A Prompt Collection

We collect the prompt from various sources. The details for each domain are as follows:

World Knowledge: We collect prompts encompassing a broad spectrum of world knowledge, including historical events, common sense, news events, culture, and society. A big part of these prompts are sourced from TruthfulQA (Lin et al., 2022), with a smaller portion being contributed by online platforms such as Quora, Twitter, online blogs and documented error archives.[2] A minor fraction were manually drafted by the authors and by ChatGPT. A few prompts are from hc3 (Guo et al., 2023) and MMLU (Hendrycks et al., 2020). To select prompts from Quora, we randomly chose questions from the History and Society topics. For TruthfulQA, we selected questions from a variety of categories, such as Sociology, Economics, Politics, and Law.

Science and Technology: In this domain, we collect questions about science, technology, and research mainly from Quora, MMLU, and online sources mentioned above, alongside questions generated by ChatGPT and manually designed by us. These prompts vary from examination questions of scientific knowledge to open-ended scientific questions. On Quora, we pick questions from scientific topics such as Scientific Research, Science of everyday life, Technology, and Physics. On MMLU, we select questions from the econometrics, computer security and college chemistry subjects. We also select some questions from the online blogs mentioned above. And we manually design 9 prompts for this domain.

Recommendation and Writing: We use ChatGPT to auto-generate prompts for recommendation. We first draft some prompts as few-shot exemplars, then feed them into ChatGPT to generate more prompts in a self-instruct manner (Wang et al., 2022). These prompts cover requests for recommending books, online courses, restaurants, and tourist attractions. Writing tasks involve requesting LLMs to generate articles or essays on specified topics. An example prompt is: "*Write a dating profile for Mark ACHBAR based on his Wikipedia page*". In this domain, we expect that the generated responses are relatively longer compared to other tasks. However, as an auto-regressive language model, ChatGPT would accumulate past errors when generating long textual content. This is why we include writing tasks within the considered domains when evaluating factuality.

Reasoning: Most of the prompts in this domain are from the GSM8K dataset (Grade School 8K) (Cobbe et al., 2021), which is a dataset of more than 8k highly diverse problems. These questions consist of basic numerical problems that require multi-step reasoning. We pick more than 200 challenging questions where the `text-Davinci-003` model makes mistakes, as shown in the HELM (Liang et al., 2022) website. In addition, a small part of the prompts are from the online sources and designed by authors.

Math: We collect problems mainly by picking questions from MATH (Hendrycks et al., 2021) where the `text-davinci-003` model makes mistakes as shown on the HELM website, similar to how we collect prompts in the reasoning domain. We select questions from algebra, counting, and probability subjects . A small part of prompts are from online sources.

B Annotation Page

We develop the annotation tool as shown in Figure 5. The tool is developed using HTML/JavaScript. The tool is designed for annotators to label the factuality, identify error types, provide reasons for the errors, and include reference links.

[2]The online blog, github repository, twitter thread and documented archive we take as reference are `https://garymarcus.substack.com/p/large-language-models-like-chatgpt`, `https://github.com/giuven95/chatgpt-failures`, `https://twitter.com/DieterCastel/status/1598727145416790028?lang=en`, `https://twitter.com/zhou_yu_ai/status/1644697590586384384?s=46&t=7b5KyE0RBwd0oyYd2mHqfA` and Borji (2023). We use "online sources" to refer to them throughout the paper unless otherwise specified.

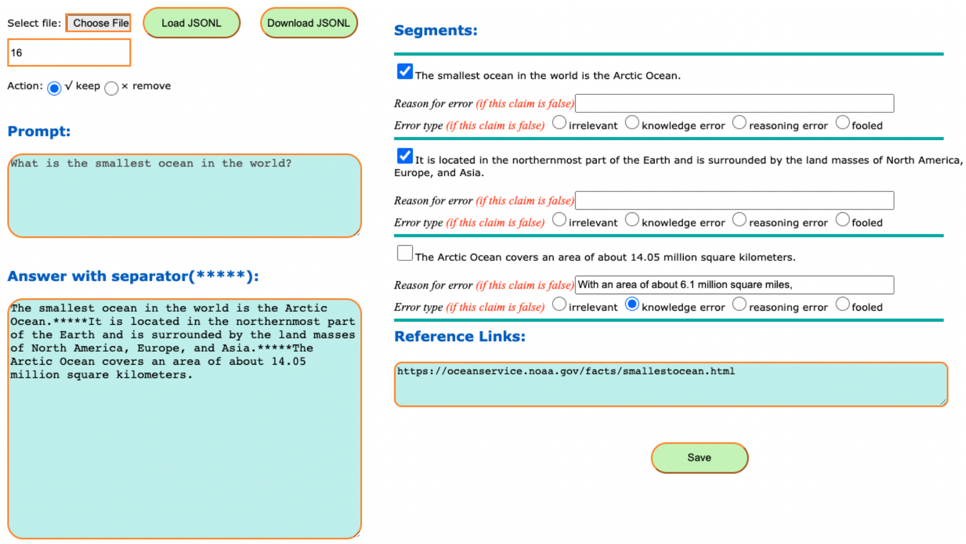

Figure 5: Annotation Page

C Experimental Setup and Prompts

Setup: In our experiments, we use greedy decoding to obtain the results. We established the maximum token limit at 1500 for claim extraction tasks and 100 for factuality detection tasks. For retrieval-augmented methods that use reference documents, we divided the retrieved documents into 512-token chunks and selected the most relevant chunk using the BM25 algorithm. In cases where there were multiple reference links, we concatenated the retrieved chunks.

Prompts: In the following tables, we present the prompts used in our experiments to evaluate the factuality assessment performance of ChatGPT and GPT-4 on world knowledge and writing/recommendation domain. These prompts encompass those used to extract claims from text segments which are shown at Table 6, as well as those used to evaluate the factuality of claims or segments which are shown at Table 7, 8, 9, and 10. We utilized the same extraction and factuality determination prompts for both the world knowledge and writing/recommendation domains, as the response formats are similar in these two domains. This approach allowed us to maintain consistency across both domains, which is important for reliable comparison of the performance of the two models. We use the exact wording of instructions here in our experiments.

Prompts for extracting claims of responses
Prompting methods for extracting claims of responses in world knowledge, writing/recommendation domains: I will show you a question and a list of text segments. The text segments can be concatenated to form a complete answer to the question. Your task is to extract factual claims from each text segment. Here is one example: Question: Tell me about the World Happiness Report. Segments: 1. The World Happiness Report is an annual report published by the United Nations Sustainable Development Solutions Network that ranks countries by their level of happiness or subjective well-being. 2. The report aims to provide policymakers with information and analysis to help them make informed decisions about promoting happiness and well-being in their countries. Below are your outputs: Answer: Segment 1: Claim 1. The World Happiness Report is an annual report. Claim 2. The World Happiness Report is published by the United Nations Sustainable Development Solutions Network. Claim 3. The World Happiness Report ranks countries by their level of happiness or subjective well-being. Segment 2: Claim 1. The World Happiness Report aims to provide policymakers with information and analysis. Claim 2. The World Happiness Report aims to help policymakers make informed decisions. Claim 3. The World Happiness Report aims to help policymakers promote happiness and well-being in their countries. Below are my inputs:
Prompting methods for extracting claims of responses in science/tech domain: I will show you a question and a list of text segments. The text segments can be concatenated to form a complete answer to the question. Your task is to extract factual claims from each text segment. Here is one example: Question: What is the diffusion model in computer science? Segments: 1. In computer science, the diffusion model is a mathematical model used to simulate the spread of information or data through a network or system. 2. It is often used to study phenomena such as the spread of viruses, the adoption of new technologies, or the dissemination of information in social networks. Below are your outputs: Answer: Segment 1: Claim 1. The diffusion model is a mathematical model. Claim 2. The diffusion model is used in computer science. Claim 3. The diffusion model is used to simulate the spread of information or data through a network or system. Segment 2: Claim 1. The diffusion model is often used to study the spread of viruses. Claim 2. The diffusion model is often used to study the adoption of new technologies. Claim 3. The diffusion model is often used to study the dissemination of information in social networks. Below are my inputs:

Table 6: A one-shot prompting example to extract claims for the response segments. We use the exact wording of instructions here in our experiments.

Vanilla prompts for factuality detection in world knowledge and writing/recommendation domains
Segment-based Vanilla Prompting for world knowledge and writing/recommendation: I will show you a question and a list of text segments. All the segments can be concatenated to form a complete answer to the question. Your task is to assess whether each text segment contains factual errors or not. Please generate using the following format: Answer: List the ids of the segments with errors (separated by commas). Please only output the ids, no more details. If all the segments are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Segments: 1. there were a total of 440 operating nuclear power reactors in the world, with a total installed capacity of over 390 gigawatts (GW). 2. These reactors are located in 30 countries around the world, with the highest number of reactors in the United States, followed by France, China, Japan, and Russia. Below are your outputs: Answer: 1,2 It means segment 1,2 contain errors. Below are my inputs:
Claim-based Vanilla Prompting for world knowledge and writing/recommendation: I will show you a question and a list of claims. All the claims are extracted from an answer to the question. Your task is to assess whether each claim contains factual errors or not. Please generate using the following format: Answer: List the ids of the claims with errors (separated by commas). Please only output the ids, no more details. If all the claims are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Claims: 1. There were 440 operating nuclear power reactors in the world. 2. The total installed capacity of these reactors was over 390 gigawatts (GW). 3. The reactors are located in 30 countries around the world. 4. The highest number of reactors is in the United States. 5. France has the second-highest number of reactors. 6. China has a significant number of reactors. 7. Japan has a significant number of reactors. 8. Russia has a significant number of reactors. Below are your outputs: Answer: 1,2,3 It means claim 1,2,3 contain errors. Below are my inputs:

Table 7: Evaluation prompts for one-shot vanilla methods on both segment-based and claim-based settings. We use the exact wording of instructions here in our experiments.

Chain-of-Thought prompts for factuality detection in world knowledge and writing/recommendation domains
Segment-based Chain-of-Thought Prompting for world knowledge and writing/recommendation: I will show you a question and a list of text segments. All the segments can be concatenated to form a complete answer to the question. Your task is to assess whether each text segment contains factual errors or not. Please generate using the following format: Thought: Your reasoning process for the segments with errors. If all the segments are correct, output nothing. Answer: List the ids of the segments with errors (separated by commas). Please only output the ids, no more details. If all the segments are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Segments: 1. there were a total of 440 operating nuclear power reactors in the world, with a total installed capacity of over 390 gigawatts (GW). 2. These reactors are located in 30 countries around the world, with the highest number of reactors in the United States, followed by France, China, Japan, and Russia. Below are your outputs: Thought: For segment 1, there are only 410 operable power reactors in the world, not 440. And the total installed capacity of these reactors was only 368.6 GW, not 390. For Segment 2, the reactors are located in 32 countries around the world, not 30. Answer: 1,2 It means segment 1,2 contain errors. Below are my inputs:
Claim-based Chain-of-Thought Prompting for world knowledge and writing/recommendation : I will show you a question and a list of claims. All the claims are extracted from an answer to the question. Your task is to assess whether each claim contains factual errors or not. Please generate using the following format: Thought: Your reasoning process for the claims with errors. If all the claims are correct, output nothing. Answer: List the ids of the claims with errors (separated by commas). Please only output the ids, no more details. If all the claims are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Claims: 1. There were 440 operating nuclear power reactors in the world. 2. The total installed capacity of these reactors was over 390 gigawatts (GW). 3. The reactors are located in 30 countries around the world. 4. The highest number of reactors is in the United States. 5. France has the second-highest number of reactors. 6. China has a significant number of reactors. 7. Japan has a significant number of reactors. 8. Russia has a significant number of reactors. Below are your outputs: Thought: For claim 1, there are only 410 operable power reactors in the world, not 440. For claim 2, The total installed capacity of these reactors was only 368.6 GW., not 390. For claim 3, the reactors are located in 32 countries around the world, not 30. Answer: 1,2,3 It means claim 1,2 and 3 contain errors. Below are my inputs:

Table 8: Evaluation prompts for one-shot chain-of-thought methods on both segment-based and claim-based settings. We use the exact wording of instructions here in our experiments.

Retrieval-augmented (link) prompts for factuality detection in world knowledge and writing/recommendation domains
Segment-based Retrieval Method with reference links for world knowledge and writing/recommendation: I will show you a question, a list of text segments, and reference links. All the segments can be concatenated to form a complete answer to the question. Your task is to assess whether each text segment contains factual errors or not with the help of the reference doc. Please generate using the following format: Answer: List the ids of the segments with errors (separated by commas). Please only output the ids, no more details. If all the segments are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Segments: 1. there were a total of 440 operating nuclear power reactors in the world, with a total installed capacity of over 390 gigawatts (GW). 2. These reactors are located in 30 countries around the world, with the highest number of reactors in the United States, followed by France, China, Japan, and Russia. Reference Links: https://en.wikipedia.org/wiki/Nuclear_power_by_country, https://en.wikipedia.org/wiki/List_of_commercial_nuclear_reactors Below are your outputs: Answer: 1,2 It means segment 1,2 contain errors. Below are my inputs:
Claim-based Retrieval Method with reference links for world knowledge and writing/recommendation: I will show you a question, a list of claims, and reference links relevant to the question and claims. All the claims are extracted from an answer to the question. Your task is to assess whether each claim contains factual errors or not with the help of the reference links. Please generate using the following format: Answer: List the ids of the claims with errors (separated by commas). Please only output the ids, no more details. If all the claims are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Claims: 1. There were 440 operating nuclear power reactors in the world. 2. The total installed capacity of these reactors was over 390 gigawatts (GW). 3. The reactors are located in 30 countries around the world. 4. The highest number of reactors is in the United States. 5. France has the second-highest number of reactors. 6. China has a significant number of reactors. 7. Japan has a significant number of reactors. 8. Russia has a significant number of reactors. Reference Links: https://en.wikipedia.org/wiki/Nuclear_power_by_country, https://en.wikipedia.org/wiki/List_of_commercial_nuclear_reactors Below are your outputs: Answer: 1,2,3 It means claim 1,2 and 3 contain errors. Below are my inputs:

Table 9: Evaluation prompts for one-shot retrieval-augmented methods with reference links on both segment-based and claim-based settings. We use the exact wording of instructions here in our experiments.

Retrieval-augmented (doc) prompts for factuality detection in world knowledge and writing/recommendation domains
Segment-based Retrieval Method with reference doc for world knowledge and writing/recommendation: I will show you a question, a list of text segments, and a reference doc. All the segments can be concatenated to form a complete answer to the question. Your task is to assess whether each text segment contains factual errors or not with the help of the reference doc. Please generate using the following format: Answer: List the ids of the segments with errors (separated by commas). Please only output the ids, no more details. If all the segments are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Segments: 1. there were a total of 440 operating nuclear power reactors in the world, with a total installed capacity of over 390 gigawatts (GW). 2. These reactors are located in 30 countries around the world, with the highest number of reactors in the United States, followed by France, China, Japan, and Russia. Reference doc: Nuclear power plants operate in 32 countries and generate about a tenth of the world's electricity.[1] Most are in Europe, North America, East Asia and South Asia. The United States is the largest producer of nuclear power, while France has the largest share of electricity generated by nuclear power, at about 70%.[2] China has the fastest growing nuclear power programme with 16 new reactors under construction, followed by India, which has 8 under construction.[3]. As of May 2023, there are 410 operable power reactors in the world, with a combined electrical capacity of 368.6 GW. Below are your outputs: Answer: 1,2 It means segment 1,2 contain errors. Below are my inputs:
Claim-based Retrieval Method with reference doc for world knowledge and writing/recommendation: I will show you a question, a list of claims, and a reference doc relevant to the question and claims. All the claims are extracted from an answer to the question. Your task is to assess whether each claim contains factual errors or not with the help of the reference doc. Please generate using the following format: Answer: List the ids of the claims with errors (separated by commas). Please only output the ids, no more details. If all the claims are correct, output "ALL_CORRECT". Here is one example: Question: What is the total number of nuclear power plants worldwide? Claims: 1. There were 440 operating nuclear power reactors in the world. 2. The total installed capacity of these reactors was over 390 gigawatts (GW). 3. The reactors are located in 30 countries around the world. 4. The highest number of reactors is in the United States. 5. France has the second-highest number of reactors. 6. China has a significant number of reactors. 7. Japan has a significant number of reactors. 8. Russia has a significant number of reactors. Reference doc: Nuclear power plants operate in 32 countries and generate about a tenth of the world's electricity.[1] Most are in Europe, North America, East Asia and South Asia. The United States is the largest producer of nuclear power, while France has the largest share of electricity generated by nuclear power, at about 70%.[2] China has the fastest growing nuclear power programme with 16 new reactors under construction, followed by India, which has 8 under construction.[3]. As of May 2023, there are 410 operable power reactors in the world, with a combined electrical capacity of 368.6 GW. Below are your outputs: Answer: 1,2,3 It means claim 1,2 and 3 contain errors. Below are my inputs:

Table 10: Evaluation prompts for one-shot retrieval-augmented methods with reference doc on both segment-based and claim-based settings. We use the exact wording of instructions here in our experiments.

D Additional Results

We report the balanced accuracy of all the evaluators on both the segment level and the response level under all the settings in §4 at Table 11 and Table 12.

	Method	Overall	World Knowledge	Science/ Tech	Writing/ Rec.	Math	Reasoning
				ChatGPT			
Segment	Vanilla	50.2	50.0	47.8	49.3	50.2	50.5
	Cot	50.8	50.2	51	49.8	51.7	50.2
	Link	51.3	51.5	50.1	49.9	–	–
	Doc	52.3	52.9	51	49.9	–	–
Claim	Vanilla	51.4	51.3	52.4	50.5	–	–
	Cot	50.2	50.4	49.8	49.3	–	–
	Link	53.2	53.4	**57.6**	48.8	–	–
	Doc	53.0	54.7	51.6	50.4	–	–
				GPT-4			
Segment	Vanilla	62.4	62.0	55.2	50.8	61.5	70.5
	Cot	**65.9**	66.5	57.1	51.0	**61.6**	**77.7**
	Link	64.5	66.6	58.1	**55.9**	–	–
	Doc	65.1	**68.2**	**60.8**	51.7	–	–
Claim	Vanilla	60.6	58.1	54.8	48.3	–	–
	Cot	61.8	63	54.3	48.4	–	–
	Link	61.8	60.6	56.3	49.8	–	–
	Doc	**62.9**	**63.4**	56.1	**54.5**	–	–

Table 11: Segment-level balanced accuracy of factual error detectors powered by ChatGPT and GPT-4 on FELM. We do not involve claim-based methods for math and reasoning domains cause it is often difficult to extract self-contained, atomic claims from these two domains. There is no reference for math and reasoning either. To compute the overall average for "Link" and "Doc", we account for the vanilla numbers for math and reasoning domains since these two methods degenerate to vanilla in this case. For claim-based method, we use segment-based numbers on math and reasoning domains to compute the overall average since claim-based method degenerates to segment-based in these domains. We bold the best results by claim-based methods and by segment-based methods.

	Method	Overall	World Knowledge	Science/ Tech	Writing/ Rec.	Math	Reasoning
				ChatGPT			
Segment	Vanilla	49.8	49.3	48.3	43.8	50.1	51.3
	Cot	50.9	51.3	52.6	47.2	50.9	50.1
	Link	51.9	51.3	50.1	48.9	–	–
	Doc	53.6	53.6	49.9	48.9	–	–
Claim	Vanilla	49.6	49.5	52.1	52.3	–	–
	Cot	50.2	51.3	49.4	48.1	–	–
	Link	53.3	55	**59.5**	39.9	–	–
	Doc	52.9	52.6	52.4	48.1	–	–
				GPT-4			
Segment	Vanilla	64.5	63.7	59.3	51.3	61.1	79.9
	Cot	**67.9**	69.6	**61.9**	52.4	**61.9**	**82.7**
	Link	66.4	67.7	61.2	**54.8**	–	–
	Doc	66.3	**69.7**	57.4	53.1	–	–
Claim	Vanilla	62.2	60.9	58.6	52	–	–
	Cot	62.4	64.8	57.5	54.4	–	–
	Link	64.1	66.3	56.3	51.2	–	–
	Doc	**65.2**	**66.5**	57.7	**60.0**	–	–

Table 12: Response-level balanced accuracy of factual error detectors powered by ChatGPT and GPT-4 on FELM. We do not involve claim-based methods for math and reasoning domains cause it is often difficult to extract self-contained, atomic claims from these two domains. There is no reference for math and reasoning either. To compute the overall average for "Link" and "Doc", we account for the vanilla numbers for math and reasoning domains since these two methods degenerate to vanilla in this case. For claim-based method, we use segment-based numbers on math and reasoning domains to compute the overall average since claim-based method degenerates to segment-based in these domains. We bold the best results by claim-based methods and by segment-based methods.