should flow smoothly  
1070 and be easy to follow. Relevance means the extent  
1071 to which the answer addresses the original question.  
1072 The desired answer should stay on topic, providing  
1073 information that is directly relevant to the question.  
1074 It should not include unnecessary or tangential in-  
1075 formation. Together, coherence and relevance help  
1076 guarantee that the answer is easy to understand and  
1077 stays focused on the main topic, making it more  
1078 useful and informative for the reader.

1079 **Objectivity.** The content of the answer should  
1080 be based on the information obtained during the  
1081 search process. The desired answer should present  
1082 information and viewpoints in an unbiased man-  
1083 ner, without expressing personal opinions or pre-  
1084 ferences. While the annotation process inevitably  
1085 involves subjectivity, the questions are relatively  
1086 straightforward and it should not be difficult to  
1087 maintain a degree of objectivity. Please be neutral  
1088 and fair, and present multiple sides of an issue if  
1089 applicable.

1090 **Conciseness.** There is no specific word count  
1091 requirement for answers, but it is important to pro-  
1092 vide concise, comprehensive, and in-depth answers  
1093 that include necessary auxiliary information. It  
1094 is generally best to avoid extremely long or short  
1095 answers. In addition, the sentences in the answer

1096 should be concise and clear and should avoid re-  
1097 dundancy. For example, the question “How toxic is  
1098 barium chloride?” should not be answered simply  
1099 with “very toxic”. Instead, a more detailed descrip-  
1100 tion of the toxicity of barium chloride, including  
1101 the poisoning dose, poisoning symptoms, and poi-  
1102 soning mechanism, would be more informative and  
1103 useful. It is important to provide a well-rounded  
1104 and thorough answer to the question, rather than  
1105 just a brief or overly general response.

1106 **Normative.** It is important to answer questions in  
1107 written language, as this can help make the answer  
1108 more formal. Annotators should avoid using irreg-  
1109 ular or unconventional expressions that may not  
1110 be understood by everyone. Typos or grammatical  
1111 errors are not allowed.

## 1112 **C More Details for Data Collection**

1113 We limit our annotators and quality inspectors to  
1114 native Chinese speakers. In the meantime, they  
1115 must also speak and write fluent English. We make  
1116 sure all our annotators are fairly compensated by  
1117 the market price.

1118 **Question Creation.** Chinese QA forums, such  
1119 as [Zhihu](#) and [Baidu Zhidao](#), are known for their  
1120 abundance of long-form questions. However, when  
1121 these questions are utilized as direct queries on  
1122 Bing, users can often access multiple websites that  
1123 contain well-organized answers, thus making the  
1124 web search process less challenging. Such an issue  
1125 is not mitigated even if we block the source from  
1126 [Zhihu](#) and [Baidu Zhidao](#). In view of this, we strive  
1127 to annotate new open-ended questions that have not  
1128 been answered on Chinese QA forums.

1129 Following ELI5 ([Fan et al., 2019](#)), we turn to  
1130 creating questions from Reddit forums<sup>2</sup> as an alter-  
1131 native. We closely follow the way ELI5 collects  
1132 the source questions. After collection, we engage  
1133 annotators to refer to these questions and then ask  
1134 new questions in Chinese. This way significantly  
1135 improves the productivity of question creation.

1136 For quality control, our quality inspectors would  
1137 check whether the created question is meaningful,  
1138 semantically coherent, comprehensible, and reason-  
1139 able. Only those questions that satisfy the above  
1140 requirements would be retained. In addition, we  
1141 also remove the questions that are politically sen-  
1142 sitive. In total, 22.4% newly created questions are

<sup>2</sup><https://www.reddit.com/r/explainlikeimfive>

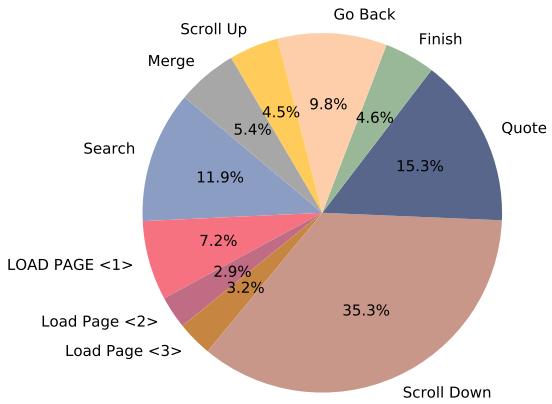


Figure 6: Proportion of different pre-defined actions in our IWSQA.

discarded. Each question is at least checked and approved by two quality inspectors.

**Web Search and Answer Annotation.** Before annotation, we provide our annotators with detailed annotation guidance. They got paid based on the number of instances they annotate instead of the time spent during annotation. Note for answer annotation, we did not require annotators to use all the collected facts when composing the answer but asked them to record which facts are leveraged in their answer. According to our estimate, each question takes about 16 minutes for our annotators to finish the web search and answer annotation.

**Proportion for Different Actions.** We record the proportion of different pre-defined actions in our collected dataset in Figure 6. As can be seen, Scroll Down, Quote, and Search are the most frequently used actions, which take up 62.5%. The proportion of Load Page <1> is larger than those of Load Page <2> and Load Page <3>. This is because search engines rank search results based on their relevance to the query. Humans tend to visit the links according to the order recommended by search engines. If humans have collected enough supporting facts on the first page or find it to be irrelevant, they probably would not continue browsing other web pages of the current query.

## D Details for the PLMs Evaluated

We select 6 series of representative and publicly available generative PLMs that support Chinese. For all the models, we use them for their intended uses. In the following, we give a brief introduction to them:

**mT5** (Xue et al., 2021) is a multilingual encoder-decoder PLM with a general-purpose text-to-text format. Its pre-training data mC4 (Xue et al., 2021) covers 101 languages collected from the public Common Crawl web scrape. **mT5** achieves superior performance in various multilingual benchmarks.

**MT0** (Muennighoff et al., 2022) is a multi-task fine-tuned version of Google’s **MT5**. The model attained strong zero-shot performance and cross-lingual generalization ability. Through explicit multi-task learning, a variety of language capabilities are enhanced through knowledge transfer; inevitably, some capabilities, which are not required by the trained tasks, might have been impaired.

**Mengzi-T5** (Zhang et al., 2021b) is a powerful Chinese encoder-decoder PLM that achieved state-of-the-art results on the CLUE benchmark. Instead of chasing a larger scale, the authors turn to developing lightweight yet more powerful models for easier deployment. **Mengzi-T5** was trained on Chinese Wikipedia, Chinese News, and Common Crawl and the total size of the pre-training corpus is 300G.

**mBART** (Liu et al., 2020) is a multi-lingual variant of BART, which is a sequence-to-sequence denoising auto-encoder. **mBART** is pre-trained on large-scale monolingual corpora with the BART (Lewis et al., 2020) pre-training objective. The model performs extremely well in machine translation tasks and can be generalized to languages that are not in the pre-training corpora.

**C-BART** (Shao et al., 2021) is the Chinese version of BART. Compared with **mBART**, the model was pre-trained only on a Chinese corpus. The model shows superior performance on keyword recognition tasks evaluated by the Rouge-L metric.

**CPM**<sup>3</sup> is the pre-trained model series provided by OpenBMB<sup>4</sup>. We choose three PLMs **CPM<sub>2.6B</sub>** (**CPM-1** (Zhang et al., 2021a)), **CPM<sub>7B</sub>** (**CPM-Live**), and **CPM<sub>10B</sub>** (**CPM-Ant**) with increasing model sizes. The three models are trained with increasingly larger sizes of data and training computations. **CPM<sub>10B</sub>** is trained on 200G texts cleaned from 1T raw data.

**Training Details.** For all the PLMs evaluated, we choose Adam (Kingma and Ba, 2015) as the optimizer and perform a grid search for learning rate in  $\{1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$ , and the batch size

<sup>3</sup><https://github.com/OpenBMB/CPM-Live>

<sup>4</sup><https://live.openbmb.org/en/>

1226 in  $\{8, 16, 32\}$  for all the sub-tasks. The learning  
1227 rate is warmed up for the first 10% training steps,  
1228 and the weight decay is set to  $1 \times 10^{-4}$ . For other  
1229 settings, we follow the configuration recommended  
1230 by the original papers. During training, we select  
1231 the model checkpoint with the best performance  
1232 on the development set and evaluate it on the test  
1233 set. The maximum sequence length is 2048 for  
1234 **mT0<sub>BASE</sub>**, **mT5<sub>BASE</sub>**, and **Mengzi-T5<sub>BASE</sub>**, 1024  
1235 for **mBART<sub>LARGE</sub>**, 512 for **C-BART<sub>LARGE</sub>**, and  
1236 3072 for **CPM**. We truncate the input sequence if  
1237 it exceeds the maximum sequence length of a PLM.  
1238 We fine-tune all the PLMs using NVIDIA A100,  
1239 and each fine-tuning takes no more than 2 days.

## 1240 E Design Differences between WebGPT 1241 and Our Work

1242 **Interface.** Our interface supports slightly different  
1243 actions than WebGPT. To begin with, we re-  
1244 move 2 actions defined by WebGPT: (1) **Find**  
1245 in **Page**: <text>, which supports finding  
1246 the next occurrence of <text> and scroll to it, and  
1247 (2) **Scroll to Top**, which supports scrolling  
1248 to the top of the page. In our pilot studies, even if  
1249 we give the annotators the options of both actions,  
1250 our annotators seldom execute them (< 1%). Con-  
1251 sidering that it may be hard for our model to learn  
1252 those extremely low-frequency actions, we do not  
1253 include both actions in the final list of our actions.

1254 Secondly, we modify the functionalities of the  
1255 **Scroll** actions in WebGPT. Specifically, Web-  
1256 GPT merged any consecutive **Scroll Down** and  
1257 **Scroll Up** actions made by humans into new  
1258 actions **Scroll Down** <?> and **Scroll**  
1259 **Up** <?>, where ? is the number of consecutive  
1260 actions. These new actions are utilized by their  
1261 models instead of the original **Scroll Down** and  
1262 **Scroll Up** actions. Therefore, there exists a  
1263 gap between what humans actually perform and  
1264 what the model is allowed to execute. We contend  
1265 that this gap could result in problems for behavior  
1266 cloning. Specifically, humans perform consecutive  
1267 **Scroll Down** actions because after each action,  
1268 they carefully check the current window and find  
1269 nothing useful. However, when merging consecu-  
1270 tive actions, the intermediate observations would  
1271 not be shown to the model, which makes decision  
1272 making even more difficult.

1273 Finally, we also implement a new **Merge** action  
1274 to support merging two supporting facts into one.  
1275 As mentioned before, **Merge** is specially designed

1276 for extracting a supporting fact that crosses the  
1277 boundary of two windows. This action is critical  
1278 to avoid recording fragmentary supporting facts.  
1279 As shown in Figure 6, **Merge** takes up a rela-  
1280 tively large (5.4%) percentage among all the ac-  
1281 tions, which is frequently executed by our anno-  
1282 tators. This action makes it possible for our anno-  
1283 tators to extract extremely long sequences as  
1284 supporting facts.

1285 **Framework.** WebGPT does not disclose the im-  
1286 plementation details for both interactive web search  
1287 and information synthesis (i.e., BC model in the  
1288 original paper), and only claims that “*we fine-tuned*  
1289 *on the demonstrations using supervised learning,*  
1290 *with the commands issued by the human demon-*  
1291 *strators as labels*”. In view of this, we propose our  
1292 own framework from scratch, with several design  
1293 choices not mentioned by WebGPT:

1294 We decompose the web search process into 3  
1295 distinct sub-tasks, i.e., action prediction, search  
1296 query generation, and supporting fact extraction.  
1297 We train 3 modules for each sub-task, respectively.  
1298 This design is because we find empirically that  
1299 training a multi-task model leads to inferior perfor-  
1300 mance than three individual specialized modules  
1301 (see appendix F.3). This implies that the three sub-  
1302 tasks might require intrinsically different language  
1303 skills. In addition, this decomposition also allows  
1304 us to evaluate three modules in isolation and gain  
1305 a deeper understanding of the strengths and weak-  
1306 nesses of each module. Furthermore, it allows for  
1307 flexibility in the system, as different modules can  
1308 be updated or replaced independently.

1309 For our synthesis model, instead of directly fine-  
1310 tuning on the (*question, supporting fact, answer*)  
1311 data, we explore (1) how to teach the model to ig-  
1312 nore irrelevant facts (§ 6.3). We achieve this goal  
1313 by introducing noisy facts into the training data to  
1314 explicitly force the model to ignore noisy facts, and  
1315 (2) how to generate novel contents beyond the col-  
1316 lected facts (appendix F.1). We corrupt the training  
1317 data by deleting partial supporting facts and forcing  
1318 the model to generate novel content based on  
1319 its pre-trained knowledge. We have demonstrated  
1320 with sufficient experimental results to show the  
1321 effectiveness of the above two methods.

1322 **Evaluation.** WebGPT only evaluates the **the**  
1323 **whole** pipeline through human evaluation. In ad-  
1324 dition to the holistic pipeline evaluation (§ 6.2),  
1325 we also evaluate each **individual** module of our

1326 pipeline (§ 6.1). To the best of our knowledge,  
 1327 this is the first work to decompose interactive web  
 1328 search into action prediction, search query genera-  
 1329 tion, and supporting fact extraction, and design the  
 1330 evaluation metrics for the three sub-tasks. It should  
 1331 be noted that holistic evaluation requires manual  
 1332 inspection, which is time-consuming despite being  
 1333 more accurate. Additionally, the holistic evalua-  
 1334 tion can only be conducted through interaction  
 1335 with the interface, whereas the individual sub-task  
 1336 evaluation can be conducted locally (by feeding  
 1337 the ground truth  $\mathcal{S}_t$  of the test data to each mod-  
 1338 ule). As such, individual sub-task evaluation is  
 1339 more flexible to implement, making it easier for  
 1340 hyper-parameter tuning, thus accelerating the de-  
 1341 velopment and iteration of the QA system. Besides,  
 1342 individual evaluation is more fine-grained, which  
 1343 helps us better understand the contribution of each  
 1344 part of the pipeline.

1345 **Analysis.** In addition to evaluating the LFQA per-  
 1346 formance of our pipeline, we also conduct an in-  
 1347 depth analysis to understand the contribution of  
 1348 core design elements of our framework. In § 6.3,  
 1349 we conduct ablation studies for the search model  
 1350 and the synthesis model, and a case study for the  
 1351 query module. We also show that our model in-  
 1352 deed acquires humanoid behaviors when interact-  
 1353 ing with the search engine. In the appendix, we  
 1354 further explore which output format is better for the  
 1355 supporting fact extraction module (appendix F.2).

1356 **Openness.** At the time of the submission of this  
 1357 paper, the interface, training data, trained models  
 1358 of WebGPT are not public. Only the comparison  
 1359 data (model-generated answers with their human  
 1360 preference) is released<sup>5</sup>. This makes it hard for  
 1361 researchers to continue exploring this line of study.  
 1362 We aim to open-source our interface, dataset, codes,  
 1363 and trained models to facilitate the research in this  
 1364 area. We also hope that the environment we pro-  
 1365 vide could serve as the testbed for other research  
 1366 topics, such as behavior cloning (Bain and Sammut,  
 1367 1995), planning in interactive environments (Huang  
 1368 et al., 2022a), and reinforcement learning from hu-  
 1369 man feedback (RLHF) (Stienon et al., 2020). This  
 1370 paper is licensed under the Creative Commons At-  
 1371 tribution 4.0 International License.

---

<sup>5</sup>[https://huggingface.co/datasets/openai/webgpt\\_comparisons](https://huggingface.co/datasets/openai/webgpt_comparisons)

$p$	0.1	0.2	0.3	0.4	0.5	1.0
NOVELTY	0.06	0.12	0.16	0.29	0.41	0.83
Length	256	216	206	201	193	126

Table 4: Results when the training data of synthesis model is corrupted with different  $p$ . We report two metrics for the generated sequence: NOVELTY and Length.

## F Additional Experiments and Analyses

### F.1 Generating Novel Contents v.s. Copying Supporting Facts

Another fascinating research question of our synthesis model is whether it could generate novel content based on its pre-trained knowledge. This ability is important especially when the collected facts are insufficient or fragmentary. Considering that copying the supporting facts and generating novel contents are often contradictory to each other, here we propose a method to flexibly strike a balance between both.

**Framework.** Specifically, we propose another way to corrupt the training data of the synthesis model. We split each collected fact into multiple sub-sentences according to punctuation and randomly erase part of these sub-sentences. We set a hyper-parameter  $p \in [0, 1]$ , which denotes the probability of erasing a sub-sentence. A higher  $p$  means more sub-sentences would be removed. After that, we concatenate the remaining sub-sentences into a new fact keeping the original order. Finally, we optimize the model to generate the human-annotated answer conditioned on the corrupted facts, i.e., maximizing:

$$\mathcal{P}(\text{Answer} | \mathcal{Q}_0, \text{CORRUPT}\{f_1, \dots, f_N\}).$$

Since the corrupted facts are fragmentary, the model learns to reconstruct those missing sub-sentences relying on its pre-trained knowledge.

**Settings.** We experiment with **CPM<sub>7B</sub>** and follow most of the settings in § 6.1. We test when different  $p$  is applied to corrupt the training data. Ideally, a higher  $p$  encourages the model to generate more novel content instead of copying the supporting facts. Specifically, we choose  $p$  from  $\{0.1, 0.2, 0.3, 0.4, 1.0\}$ , where 1.0 means the model sees no supporting facts but is required to generate all the tokens in the annotated answer.

During the evaluation, we feed **the original intact supporting facts** to the trained synthesis

Training Manner	Micro F1	Macro F1
Multi-task	52.1	46.3
Single-task	<b>58.9</b>	<b>50.5</b>

Table 5: Comparison of action prediction module when training **CPM<sub>7B</sub>** in a multi-task manner or single-task manner.

model. For evaluation metrics, we follow Welleck et al. (2019) to test the percentage of n-grams in the generated sequence that do not exist in the supporting facts, i.e.,

$$\text{NOVELTY}_n = \frac{|\text{unique generated n-grams}|}{|\text{total n-grams in supporting facts}|}.$$

The final novelty metric is defined as the average of NOVELTY<sub>2</sub>, NOVELTY<sub>3</sub>, and NOVELTY<sub>4</sub>, i.e.,

$$\text{NOVELTY} = \frac{1}{3}(\text{NOVELTY}_2 + \text{NOVELTY}_3 + \text{NOVELTY}_4).$$

Besides NOVELTY, we also record the number of generated tokens.

**Results.** We derive from the results listed in Table 4 that: (1) with  $p$  increasing, the metric NOVELTY constantly becomes larger. This demonstrates that by deleting more content of the supporting facts during training, we gradually encourage the synthesis model to generate novel content based on its pre-trained knowledge, instead of copying the supporting facts. However, it should also be noted that the generated information that is not included in the collected facts may suffer from poor factual accuracy. We expect future work to mitigate this issue; (2) in addition, with  $p$  increasing, the generated sequence tends to be shorter. This shows that only relying on the synthesis model cannot produce diverse, abundant, and informative contents, which emphasizes the importance of information retrieval in LFQA.

## F.2 Effect of the Output Format of Supporting Fact Extraction Module

When designing our supporting fact extraction module, we choose to output only the start and end tokens of the desired supporting fact instead of generating all the tokens. In this section, we compare the performance and execution speed for both strategies.

Following most of the settings in § 6.1, we train two modules using **CPM<sub>7B</sub>** with different output formats and compare both modules on our test set.

We find that compared with the strategy that generates only the start and end tokens, generating all the tokens leads to a slight performance drop in *Rouge-L* score from 59.8 to 57.7. For the execution speed, our chosen strategy only consumes 30.6% inference time than generating all the tokens on the test set. This shows that our chosen strategy is superior in both performance and execution speed. On average, our supporting fact extraction module extracts 164.7 characters for each *Quote* action during testing, which is close to that of humans (143.6).

## F.3 Multi-task Learning v.s. Single-task Learning for the Search Model

In our framework, we choose to train 3 individual modules with distinct functionalities for the search model. In this way, we can obtain specialized models for each sub-task of web search. An alternative is to train all three sub-tasks in a multi-task way so that we can deploy only one model in practice, saving storage space.

However, we find in our pilot study that multi-task training leads to a decline in performance than deploying single-task specialized modules. Specifically, we follow most of the settings in § 6.1 and train three sub-tasks in a multi-task manner using **CPM<sub>7B</sub>**. Then we compare its performance of action prediction. As listed in Table 5, multi-task training shows inferior performance than single-task training in the task of action prediction. This shows that **the three sub-tasks in web search might require intrinsically distinct language skills**, and combining them may impair their individual performance. Given the above reasons, we choose to deploy three different modules for our search model. We expect future works to explore how to combine these three tasks without suffering from a performance decline.

## F.4 More Case Study for the Query Module

In the main paper, we only present two example questions for the case study of the query module due to the length limit, here we present other 4 results in Figure 7. From these results, we contend our query module has acquired humanoid query strategies to find diverse information sources.

## G Samples of Annotated Examples

In this section, we sample a few annotated examples as a reference, see Figure 8 - Figure 13. We

<i>Question and Generated Query</i>
<p>为什么我们会对音乐类型有偏好?为什么有些人受不了某些类型的音乐?  <i>Why do we have preferences for musical genres? Why can't some people stand listening to certain genres of music?</i></p> <p>[1] 为什么我们会对音乐类型有偏好? <i>Why do we have preferences for music genres?</i>  [2] 为什么有些人受不了某些音乐 <i>Why can't some people stand certain music</i>  [3] 音乐品味 <i>Music taste</i></p>
<p>为什么在大学辩论队中他们说话这么快?  <i>Why in college debate teams do they speak so fast?</i></p> <p>[1] 为什么在大学辩论队中他们说话这么快? <i>Why in college debate teams do they speak so fast?</i>  [2] 为什么大学辩论队的人说话都很快 <i>Why do people on the college debate team speak so fast</i>  [3] 辩论队的要求 <i>Debate team requirements</i>  [4] 辩论队说话语速快的原因 <i>The reason why the debate team speaks fast</i>  [5] 辩论队的训练 <i>Debate team training</i>  [6] 大学辩论队的人说话快 <i>People on college debate teams talk fast</i>  [7] 辩论队的人说话快 <i>People on the debate team speak fast</i></p>
<p>里和海里之间的区别? 为什么我们需要两种不同的测量值?  <i>The difference between miles and nautical miles and why we need two different measurements?</i></p> <p>[1] 里和海里之间的区别? <i>Difference between mile and nautical mile?</i>  [2] 为什么我们需要两种不同的测量值? <i>Why do we need two different measurements?</i>  [3] 海里和里的区别 <i>Difference between nautical mile and mile</i></p>

Figure 7: More case study for the query module in addition to Figure 4.

present the following items: the original question, the simplified action sequence, the collected supporting facts, and the annotated answer. We omit the specific generated queries and the links / titles of the browsed web pages. We have also translated the original Chinese version into English.

## H Samples of Generated Answers by our Pipeline

We randomly sample a few answers to the test question of IWSQA generated by our pipeline, see Figure 14 - Figure 19. It can be derived from these samples that, in general, our pipeline has learned to integrate relevant supporting facts and generate coherent answers. However, the pipeline still has some limitations. For instance, in some cases, the generated answers may not be as accurate and detailed as human-written ones. Additionally, the models may not always be able to understand the nuances of a question or interpret it in the same way as a human would.

## I Future Explorations

We expect future works to explore the following directions:

**Pre-training and Utilizing More Powerful PLMs.** Recent years have witnessed the powerful capabilities of ever-larger PLMs, which have increasingly stronger language understanding and gener-

ation abilities. The superior LFQA performance of WebGPT may largely be owing to the powerful capability of the backbone PLM GPT-3 with 175B parameters. However, GPT-3 is not publicly available, and for public large PLMs that support Chinese (e.g., GLM-130B (Zeng et al., 2022) and BLOOM (Scao et al., 2022)), the corresponding computing resources required are beyond the scope of most research groups. In this paper, we evaluate 8 representative PLMs that support Chinese, with the largest PLM exceeding 10B parameters. We expect future work to build and evaluate more powerful PLMs.

**Efficient and Scalable Use.** Despite the fascinating feature of interactive web search, such a process is inherently slower to execute than the conventional non-interactive retrieval process of open-domain QA. In this regard, we encourage further explorations in reducing the latency of our pipeline. Possible solutions include improving the speed and memory usage of the PLM, developing techniques for distributed training, and enhancing the PLM's ability in action prediction so that the search process can be finished with fewer actions.

**Extension to Other Languages and Domains.** It would be interesting to extend the current approach to other languages beyond Chinese. Considering that the search engine supports multiple languages, our interface can be easily adapted to

1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553

1554 building benchmarks for other languages. In addition,  
1555 our dataset focuses on the general domain. It  
1556 is worthwhile to extend the dataset to cover more  
1557 specific domains, such as technology, science, and  
1558 politics, which may contain more challenging ques-  
1559 tions.

1560 **Leveraging the Reference Information.** In ad-  
1561 dition to the annotated answers, we also require  
1562 the annotators to record which supporting facts are  
1563 referenced and leveraged in their answers. How-  
1564 ever, in this paper, we do not utilize this informa-  
1565 tion when training our synthesis model. Intuitively,  
1566 such information could guide the synthesis model  
1567 to better organize existing supporting facts in a  
1568 more coherent way, and to improve its ability in  
1569 selecting important information and ignoring irrel-  
1570 evant noises.

1571 **Diversify the Interactive Elements.** In this pa-  
1572 per, we focus on supporting the mainstream web  
1573 search actions for our users. It would interesting  
1574 to explore incorporating more interactive elements  
1575 into the interface, such as allowing the users to  
1576 provide feedback on the retrieved information and  
1577 supporting multimedia information retrieval. How-  
1578 ever, more actions also increase the difficulty of  
1579 behavior cloning to a certain degree.

1580 **Incorporating Additional Information Sources.**  
1581 In addition to collecting knowledge from the web  
1582 using interactive web search, other sources of infor-  
1583 mation such as structured databases and knowledge  
1584 graphs can also be integrated into the information  
1585 retrieval process to improve the relevance and com-  
1586 pleteness of the supporting facts.

1587 **Improving Model Behavior from Human Feed-**  
1588 **backs.** WebGPT has demonstrated it is promising  
1589 to use reinforcement learning from human feed-  
1590 back (RLHF) (Stiennon et al., 2020) to improve  
1591 the quality of the generated answers. RLHF can  
1592 also be used for improving the search model’s web  
1593 search behavior, and make it collect more diverse  
1594 and relevant supporting facts. As mentioned be-  
1595 fore, our provided environment can be utilized by  
1596 researchers to study RLHF in the future.

---

**Question:**

麦田怪圈是什么？它们是如何形成的？

**Translated Question:**

What are crop circles and how are they formed?

**Action Sequence:**

```
Search → Load Page <1> → 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 → Scroll Down → 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。

**Translated Supporting Facts:**

1. Crop circles, also known as “crop formations”, are geometric patterns that appear in fields of crops, often as a result of unknown forces (though most crop circles are actually created by humans). This mysterious phenomenon has sometimes been used to support the belief in extraterrestrial life.
2. Human-made theory: According to the human-made theory, crop circles are believed to be created using planks of wood. The ends of the planks are attached to ropes to form a loop, and the crops are flattened by dragging the planks across them while maintaining a fixed distance from the center of the circle using the ropes. This process can gradually create a circular pattern. To make the process easier, the diameters of most circles can be evenly divided by 6. In the past, there have been instances of crop circle makers being caught in the act, and they have used this method to create the circles.
3. Natural formation theory: According to the natural formation theory, some people believe that crop circles are a phenomenon whose cause has not yet been discovered by humans. Similar to how lightning was once thought to be the work of the god Zeus, the appearance of human-like symbols in crop circles is seen as a result of people’s preconceptions.
4. Magnetic field theory: According to the magnetic field theory, some experts believe that crop circles are created by a mysterious force in the magnetic field that generates an electric current that causes the crops to lie flat on the ground. American expert Jeffrey Wilson studied over 130 crop circles and found that 90% were located near transformers connected to high-voltage power lines and within 270 meters of a pond or other source of water. The irrigation of the soil in the fields can cause the release of ions in the soil at the bottom of the field, creating a negative charge. The transformer connected to the high-voltage power lines generates a positive charge, and the collision of the negative and positive charges creates electromagnetic energy that knocks down the wheat, forming a crop circle.

**Answer:**

麦田怪圈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。【1】而麦田怪圈的形成有三个说法，人为说、自然说与磁场说。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。【2】自然形成说认为麦田怪圈的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。【3】磁场说认为，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。【4】

**Translated Answer:**

Crop circles are geometric patterns created by flattening crops through an unknown force (most crop circles are made by humans) in a field of wheat or other crops. [1] There are three theories about the formation of crop circles: human-made, natural, and magnetic fields. The human-made theory generally believes that crop circles are made using wooden boards. The wooden boards are connected by ropes to form a hoop, and during the production process, one-foot steps on the wooden board, dragging the wooden board to flatten the wheat, and pulling the fine rope to maintain a fixed distance from the center, gradually forming a circle. In order to facilitate manufacture, the main shape of all circles has a diameter that can be evenly divided by 6. [2] The natural formation theory believes that the cause of crop circles has not yet been discovered by humans. Like lightning, ancient humans also thought it was made by the thunder and lightning mother, and for the phenomenon of human information often appearing in crop circles, they think it is just a "preconceived" illusion caused by people. [3] The magnetic field theory believes that there is a mysterious moving force in the magnetic field that can generate an electric current, causing crops to "lie flat" on the ground. [4]

---

Figure 8: Sampled example (1) from IWSQA.

---

**Question:**

根据民意调查，大多数用户忽视了网络广告。为什么公司仍然为这些网络广告付费？

**Translated Question:**

According to polls most users ignore Internet ads. So why do companies still pay for these online ads?

**Action Sequence:**

Search × 2 → Load Page <1> → Go Back → Load Page <3> → Scroll Down × 2 → Search → Load Page <2> → Quote × 8 → Scroll Down → Quote → Merge → Quote → Go Back → Load Page <1> → Scroll Down → Go Back → Scroll Down → Load Page <3> → Scroll Down × 2 → Search → Scroll Down × 2 → Finish

**Supporting Facts:**

1. 网络作为第四种媒体，有着传统媒体不可比拟的优势：互联网广告传播范围更加广泛
2. 任何人在任何地点都可以随时浏览到网络的广告信息
3. 互联网广告可直达产品核心消费群体
4. 互联网广告具有强烈的互动性
5. 受众是广告的主人，受众只会点击感兴趣的讯息，因而增强广告效果。而厂商也可以在线随时获得大量的用户反馈信息
6. 互联网广告富有创意，感官性强
7. 互联网广告以多媒体、超文本格式为载体，图、文、声、像传递多感官的信息，使受众能身临其境般感受商品或服务
8. 互联网广告更加节省成本
9. 而且可按需要变更内容或改正错误，使广告成本大大降低

**Translated Supporting Facts:**

1. The internet is the fourth type of medium and has incomparable advantages over traditional media: wider reach for internet advertising
2. Anyone can view online advertising at any time and from any location.
3. Internet advertising can directly reach the core consumer group of a product.
4. Internet advertising is highly interactive.
5. The audience is in control of the advertising and only clicks on information that interests them, which enhances the effectiveness of the advertising. Companies can also receive a large amount of user feedback online in real time.
6. Internet advertising is creative and visually appealing.
7. Internet advertising uses multimedia and hypertext formats to transmit sensory information through text, images, sound, and video, allowing the audience to feel like they are experiencing the product or service firsthand.
8. Internet advertising is more cost-effective.
9. The content or errors can be changed as needed, greatly reducing advertising costs.

**Answer:**

现在是互联网时代，比起传统的电视或纸媒广告，受众基数还是较多。【1, 2】网络广告还是有着不可比拟的优势，它可直达产品核心消费群，让厂商可以在线获得大量的用户反馈信息，从而精准统计效率和访问量。

【3, 4, 5】网络广告的成本要比传统广告低，而且也可按需要变更内容或改正错误，深得广告厂商们的喜欢。  
【8, 9】

**Translated Answer:**

We are currently in the internet age, and online advertising has a larger audience base compared to traditional television or print media.[1,2] Online advertising still has unparalleled advantages; it can directly reach the core consumer group of a product, allowing companies to receive a large amount of user feedback online in real time, improving statistical efficiency and tracking the number of visits.[3,4,5] The cost of online advertising is lower than traditional advertising, and the content or errors can be changed as needed, making it popular among advertisers.[8,9]

---

Figure 9: Sampled example (2) from IWSQA.

---

**Question:**

当猫狂暴地把所有东西都打翻，表现得像个疯子时，是怎么回事？

**Translated Question:**

What is going on with my cat when he goes on a rampage knocking everything over and acting like a lunatic?

**Action Sequence:**

Search → Search → Load Page <2> → Go Back → Load Page <1> → Scroll Down → Quote → Quote → Scroll Down → Quote → Merge → Finish

**Supporting Facts:**

1. 很多人认同这是属于猫的一种先天行为，例如狩猎本能。
2. 它的基因里也刻着猎手的本能反应。面对眼前的事物，它会展开调查，用爪子测试它是活物还是死物，能不能吃，以此来确认如何制服或者吃掉这个东西：这种调查除了狩猎，也算是一种探索行为。像是家里的摆放环境有所变化，就有可能会引起它的好奇，或者调查新鲜的东西是否会移动，或者移动方式如何。发现可以动后，有时会进入娱乐模式，就是多扒拉几下，或者更加用力的推它
3. 另外日本一项实验表明，猫很可能对引力有所了解，当然这不是说它精通物理，而说猫很有可能知道物体离开桌面会坠落，然后撞击发出声响，也就是它可能会故意推掉东西。而让它坠落后，物体的坠落过程，和撞击发出的声音，对它来说都很有吸引力。

**Translated Supporting Facts:**

1. Many people believe that this is an innate behavior for cats, such as the instinct to hunt.
2. Their genes also contain the instinctive response of a hunter. In the face of something in front of them, they will conduct an investigation, using their claws to test whether it is alive or dead, whether it can be eaten, and how to subdue or eat it: this investigation is not only hunting but also a form of exploration. For example, changes in the home environment may cause their curiosity, or they may investigate whether fresh things will move, or how they will move. After finding that they can move, they may enter the entertainment mode, that is, play with it more or push it more forcefully.
3. In addition, a Japanese experiment showed that cats are likely to have some understanding of gravity, of course, this does not mean that they are proficient in physics, but that cats are likely to know that objects will fall off the desk and make a noise when they hit, that is, they may intentionally push things off. And after it falls, the process of the object falling and the sound of the impact are very attractive to it.

**Answer:**

这是猫的一种天性，因为它的基因里刻着猎手的本能反应。所以面对眼前的事物，它会展开调查，用爪子测试它是活物还是死物，能不能吃，以此来确认如何制服或者吃掉这个东西。还有一种可能就是它想玩，因为东西坠落后，物体的坠落过程，和撞击发出的声音，对它来说都很有吸引力，而且主人一听见就会过来，他感觉很有趣。

【1, 2, 3】

**Translated Answer:**

This is a natural behavior for cats because their genes contain the instinctive response of a hunter. Therefore, in the face of something in front of them, they will conduct an investigation, using their claws to test whether it is alive or dead, whether it can be eaten, and how to subdue or eat it. Another possibility is that they want to play for fun. Because after an object falls, the process of the object falling and the sound of the impact are very attractive to them, and when the owner hears it, he will come over and the cat finds it very interesting.[1,2,3]

---

Figure 10: Sampled example (3) from IWSQA.

---

**Question:**

为什么一匹马摔断了腿就要被淘汰?

**Translated Question:**

Why do horses have to be put down when they break a leg?

**Action Sequence:**

Search → Load Page <1> → Scroll Down × 4 → Go Back → Load Page <3> → Go Back → Load Page <2> → Go Back → Scroll Down → Load Page <1> → Scroll Down → Quote × 3 → Scroll Down → Quote → Merge → Finish

**Supporting Facts:**

1. 马是一种十分特别的动物，马的体重重达半吨，这么重的重量却仅仅靠一个细细的腿上一个指节站立
2. 这样的特殊构造，决定了马腿的精密，如果长时间只用三只脚站立，不仅断腿好不了，好腿由于长时间的代偿作用也会被折腾坏了。
3. 总而言之，断了腿的赛马，不仅断腿治不好，而且会影响其他腿的健康，如果坚持治疗，只会给马匹带来无尽的痛苦，最后死得更惨。

**Translated Supporting Facts:**

1. Horses are a very special type of animal, with a weight of up to half a ton but supported by just one thin leg and on a single finger joint.
2. This special construction determines the delicacy of a horse's legs. If a horse stands on only three legs for a long time, not only will the broken leg not heal properly, but the good legs will also be damaged due to the long-term compensation effect.
3. In summary, a horse with a broken leg not only will not heal properly, but it will also affect the health of the other legs. If treatment is insisted on, it will only bring endless pain to the horse and ultimately result in a more miserable death.

**Answer:**

因为马所有的重量都在腿上，摔断了腿，就意味着如果马儿持续活下来，那么好腿由于长时间的代偿作用也会被折腾坏了。所以，断了腿的马，不仅断腿治不好，而且会影响其他腿的健康，如果坚持治疗，只会给马匹带来无尽的痛苦，最后死得更惨，还不如淘汰。【1, 2, 3】

**Translated Answer:**

Because all of a horse's weight is on its legs, breaking a leg means that if the horse continues to live, the good leg will also be damaged due to the long-term compensation effect. Therefore, a horse with a broken leg not only will not heal properly, but it will also affect the health of the other legs. If treatment is insisted on, it will only bring endless pain to the horse and ultimately result in a more miserable death. It is better to eliminate the horse rather than continue treatment. [1,2,3]

---

Figure 11: Sampled example (4) from IWSQA.

---

**Question:**

如果我对家猫过敏，这是否意味着我对狮子或老虎等大型猫科动物过敏？

**Translated Question:**

If I'm allergic to house cats, does that mean I'm allergic to big cats such as lions or tigers?

**Action Sequence:**

Search → Search → Load Page <1> → Quote → Go Back → Load Page <2> → Quote → Go Back → Scroll Down → Search → Finish

**Supporting Facts:**

1. 大致而言，大部分对家猫过敏者也会对大型猫科动物产生过敏反应，只是反应现象可能会稍许轻微一些
2. 并不是对猫毛过敏，而是对猫身上分泌的猫蛋白Feld 1 过敏，这种过敏原物质会通过猫身上的唾液腺或者皮肤腺体，释放到皮肤上，或者猫咪在舔毛发给自己梳理猫毛时候，这些过敏原就都会粘连到上面。是放到猫猫。最后在猫咪到处活动或，猫毛四处散落，猫毛身上的过敏原也随之四处散落，在你不经意间导致你发生过敏。也并不会对所有的猫科动物都过敏，有的时候，你可能只对同一品种的猫过敏，而对另一种猫就不过敏，这都是有可能的。

**Translated Supporting Facts:**

1. Generally speaking, most people who are allergic to domestic cats will also have allergic reactions to big cats, but the reaction may be slightly milder
2. It is not allergic to cat fur, but allergic to the cat protein Feld 1 secreted by cats. This allergen substance will be released to the skin through the salivary glands or skin glands of the cat, or these allergens will stick to furs when the cat is licking its own hair. Is to put cats. Finally, when the cat is moving around or the cat fur is scattered everywhere, the allergens on the cat fur are also scattered everywhere, causing you to have allergies inadvertently. You're not allergic to all cats, and it's possible that you're allergic to a cat of the same breed but not to another.

**Answer:**

这主要是看过敏源和个人体质。大致而言，大部分对家猫过敏者也会对大型猫科动物产生过敏反应，只是反应现象可能会稍许轻微一些。还有一种是对猫身上分泌的猫蛋白Feld 1 过敏，这种并不会对所有的猫科动物都过敏。  
【1, 2】

**Translated Answer:**

This mainly depends on the allergens and personal constitution. In general, most people who are allergic to domestic cats will also have an allergic reaction to big cats, although the reaction may be slightly milder. There is also an allergy to Feld 1, a protein secreted by cats, which is not allergic to all cats. [1,2]

---

Figure 12: Sampled example (5) from IWSQA.

---

**Question:**

青春期后有没有办法长高?

**Translated Question:**

Is there a way to grow taller after puberty?

**Action Sequence:**

```
Search → Load Page <3> → Scroll Down × 2 → Go Back → Scroll Down × 3 → Search → Load Page <1> → Quote → Scroll Down → Quote → Merge → Scroll Down → Scroll Up → Quote → Scroll Down → Quote → Go Back → Load Page <2> → Scroll Down × 6 → Go Back → Finish
```

**Supporting Facts:**

1. 成年后还能长高的原理吧。一、运动增高的原理 其实成年后还能不能长高，从科学的角度来看，主要看骨骼线的闭合程度，特别是大腿关节处的骨骼，如果你去医院拍ct，发现你的骨骼线还没有完全闭合，那你就还有可能长高，特别是找到适合自己的运动，长期坚持运动加健身，非常有利于长高。
2. 那如果是骨骼线已经完全闭合呢？其实骨骼线完全闭合的成年人也可以长高，只不过如果骨骼线没有闭合的话，可以通过有氧运动来增高，但是骨骼线已经闭合的成年人只能通过调整体态来达到增高的目的，这个调整体态可是视觉上的增高，对于自身形象也有所提高。所以，如果你还有一些增高的想法的话。先去医院照一组ct，看自己的骨骼有没有完全闭合，成年人的骨骼大部分已经完全闭合，那么可以通过调整体态的方式来增高。
3. 就比如驼背，溜肩还有头前伸。这几个不良体态在现代人中很常见。如果你有不良体态，那从视觉上你就比别人矮了一截，造成不良体态的原因是因为现代人在电脑前，在手机前的时间越来越长。导致很多人都形成了不良的体态。

**Translated Supporting Facts:**

1. The principle of growing taller after adulthood. 1. The principle of exercise heightening. In fact, whether you can grow taller after adulthood, from a scientific point of view, mainly depends on the degree of closure of the bone line, especially the bones at the thigh joints. If you go to the hospital for a CT scan and find that your bone line is not completely closed, then It is still possible for you to grow taller, especially if you find suitable sports for you. Long-term exercise and fitness are very conducive to growing taller.
2. What if the bone line is completely closed? In fact, adults with completely closed bone lines can also grow taller, but if the bone line is not closed, they can increase their height through aerobic exercise, but adults with closed bone lines can only achieve the purpose of increasing height by adjusting their posture. Adjusting the posture can increase the visual height and improve the self-image. So, if you still have some ideas for heightening. First, go to the hospital to take a set of CT scans to see if your bones are completely closed. Most of the bones of adults have been completely closed, so you can increase your height by adjusting your posture.
3. Such as hunchback, sloping shoulders, and forward head. These bad postures are very common among modern people. If you have a bad posture, you are visually shorter than others. The reason for the bad posture is that modern people spend more and more time in front of computers and mobile phones, leading to a lot of people forming bad posture.

**Answer:**

成年后还能不能长高，从科学的角度来看，主要看骨骼线的闭合程度，特别是大腿关节处的骨骼，如果你去医院拍ct，发现你的骨骼线还没有完全闭合，那你就还有可能通过有氧运动长高，特别是找到适合自己的运动，长期坚持运动加健身，非常有利于长高。【1】而成年人的骨骼线大部分已经完全闭合，那么可以通过调整体态的方式来增高。【2】就比如驼背，溜肩还有头前伸。这几个不良体态在现代人中很常见。如果你有不良体态，那从视觉上你就比别人矮了一截。【3】通过改善不良体态，也可以达到一定的增高效果。

**Translated Answer:**

Whether you can grow taller after adulthood, from a scientific point of view, mainly depends on the degree of closure of the bone line, especially the bones at the thigh joints. If you go to the hospital for a CT scan and find that your bone line is not completely closed. It is also possible to grow taller through aerobic exercise, especially if you find the exercise that suits you. Long-term exercise and fitness are very conducive to growing taller. [1]Most of the bone line of adults has been completely closed, so the height can be increased by adjusting the body posture. [2]It's like a hunchback, with sloping shoulders, and protruding head. These bad postures are very common among modern people. If you have bad posture, you will be visually shorter than others. [3]By improving bad posture, a certain increased effect can also be achieved.

---

Figure 13: Sampled example (6) from IWSQA.

---

**Question:**  
唾液真的没味道吗?

**Translated Question:**  
Is saliva really tasteless?

**Pipeline-generated Answer:**  
唾液无色无味，PH为6.6-7.1。正常的口水是没有味道的。当口水离开你的嘴后，水分就会被蒸发掉，空气中的微生物和口水中的各种酶发生一系列的反应，才会有臭味。

**Translated Answer:**  
Saliva is colorless and tasteless, with a pH of 6.6-7.1. Normal saliva is tasteless. When the saliva leaves your mouth, the water will be evaporated, and a series of reactions will occur between the microorganisms in the air and various enzymes in the saliva, and then there will be a bad smell.

---

Figure 14: Sampled answer generated by our pipeline on our test set (1).

---

**Question:**  
为什么醒来时很难握紧拳头，但随着时间的推移这会变得更容易?

**Translated Question:**  
Why is it so hard to make a fist when you wake up, but it gets easier over time?

**Pipeline-generated Answer:**  
刚醒来时很难握紧拳头，是因为手部肌肉中的电解质含量不足。当我们睡觉时，身体会放松下来，心率会下降，流向四肢的血液减少，流向四肢的电解质也自然减少。当我们醒来时，所有的系统会恢复工作，但那些在睡觉时获得更少电解质的部位，比如四肢需要更长的时间来获得足够的电解质才能完全恢复工作，因此当我们刚醒来时，手会显得没那么有力气，很难握紧。

**Translated Answer:**  
It is difficult to make a fist when you first wake up, it is because there are not enough electrolytes in the hand muscles. When we sleep, the body relaxes, the heart rate drops, there is less blood flow to the extremities, and there is naturally less electrolyte flow to the extremities. When we wake up, all the systems go back to work, but those parts that get fewer electrolytes while sleeping, like the extremities, take longer to get enough electrolytes to fully go back to work, so when we just wake up, the hand will appear weak and difficult to hold.

---

Figure 15: Sampled answer generated by our pipeline on our test set (2).

---

**Question:**  
如果食物进入肺部会发生什么?

**Translated Question:**  
What happens if food gets into the lungs?

**Pipeline-generated Answer:**  
如果真有东西进入了肺部，极大几率会有肺炎，这时是人体免疫系统在对抗外来异物。在没有就医的情况下，液体般水溶类异物可能会被消耗掉，或者死于肺炎感染。如果异物像砂石这些人体不能被消耗掉的东西，要么成了肉裹后并有几率演变成肺炎然后死于肺炎感染，要么直接死于肺炎感染。也就是有异物进入了肺部，就有生命危险。食物进入肺部以后，会引起气管刺激性的咳嗽，这是典型的咳嗽刺激性的干咳，通常是阵发性的发作。

**Translated Answer:**  
If something really gets into the lungs, there is a high chance of pneumonia. At this time, the body's immune system is fighting against foreign item. In the absence of medical attention, the liquid-like water-soluble xenoid may be consumed, or die from a pneumonia infection. If the foreign item is something that cannot be consumed by the human body, such as sand and gravel, it will either become wrapped in meat and have a chance to develop into pneumonia and then die from pneumonia infection, or die directly from pneumonia infection. That is, if a foreign item enters the lungs, it is life-threatening. After food enters the lungs, it will cause a tracheal irritating cough, which is a typical irritating dry cough, usually in paroxysmal attacks.

---

Figure 16: Sampled answer generated by our pipeline on our test set (3).

---

**Question:**

当鼻子充血时，为什么很难尝出东西的味道？

**Translated Question:**

Why is it so hard to taste things when your nose is congested?

**Pipeline-generated Answer:**

一方面是因为，感冒时鼻腔黏膜充血肿胀、流涕，阻塞鼻腔气道，限制了吸气时的气流量，空气到达不了嗅上皮这一味觉灵敏区。另一方面可能是病毒直接侵袭了嗅觉神经组织，使嗅觉传导通路受损。但是感冒引起的嗅觉障碍多是短暂性的，即使出现了较长时间的嗅觉障碍大约有1/3的人也会在半年内恢复，所以并不可怕。

**Translated Answer:**

On one hand, it is because the nasal mucosa is congested and swollen during a cold, and the nasal airway is blocked, which limits the airflow during inhalation, and the air cannot reach the taste-sensitive area of the olfactory epithelium. On the other hand, it may be that the virus directly invaded the olfactory nerve tissue and damaged the olfactory conduction pathway. However, most of the anosmia caused by a cold is short-lived. Even if there is long-term anosmia, about 1/3 of the people will recover within half a year, so it is not terrible.

---

Figure 17: Sampled answer generated by our pipeline on our test set (4).

---

**Question:**

为什么我们打哈欠时会流泪？

**Translated Question:**

Why do we cry when we yawn?

**Pipeline-generated Answer:**

人在打哈欠的时候嘴巴张得很大，尽可能多地让气呼出体外，为的是呼出二氧化碳，吸进氧气。随着打哈欠的动作一股气体有力地从嘴巴呼出。面颊部、舌和咽喉部肌肉紧张收缩，这时口腔及鼻腔的压力增大。鼻腔压力增高的结果就会一时性地阻挡鼻泪管排泄泪水的工作，“下水道”不通了，泪水流不下去，积在眼睛里的泪水就只能从眼眶中流出来。虽然我们的眼泪一直都在分泌，但是并不会从眼睛中流出，这是因为我们一直在眨眼睛，我们的眼睛下面是一个通道的，叫做鼻泪管，当我们眨眼睛时泪水就会进入到鼻泪管中，因此我们的泪水就不会从眼睛中流出了，但是当我们打哈欠时，鼻腔压力增高就会导致鼻泪管堵塞，这样泪水就流不进去了，泪水无法正常分泌，就只能从我们的眼睛中流出了，这就是我们打哈欠会流泪的原因了。

**Translated Answer:**

When a person yawns, he opens his mouth wide, exhaling as much air as possible, in order to exhale carbon dioxide and inhale oxygen. With a yawning movement, a puff of air is exhaled forcefully from the mouth. Muscles in the cheeks, tongue, and throat tense and contract, increasing pressure in the oral and nasal cavities. As a result of increased nasal cavity pressure, the nasolacrimal duct will temporarily block the work of draining tears. The "sewer" is blocked, and the tears cannot flow down. The tears accumulated in the eyes can only flow out of the eye sockets. Although our tears are secreted all the time, they don't flow out of our eyes. This is because we have been blinking. There is a channel under our eyes called the nasolacrimal duct. When we blink, the tears flow out. Into the nasolacrimal duct, so our tears will not flow out of the eyes, but when we yawn, the increased pressure in the nasal cavity will cause the nasolacrimal duct to block, so that the tears cannot flow in, and the tears cannot be secreted normally, it can only flow out of our eyes, which is why we cry when we yawn.

---

Figure 18: Sampled answer generated by our pipeline on our test set (5).

---

**Question:**

如果在太空中死亡，会被分解还是保持不变？

**Translated Question:**

If you were to die in space, would you decompose or stay the same?

**Pipeline-generated Answer:**

根据死亡状况来判断，遗体会结冰，脱水，缓慢腐烂或者正常降解。如果遗体位于近地轨道，那么过不了多久就会往地球坠落，在经历大气层的时候被烧毁。若所处的轨道高度再高一些，那就是太空垃圾最密集的区域了，在这个轨道，遗体会遭到微小碎片的撞击，也许会在未来跟某个碎片剧烈相撞。而如果处于更高的轨道，遗体就相对比较安全了，比如位于地球同步轨道周围，被太空垃圾撞击的可能性大大降低。

**Translated Answer:**

Depending on the state of death, the body freezes dehydrates, decays slowly, or degrades normally. If the remains are in low-Earth orbit, it won't be long before they fall to the Earth and burn up as they pass through the atmosphere. If the altitude of the orbit is higher, it will be the densest area of space junk. In this orbit, the body will be hit by tiny debris and may collide violently with certain debris in the future. And if it is in a higher orbit, the remains will be relatively safe, for example, if it is located around the geosynchronous orbit, the possibility of being hit by space junk is greatly reduced.

---

Figure 19: Sampled answer generated by our pipeline on our test set (6).