AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators

Xingwei He[1]*   Zhenghao Lin[2]   Yeyun Gong[3]   A-Long Jin[1]   Hang Zhang[3],
Chen Lin,[2]   Jian Jiao,[4]   Siu Ming Yiu,[1]   Nan Duan,[3]   Weizhu Chen[4]
[1]The University of Hong Kong, [2]Xiamen University,
[3]Microsoft Research Asia, [4]Microsoft
hexingwei15@gmail.com, ajin@eee.hku.hk, smyiu@cs.hku.hk,
zhenghaolin@stu.xmu.edu.cn, chenlin@xmu.edu.cn,
{yegong, v-zhhang, jian.jiao, nanduan, wzchen}@microsoft.com

Abstract

Many natural language processing (NLP) tasks rely on labeled data to train machine learning models to achieve high performance. However, data annotation can be a time-consuming and expensive process, especially when the task involves a large amount of data or requires specialized domains. Recently, GPT-3.5 series models have demonstrated remarkable few-shot and zero-shot ability across various NLP tasks. In this paper, we first claim that large language models (LLMs), such as GPT-3.5, can serve as an excellent crowdsourced annotator by providing them with sufficient guidance and demonstrated examples. To make LLMs to be better annotators, we propose a two-step approach, 'explain-then-annotate'. To be more precise, we begin by creating prompts for every demonstrated example, which we subsequently utilize to prompt a LLM to provide an explanation for why the specific ground truth answer/label was chosen for that particular example. Following this, we construct the few-shot chain-of-thought prompt with the self-generated explanation and employ it to annotate the unlabeled data. We conduct experiments on three tasks, including user input and keyword relevance assessment, BoolQ and WiC. The annotation results from GPT-3.5 surpasses those from crowdsourced annotation for user input and keyword relevance assessment. Additionally, for the other two tasks, GPT-3.5 achieves results that are comparable to those obtained through crowdsourced annotation.

1 Introduction

Labeled data refers to a dataset that has been manually annotated with predefined target labels or categories. It is crucial to develop machine learning models for many NLP tasks, such as sentiment analysis (Socher et al., 2013), machine translation

(Sutskever et al., 2014) and word sense disambiguation (He and Yiu, 2022). The process of labeling data is typically done by human annotators who are given specific guidelines and criteria on how to assign labels to each instance in the dataset. For example, in sentiment analysis, each sentence or document may be labeled with a polarity score such as "positive", "negative", or "neutral". However, it is very labor-intensive and time-consuming to create a large dataset with human annotation, which limits the availability of such data and the applicability of machine learning models in various NLP tasks.

Previous works have demonstrated that large-scale pre-trained language models (LLMs), such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022), achieve impressive results in various downstream tasks without collecting large-scale task-specific data or tuning model parameter but only with a few examples as instructions. OpenAI has recently launched the GPT-3.5 series, which are the upgraded versions of GPT-3, and trained on a blend of text and code published before the end of 2021. In the meanwhile, OpenAI also unveiled ChatGPT, another fine-tuned version of GPT-3.5. Within just two months since its release, ChatGPT has garnered a massive following of 100 million users worldwide, garnering significant global attention.

Wang et al. (2021) showed that augmenting manually labeled data with pseudo-labeled data from GPT-3 could enhance the performance of models, particularly when the labeling budget is restricted. However, the quality of GPT-3's labeled data still lags behind that of manually labeled data. Considering the GPT-3.5 model's remarkable few-shot and zero-shot capabilities across various tasks and the expensive nature of manual annotation, we raise an essential and significant inquiry: Can GPT-3.5 potentially replace crowdsourced annotators?

Before answering this question, let us go over

---
*Work in progress

the process of crowdsourced data annotation. First, we need to provide the annotators with a specific definition of the task. Then, for classification tasks, we need to tell the annotators the specific meanings of each category. Finally, we need to provide the annotators with a few examples that have already been annotated as references. Naturally, we can guide GPT-3.5 to annotate data using the same approach as with human annotators by providing it with task definitions and example samples. Furthermore, we found that requesting a LLM to furnish the rationale behind the ground truth label or answer for a particular example can prompt the LLM to produce high-quality explanations. Based on this, we construct the few-shot chain-of-thought prompt (Wei et al., 2022) with the self-generated explanations to annotate data. We refer to this method as 'explain-then-annotate', which further improves the annotation quality.

We summarize our contributions as follows:
(1) For the first time, we propose that AnnoLLM, a **Anno**tation system powered by **L**arge **L**anguage **M**odels, can replace crowdsourced annotators to annotate data.
(2) To enhance the data annotation capabilities of LLMs, we suggest a two-step approach called 'explain-then-annotate'. In this approach, we leverage ChatGPT to generate a few-shot chain-of-thought prompt, which we then use to annotate unlabeled data.
(3) Our results on three datasets verify the feasibility of substituting crowdsourced annotators with GPT-3.5[1], where it either surpasses or matches crowdsourced annotators.

2 Approach

Providing detailed instructions to annotators is crucial when using crowdsourcing to annotate data, as it can help crowd workers better understand task requirements and annotation standards, ultimately improving the quality and accuracy of annotated data. The instructions for each task mainly includes three parts: task description, category definition, and demonstrated examples.

Motivated by the guidance to human annotators, we will introduce how to convert GPT-3.5 into a zero-shot data annotator by providing guidance on the task description and category definitions in Section 2.1. Then, we will show how to transform

GPT-3.5 into a few-shot data annotator with the demonstrated examples in Section 2.2. We show the crowdsourcing annotation and our proposed 'explain-then-annotate' processes in Figure 1.

2.1 GPT-3.5 as a Zero-shot Data Annotator

In the zero-shot setting, we can only provide the task description and category definitions to annotators. The task description includes information on task definition, task purpose, and so on. Category definitions require clear definitions for each category, so that the crowd workers can understand the meaning and standard of each category.

In the same vein, we furnish GPT-3.5 with the task description and category definitions, which enable GPT-3.5 to function as a zero-shot data annotator. We show the zero-shot prompts for GPT-3.5 on the user query and keyword relevance assessment, WiC and BoolQ tasks in Tables 10, 11 and 12, respectively.

2.2 GPT-3.5 as a Few-shot Data Annotator

When annotating data, providing annotation examples for each category to annotators can help crowd workers better understand how to annotate and classify the data accurately. Similarly, we also provide the demonstrated examples to GPT-3.5, thus changing it into a few-shot data annotator. We show the few-shot prompts for GPT-3.5 on the user query and keyword relevance assessment, WiC and BoolQ tasks in Tables 13, 14 and 15, respectively.

Recent work (Wei et al., 2022) has shown that adding human written rationales to the demonstrated examples, called as chain-of-thought (CoT), can elicit the LLMs' reasoning ability, thus gaining improvements on reasoning tasks. In this paper, we find that GPT-3.5 [2] is a good reasoner who can automatically generate reasonable explanations for demonstrated examples. In the following, we will introduce how to generate explanations with GPT-3.5, and then create few-shot CoT prompts with the generated explanations.

Generating Explanations with GPT-3.5. In this step, we simulate the way humans explain problems to induce GPT-3.5 to explain the annotated examples. To be concrete, we present the task description, specific example, and the corresponding true labels to GPT-3.5, and then ask it to answer why the corresponding answer for that example is

---

[1] In this paper, we focus on using GPT-3.5 series models to annotate data for classification tasks.

[2] We resort to ChatGPT, the latest version of GPT-3.5, to generate explanations.
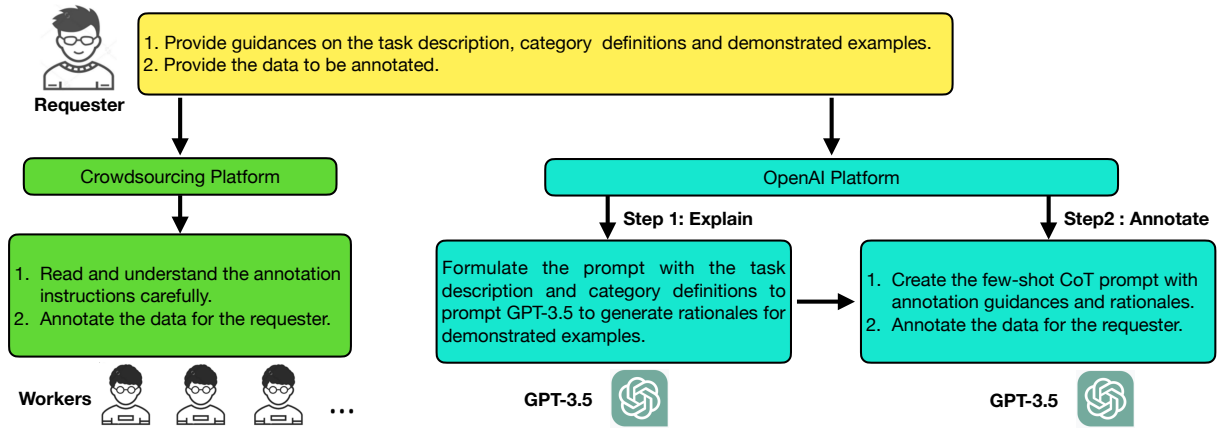
Figure 1: On the left is the annotation process used by crowdsourced workers, while on the right is AnnoGPT's process. AnnoGPT mimics the manual annotation process, with the exception that it generates explanations for each example before annotation. This ensures that each demonstrated example is accompanied by helpful explanations, making the annotation guidelines more informative and useful."

the given label. By doing so, GPT-3.5 will generate reasonable explanations. For the user query and keyword relevance assessment task, we show how to use GPT-3.5 to explain why the label between the user query "**google data studio sharepoint**" and the keyword "**sharepoint migration tool file share**" is "**Bad**" in Table 1. Please refer to Table 8 and Table 7 for how to generate explanations for the demonstrated examples of BoolQ and WiC.

Creating Few-shot CoT Prompts. After getting the explanations generated by GPT-3.5, we can construct the few-shot CoT prompt. We show the few-shot CoT prompts for GPT-3.5 on the user query and keyword relevance assessment, WiC and BoolQ tasks in Tables 16, 17 and 18, respectively.

3 Experiment

3.1 Tasks and Datasets

We evaluate AnnoLLM on three different tasks, including the user query and keyword relevance assessment, BoolQ and WiC. The basic statistics of these datasets are shown in Table 2.

The QK (user query and keyword relevance assessment) task aims to judge whether the user input query is related to the given keywords.

BoolQ, which stands for Boolean Questions and was introduced by Clark et al. (2019), is a question-answering task. In this task, each example comprises a brief passage and a yes/no question related to the passage. The users of the Google search engine anonymously and without solicitation submit the questions, which are then matched with a paragraph from a Wikipedia article that provides

the answer.

The WiC (Word-in-Context) task, developed by Pilehvar and Camacho-Collados (2019), involves disambiguating word senses through binary classification of sentence pairs. In this task, two text snippets are provided along with a polysemous word that occurs in both sentences. The objective is to determine whether the target word shares the same sense in both sentences.

Since all three tasks are binary classification tasks, accuracy is used to evaluate their results.

3.2 Human Performances

To evaluate the human performance on the QK dataset, we use UHRS[3], a crowdsourcing platform, to annotate this data. Before annotation, we first present the annotators with the task description, category definitions, and annotated examples. Then, we invite three annotators to label a data instance. If the annotated results of the three workers are consistent, then this result will be considered as the annotated label. Otherwise, we would invite other annotators to continue annotating this data instance until three people have consistent annotation results. We require the crowdsourced annotators to annotate all development and test sets.

BoolQ and WiC are two of the most challenging datasets in superGLUE (Wang et al., 2019). In the case of BoolQ, three authors labeled 110 randomly chosen examples, with human performance achieving 89%. Regarding WiC, Pilehvar and Camacho-Collados (2019) selected four groups of 100 instances from the test set, assigning each

---

[3] https://prod.uhrs.playmsn.com/uhrs/

Input	Directions: Given a search engine query: "**google data studio sharepoint**", first, consider what the user could have in mind when they type in the query and allow for misspellings or other ambiguity, then classify the relevance of keyword: "**sharepoint migration tool file share**" to the query into one of the following categories: "Not bad", "Bad". The definitions of the categories are "Not bad": the keyword is relevant to the user's search query. This can include: broader or narrower product selection, competitor or alternative products, accessories, products often purchased together and related topics as well as direct matches to the user's search. "Bad": the keyword is not relevant to the user's search query. There is no relationship between the query and keyword. This includes but is not limited to: incorrect/unrelated product, unrelated topic, wrong location when location is important, cannot be used in place of query product nor are they commonly purchased together, etc. Briefly explain why the relevance is "**Bad**", with a response length not exceeding 100 words.
Output 1	The relevance of the keyword "sharepoint migration tool file share" to the search query "google data studio sharepoint" is "Bad". This keyword is not relevant to the user's search query as it pertains to a different tool or technology (SharePoint) that is not mentioned in the search query. The keyword focuses on a migration tool for SharePoint file sharing, while the search query is focused on Google Data Studio. There is no relationship between the two, so the keyword is not relevant to the user's search query.
Output 2	The relevance of "sharepoint migration tool file share" to the search query "google data studio sharepoint" is "Bad." The keyword is not relevant to the user's search query because it deals with SharePoint migration and file sharing, whereas the user's query is focused on Google Data Studio and SharePoint. There is no relationship between these two topics, and the keyword cannot be used in place of the query product.
Output 3	The relevance is "Bad". The keyword "sharepoint migration tool file share" is not directly related to the query "google data studio sharepoint". The keyword refers to tools and services for migrating data from SharePoint to another file share platform, whereas the query is about using Google Data Studio to connect with SharePoint data. There is no clear relationship between the two topics.
Output 4	The relevance of the keyword "sharepoint migration tool file share" to the search query "google data studio sharepoint" is "Bad". The keyword is about a tool for migrating data from SharePoint, which is not directly related to Google Data Studio. Google Data Studio is a data visualization and reporting tool, not a migration tool for SharePoint.
Output 5	The relevance is "Bad" as the keyword "sharepoint migration tool file share" has no direct relation to the user's search query "google data studio sharepoint". The keyword pertains to a tool used for migrating data within a SharePoint platform, while the query is focused on Google Data Studio and SharePoint. These two concepts are not directly related and therefore, the keyword is not relevant to the user's search query.

Table 1: Illustration of how to use a LLM to generate explanations for the user query and keyword relevance assessment task. "Input" and "Output" refer to the prompt provided to the LLM and the generated explanations, respectively.

Task / Partition	Dev	Test
QK	350	1000
BoolQ	3270	3245
WiC	638	1400

Table 2: The basic statistics of the QK, BoolQ and WiC datasets.

group to an annotator, and achieving a human performance of 80%.

3.3 Experimental Results

User Query and Keyword Relevance Assessment. Table 3 shows our experimental results on the QK development and test sets. Surprisingly, GPT-3.5 performs worse under the few-shot setting than under the zero-shot setting on this task. Fu and Khot (2022) speculate that the instruction tuning on GPT-3.5 may decrease the model's in-context learning ability but increase the model's zero-shot

ability.

On the other hand, GPT-3.5 with a 4-shot CoT prompt outperforms its counterparts under the zero-shot and few-shot settings by around 6 and 8 points, respectively. Impressively, it even surpasses the crowdsourced annotators. The experimental results presented herein provide compelling evidence of the effectiveness of our proposed method, AnnoLLM. Previous studies (Wei et al., 2022; Wang et al., 2022) have shown that the few-shot CoT, constructed using human-written explanations, can enhance the model's reasoning ability on reasoning tasks. However, our approach differs from previous methods in that we utilize explanations generated by the large model itself, thereby allowing the model's reasoning ability to emerge. Additionally, we have demonstrated, for the first time, the effectiveness of the CoT method on tasks other than typical reasoning tasks.

Models	Dev Set	Test Set
Crowdsourced Annotator	65.58	71.5
GPT-3.5 + zero-shot	67.71	70.00
GPT-3.5 + 8-shot	65.71	67.80
GPT-3.5 + 4-shot CoT	**74.17***	**75.60***

Table 3: Evaluation results on the QK task. Accuracy is used as the evaluation metric. Results marked with an asterisk (*) represent the average result of five few-shot CoT prompts constructed with different generated explanations.

Word Sense Disambiguation. Table 4 presents our experimental results on the WiC development and test sets, from which we also see that An-noLLM, i.e., GPT-3.5 + 8-shot CoT, outperforms its few-shot counterpart significantly. Nevertheless, there remains a considerable disparity between An-noLLM and crowdsourced annotators. This is primarily due to the task's inherent complexity, and currently, even the supervised SOTA models still exhibit a substantial gap compared to manual annotation.

Question Answering. As shown in Table 5, An-noLLM surpasses crowdsourced annotators on the BoolQ development and test sets, but does not show significant improvement compared to the few-shot method. However, this does not imply that CoT with generated explanation is not useful for this task. In Section 3.5, we found that AnnoLLM with CoT exhibits better stability across various prompts, while its counterpart with the few-shot setting is highly sensitive to templates.

3.4 Ablation Study

In this section, we perform an ablation study to compare the effect of different methods used to generate explanations on AnnoLLM's performance.

Firstly, we want to investigate whether the use of ground truth labels is beneficial for generating explanations for demonstrated examples. To answer this question, we induce LLM to generate explanations using prompts that contain ground truth labels, and prompts that do not contain labels (we only replace the last sentence of the original prompt "*Briefly explain why the relevance is "Bad", with a response length not exceeding 100 words.*" with "*Briefly explain the relevance between the keyword and query, with a response length not exceeding 100 words.*"). From Table 6, we found that not us-

Models	Dev Set	Test Set
Crowdsourced Annotator	80.0	80.0
Zero/Few-shot		
PaLM 540B + zero-shot	59.1‡	-
PaLM 540B + 5-shot	64.6‡	-
GPT-3.5 + zero-shot	57.52	59.79
GPT-3.5 + 8-shot	67.71	66.36
GPT-3.5 + 8-shot CoT	**71.47***	**69.17***
Fine-tune		
T5 11B (Raffel et al., 2020)	77.3‡	76.9†
PaLM 540B	78.8‡	77.4†
ST-MoE 32B (Zoph et al., 2022)	**81.0**‡	**77.7**†

Table 4: Evaluation results on the WiC task. Accuracy is used as the evaluation metric. Results marked with † and ‡ are from the official SuperGLUE leaderboard[4] and PaLM (Chowdhery et al., 2022), respectively. Results marked with an asterisk (*) represent the average result of five few-shot CoT prompts constructed with different generated explanations. The number behind the model represents the size of the model's parameters.

ing true labels when generating explanations will impair the performance of AnnoLLM by about 3 points on the QK test set (row 4 vs. row 1). This is because the model may give explanations for an incorrect answer without the guidance of ground labels. Please refer to Table 1 and Table 9 for the specific prompts and generated explanations.

Based on the method in the fourth row, we filter out the incorrect explanations using true labels and retain three correct explanations for three out of four demonstrated samples. However, for the third sample, all five generated explanations are wrong, and we have to keep three incorrect explanations for this demonstrated example. That is why using the golden label to filter explanations does not bring significant improvement (row 5 vs. row 4).

As shown in Table 1, we found that LLM reveals the true answer at the beginning of the generated explanation, and then provides an explanation for it. This is different from previous work (Wei et al., 2022; Wang et al., 2022), which induces the model to provide an explanation first and then output the answer. Therefore, we remove the beginning sentence containing the label from the generated explanations (the underlined text in Table 16). However, this does not lead to improvement (row 2 vs. row 1). We speculate that this may be due to the difference in the nature of our task and traditional reasoning tasks. We remove the last sentence containing the answer to the demonstrated examples

Models	Dev Set	Test Set
Crowdsourced Annotator	89.0	89.0
Zero/Few-shot		
GPT-3 175B + zero-shot	60.5	-
Gopher 280B + zero-shot (Rae et al., 2021)	79.3	-
Chinchilla 70B + zero-shot (Hoffmann et al., 2022)	83.7	-
PaLM 62B + zero-shot	84.8	-
PaLM 540B + zero-shot	88.0	-
LLaMA 65B + zero-shot (Touvron et al., 2023)	85.3	-
GPT-3.5 + zero-shot	84.28	84.30
GPT-3.5 + 8-shot	89.17	89.10
GPT-3.5 + 8-shot CoT	**89.69**	**89.20**
Fine-tune		
T5 11B (Raffel et al., 2020)	90.8‡	91.2†
PaLM 540B	92.2‡	91.9†
ST-MoE 32B (Zoph et al., 2022)	**93.1**‡	**92.4**†

Table 5: Evaluation results on the BoolQ task. Accuracy is used as the evaluation metric. Results marked with † and ‡ are from the official SuperGLUE leaderboard and PaLM, respectively. The number behind the model represents the size of the model's parameters.

(the Italicized text in Table 16), yet it does not have too much impact on the performance (row 3 vs. row 1). That is because the generated explanations contains the gold answer. To be in line with the prompt format of previous work (Wei et al., 2022), we still append the ground truth label to the generated explanation.

3.5 More Analysis and Discussion

Consistency Analysis of Generated Explanations. In the ablation study, we found that the performance of AnnoLLM is highly dependent on the generated explanations. This leads to a natural inquiry: Are the explanations produced by the LLM consistent enough for the same demonstrated sample? To answer this question, we generate five explanations for each demonstrated sample, and obtain five different few-shot CoT prompts. The results obtained using different few-shot CoT prompts are presented in Figure 2 (a). It can be observed that different few-shot CoT prompts achieve similar performance in the QK, WiC, and BoolQ tasks. This indicates that the quality of the explanations generated by the LLM is sufficiently consistent.

Stability Analysis of Generated Explanations. From Figure 2 (a), we can see that AnnoLLM with few-shot CoT prompts significantly outperforms its

counterpart with the few-shot setting, with a high margin of 5 percentage points on the QK and WiC datasets in most cases. However, the improvement is quite modest on BoolQ, where it is mostly less than 0.5. This does not mean that AnnoLLM with few-shot CoT prompts has no effect on the BoolQ task.

To further analyze this, we make slight modifications to the existing prompts for BoolQ to obtain three corresponding few-shot CoT and few-shot prompts (Please refer to Appendix E for the few-shot and few-shot CoT prompts). The experimental results in Figure 2 show that the few-shot method is very sensitive to the templates, since even with slight modifications to the templates, the experimental performances will drop from around 89 to below 80 points. In comparison, AnnoLLM with few-shot CoT prompts suffer less performance loss, and we found that in these cases, AnnoLLM with few-shot CoT prompts outperforms its counterpart with few-shot templates by around 4 points. Therefore, we conclude that few-shot settings are more picky about templates, but few-shot CoT exhibits better stability across different templates.

4 Related Work

4.1 Large-scale Pre-trained Language Models

GPT (Generative Pre-trained Transformer) is a family of language models developed by OpenAI, designed to generate human-like natural language text. The GPT models are based on the Transformer architecture (Vaswani et al., 2017), which are pre-trained on an enormous corpus of text data with the standard language modeling objective to predict the next token based on the previous context. OpenAI has continuously increased the parameters and training data of its models, and has released GPT (Radford, 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020) from 2018 to 2020. GPT-3 is a powerful language model with 175 billion parameters, and make a significant advance in the field of NLP. One of the unique features of GPT-3 is its ability to perform in-context learning, where one can apply it to various tasks by simply providing few-shot demonstrations without any gradient updates or fine-tuning. Furthermore, OpenAI fine-tuned GPT-3 on the code data or instruction data, and released Codex (Chen et al., 2021) and InstructGPT (Ouyang et al.), respectively. Recently, OpenAI released the GPT-3.5 series models, in-

Variants of GPT-3.5 + 4-shot CoT					Datasets	
#	Generate E with L	Delete L from E	Filter E with L	Append L to L	Dev Set	Test Set
1	✓			✓	**74.17**	**75.60**
2	✓	✓		✓	72.97	74.76
3	✓				74.09	75.44
4				✓	72.63	72.84
5			✓	✓	$73.05^{\dagger}$	$73.20^{\dagger}$

Table 6: Ablation study on the QK task. 'E' and 'L' refer to the generated explanations and ground truth labels, respectively. Results marked with a dagger (†) represent the average outcome of three few-shot CoT prompts, each constructed with different generated explanations, while the remaining results represent the average of five.

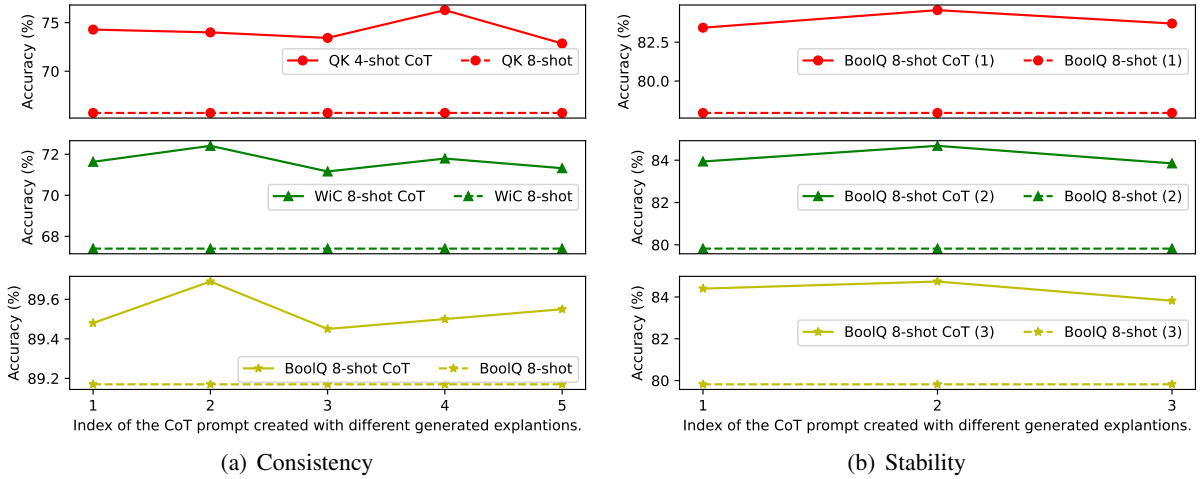(a) Consistency

(b) Stability

Figure 2: Subfigure (a) shows the performance on development sets for different few-shot CoT prompts, created with different explanations. Subfigure (b) shows the performance for different few-shot and few-shot CoT prompts on the development set of BoolQ.

cluding text-davinci-003 and ChatGPT, by training on text and code data, then tuning with supervised instructions and reinforcement learning with human feedback. Some recent work has shown that GPT-3.5 models have strong few-shot and zero-shot learning ability on various NLP tasks, such machine translation (Jiao et al., 2023), and information extraction (Wei et al., 2023).

In this paper, we first propose that we can readily change GPT-3.5 to a good data annotator for a specific task by providing the detailed annotation instructions similar to human annotators.

4.2 Pseudo Annotated Data

Creating human-annotated data is tedious and costly , particularly for complex tasks or specialized domains where there may not be sufficient data available. Therefore, creating pseudo-annotated data has been widely used to generate labeled data for a specific task when there is a limited amount of annotated data available. Back-translation involves translating a target language sentence back

into the source language, which is first proposed to improve neural machine translation models with synthetic parallel data (Sennrich et al., 2016). Beyond machine translation, the technique of back-translation has been applied to unsupervised text style transfer (Prabhumoye et al., 2018) and image style transfer (Zhu et al., 2017). In addition, rule-based methods are also widely used to construct synthetic data. For example, Zhang et al. (2020) resort to the lead bias to create paired data to pre-train the text summarization model, PEGASUS. Lee et al. (2019) pre-trained the retriever with the Inverse Cloze Task, which aims to predict the context based on the given sentence.

However, these methods are designed for specific tasks and it is difficult to generalize them to other tasks. In this paper, we study how to transform the powerful language model GPT-3.5 into a general-purpose data annotator. By providing the corresponding task description and few-shot chain-of-thought demonstrations, it can be easily used to

annotate data for various tasks.

5 Conclusion

In this paper, we introduce AnnoLLM, a novel annotation system powered by LLMs that has the potential to replace traditional crowdsourced annotators. Additionally, we propose a two-step approach, 'explain-then-annotate', to enhance the data annotation capabilities of LLMs. The approach leverages LLMs to generate a few-shot chain-of-thought prompt, which is then used to annotate unlabeled data. Our experimental results on three datasets demonstrate the feasibility of using LLMs to substitute crowdsourced annotators, which highlights the potential to facilitate the development of using LLMs like GPT-3.5 to annotate data for various NLP tasks.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Xingwei He and Siu Ming Yiu. 2022. Controllable dictionary example generation: Generating example sentences for specific targeted audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627, Dublin, Ireland. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training. *OpenAI Technical Report*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. volume 32.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. Designing effective sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Prompts for Generating Explanations

Input	Task description: The task is to identify if the occurrences of "place" in '**Do you want to come over to my "place" later?**' and '**A political system with no "place" for the less prominent groups.**' correspond to the same meaning or not. The target word "place" in the first sentence and the second sentence is enclosed in quotation marks. There is no need to consider the tense and form of "place". If "place" in the first sentence and the second sentence correspond to the same meaning, output "true", otherwise output "false". Briefly explain why the answer is "**false**", with a response length not exceeding 100 words.
Output 1	The answer is "false". In the first sentence, "place" refers to a physical location or a person's home. In the second sentence, "place" refers to a position or status in a political system. Therefore, although the spelling and punctuation of "place" are the same in both sentences, the context and meaning of the word are different.
Output 2	The answer is "false". In the first sentence, "place" refers to a physical location, meaning a house or an apartment. In contrast, in the second sentence, "place" refers to a figurative location, meaning a position or role within a political system. Therefore, the two occurrences of "place" have different meanings and do not correspond to the same meaning.
Output 3	The answer is "false". In the first sentence, "place" refers to a physical location or someone's home, while in the second sentence, "place" refers to a position or status in a political system. These are different meanings, and therefore, the occurrences of "place" in the two sentences do not correspond to the same meaning.
Output 4	The answer is "false". In the first sentence, "place" refers to a physical location, as in a person's home. In contrast, in the second sentence, "place" refers to a position or status within a system or society, as in a group's role or standing. Thus, the two uses of "place" have different meanings, and do not correspond to the same meaning.
Output 5	The answer is "false". In the first sentence, "place" refers to a physical location where the speaker is inviting someone to come over to. In contrast, in the second sentence, "place" refers to a position or role in a political system. Therefore, the two occurrences of "place" have different meanings and cannot be considered to correspond to the same meaning.

Table 7: Illustration of how to use LLM to generate explanations for the WiC task. "Input" and "Output" refer to the prompt provided to LLM and the generated explanations, respectively.

Input	Task description: Based on the information provided in the passage: "**The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.**", determine if the answer to the question: "**is elder scrolls online the same as skyrim**" is "yes" or "no". If the answer is affirmative, indicate "true", otherwise indicate "false". Briefly explain why the answer is "**false**", with a response length not exceeding 100 words.
Output 1	The answer is "false". Although The Elder Scrolls Online is set on the continent of Tamriel, like Skyrim and other games in The Elder Scrolls series, it takes place a millennium before Skyrim and around 800 years before Morrowind and Oblivion. While it has a similar structure to Skyrim, with two conflicts progressing at the same time, the primary struggle in The Elder Scrolls Online is against the Daedric Prince Molag Bal, who has stolen the player character's soul, and the second is to capture the vacant imperial throne. Therefore, The Elder Scrolls Online is not the same game as Skyrim.
Output 2	The answer is "false". While The Elder Scrolls Online is set in the same continent of Tamriel as Skyrim, and shares a similar structure with two conflicts happening simultaneously, the events of the game occur a millennium before Skyrim, and its primary antagonist is the Daedric Prince Molag Bal, not Alduin as in Skyrim. The gameplay mechanics and overall experience of The Elder Scrolls Online are also distinct from Skyrim, as it is a massively multiplayer online role-playing game (MMORPG), while Skyrim is a single-player game.
Output 3	The answer is "false". While The Elder Scrolls Online and Skyrim share the same fictional universe and are part of The Elder Scrolls series, they are not the same game. The Elder Scrolls Online is set a millennium before Skyrim and has a different storyline and gameplay mechanics. In The Elder Scrolls Online, the player fights against Molag Bal and attempts to recover their stolen soul, while in Skyrim, the player character is a Dragonborn who fights against Alduin, the World-Eater. Additionally, The Elder Scrolls Online is a massively multiplayer online game, while Skyrim is a single-player game.
Output 4	The answer is "false". While The Elder Scrolls Online and Skyrim share some similarities, such as being set in the same continent of Tamriel and having a similar structure, they are different games with different stories and timelines. The events in The Elder Scrolls Online take place a millennium before Skyrim and revolve around two conflicts, one against the Daedric Prince Molag Bal and the other for the vacant imperial throne, while Skyrim has its own unique story centered around the return of the dragons to the world.
Output 5	The answer is "false". While The Elder Scrolls Online and Skyrim share the same fictional world of Tamriel, they are not the same game. The events of The Elder Scrolls Online take place a millennium before Skyrim, and the game features a different storyline and game structure. In The Elder Scrolls Online, the player character is sacrificed to the Daedric Prince Molag Bal and must recover their soul, while in Skyrim, the player character is a Dragonborn on a quest to defeat the dragon Alduin and save Tamriel from destruction.

Table 8: Illustration of how to use a LLM to generate explanations for the BoolQ task. "Input" and "Output" refer to the prompt provided to the LLM and the generated explanations, respectively.

Input	Directions: Given a search engine query: "**google data studio sharepoint**", first, consider what the user could have in mind when they type in the query and allow for misspellings or other ambiguity, then classify the relevance of keyword: "**sharepoint migration tool file share**" to the query into one of the following categories: "Not bad", "Bad". The definitions of the categories are "Not bad": the keyword is relevant to the user's search query. This can include: broader or narrower product selection, competitor or alternative products, accessories, products often purchased together and related topics as well as direct matches to the user's search. "Bad": the keyword is not relevant to the user's search query. There is no relationship between the query and keyword. This includes but is not limited to: incorrect/unrelated product, unrelated topic, wrong location when location is important, cannot be used in place of query product nor are they commonly purchased together, etc. Briefly explain the relevance between the keyword and query, with a response length not exceeding 100 words.
Output 1	The user's query "google data studio sharepoint" suggests they are looking for information on how to integrate or connect Google Data Studio with SharePoint, possibly for data visualization or reporting purposes. The keyword "sharepoint migration tool file share" is not directly relevant to the user's query as it pertains to a tool used for migrating files between SharePoint instances, rather than integrating Google Data Studio with SharePoint. Therefore, the keyword is considered "Bad" in relation to the user's query.
Output 2	When a user searches for "Google Data Studio SharePoint," they are likely looking for information on how to integrate Google Data Studio with SharePoint or how to use Google Data Studio to create reports on SharePoint data. The keyword "SharePoint migration tool file share" is not directly relevant to the query as it pertains to a tool for migrating files to SharePoint rather than using Google Data Studio with SharePoint. Therefore, the keyword is considered "bad" in this context.
Output 3	When a user searches for "google data studio sharepoint," they are likely looking for information on how to integrate or connect Google Data Studio with SharePoint, a popular document management and collaboration platform. The keyword "sharepoint migration tool file share" is somewhat relevant to the query, as it pertains to migrating files from one SharePoint instance to another, which could be useful information for someone looking to integrate the two platforms. Therefore, I would classify this keyword as "Not bad."
Output 4	The search query "google data studio sharepoint" likely suggests that the user is looking for information about how to integrate or use Google Data Studio with SharePoint. They may be interested in creating reports or visualizations using data from SharePoint in Google Data Studio. The keyword "sharepoint migration tool file share" is not relevant to the query as it refers to a tool for migrating files from one SharePoint site to another, and does not address the query's focus on integrating Google Data Studio with SharePoint. Therefore, the keyword is considered "Bad" for this search query.
Output 5	The search engine query "google data studio sharepoint" suggests that the user may be looking for information on how to integrate or use Google Data Studio with SharePoint, which is a web-based collaborative platform used for document management and storage. The keyword "sharepoint migration tool file share" is not directly relevant to the user's search query, as it is related to a tool used for migrating files between different file-sharing platforms, and does not address the user's original intent of using Google Data Studio with SharePoint. Therefore, the keyword is classified as "Bad."

Table 9: Illustration of how to use a LLM to generate explanations for the user query and keyword relevance assessment task without using the ground truth labels. "Input" and "Output" refer to the prompt provided to the LLM and the generated explanations, respectively. The red outputs indicate incorrect generated explanations.

B   Zero-shot Prompts

Given a search engine query, first, consider what the user could have in mind when they type in the query and allow for misspellings or other ambiguity, then classify the relevance of keyword to the query into one of the following categories: "Not bad", or "Bad". The definitions of the categories are

"Not bad": the keyword is relevant to the user's search query. This can include: broader or narrower product selection, competitor or alternative products, accessories, products often purchased together and related topics as well as direct matches to the user's search.

"Bad": the keyword is not relevant to the user's search query. There is no relationship between the query and keyword. This includes but is not limited to: incorrect/unrelated product, unrelated topic, wrong location when location is important, cannot be used in place of query product nor are they commonly purchased together, etc.

Please predict whether the keyword is relevant to the query or not. The answer should be exact "Not bad" or "Bad".

Table 10: Zero-shot prompt for the QK task.

The goal of this task is to determine whether the targeted word in the first sentence and the second sentence conveys the same meaning. Please note that if the targeted word appears multiple times in the sentences, only the instance of the word surrounded by quotation marks should be considered. Additionally, the tense and form of the targeted word should not be taken into account. If the targeted word in the first sentence and the second sentence corresponds to the same meaning, output "True"; otherwise, output "False".

To complete this task, you will need to predict whether the targeted word "w" in the first sentence "s1" and the second sentence "s2" convey the same meaning. Your answer should be either "True" or "False".

Table 11: Zero-shot exemplars prompt for the WiC task.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer. If there exists evidence in the passage that supports the facts in the question, the answer should be "Yes". If there exists evidence in the passage that denies the facts in the question, the answer should be "No".

Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Table 12: Zero-shot prompt for the BoolQ task.

C Few-shot Prompts

Given a search engine query, first, consider what the user could have in mind when they type in the query and allow for misspellings or other ambiguity, then classify the relevance of keyword to the query into one of the following categories: "Not bad", or "Bad". The definitions of the categories are
"Not bad": the keyword is relevant to the user's search query. This can include: broader or narrower product selection, competitor or alternative products, accessories, products often purchased together and related topics as well as direct matches to the user's search.
"Bad": the keyword is not relevant to the user's search query. There is no relationship between the query and keyword. This includes but is not limited to: incorrect/unrelated product, unrelated topic, wrong location when location is important, cannot be used in place of query product nor are they commonly purchased together, etc.
Please predict whether the keyword is relevant to the query or not. The answer should be exact "Not bad" or "Bad".

Query: google data studio sharepoint
Keyword: sharepoint migration tool file share
Answer: Bad

Query: motorhomes sale
Keyword: rv sale used class c
Answer: Not bad

Query: southern exposure seed exchange company
Keyword: uk poppy seeds
Answer: Not bad

Query: nissan parts canada
Keyword: purchase tires
Answer: Bad

Query: alcohol detoxing
Keyword: inpatient drug rehab
Answer: Not bad

Query: loudmouth clothing sale
Keyword: levis jeans
Answer: Bad

Query: firefox mac sierra
Keyword: opera browser mac
Answer: Not bad

Query: google images
Keyword: buy photo
Answer: Bad

Table 13: Few-shot exemplars prompt for the QK task.

The goal of this task is to determine whether the targeted word in the first sentence and the second sentence conveys the same meaning. Please note that if the targeted word appears multiple times in the sentences, only the instance of the word surrounded by quotation marks should be considered. Additionally, the tense and form of the targeted word should not be taken into account. If the targeted word in the first sentence and the second sentence corresponds to the same meaning, output "True"; otherwise, output "False".
To complete this task, you will need to predict whether the targeted word "w" in the first sentence "s1" and the second sentence "s2" convey the same meaning. Your answer should be either "True" or "False".

w: "place"
s1: Do you want to come over to my "place" later?
s2: A political system with no "place" for the less prominent groups.
Answer: False

w: "hold"
s1: The general ordered the colonel to "hold" his position at all costs.
s2: "Hold" the taxi.
Answer: True

w: "summer"
s1: We like to "summer" in the Mediterranean.
s2: We "summered" in Kashmir.
Answer: True

w: "approach"
s1: "Approach" a task.
s2: To "approach" the city.
Answer: False

w: "run"
s1: "Run" rogue.
s2: She "ran" 10 miles that day.
Answer: False

w: "head"
s1: His horse won by a "head".
s2: He is two "heads" taller than his little sister.
Answer: True

w: "meet"
s1: The company agrees to "meet" the cost of any repairs.
s2: This proposal "meets" my requirements.
Answer: True

w: "development"
s1: The organism has reached a crucial stage in its "development".
s2: Our news team brings you the latest "developments".
Answer: False

Table 14: Few-shot exemplars prompt for the WiC task.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer. If there exists evidence in the passage that supports the facts in the question, the answer should be "Yes". If there exists evidence in the passage that denies the facts in the question, the answer should be "No". Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Question: is elder scrolls online the same as skyrim
Answer: No

Passage: Good Samaritan law – Good Samaritan laws offer legal protection to people who give reasonable assistance to those who are, or who they believe to be, injured, ill, in peril, or otherwise incapacitated. The protection is intended to reduce bystanders' hesitation to assist, for fear of being sued or prosecuted for unintentional injury or wrongful death. An example of such a law in common-law areas of Canada: a good Samaritan doctrine is a legal principle that prevents a rescuer who has voluntarily helped a victim in distress from being successfully sued for wrongdoing. Its purpose is to keep people from being reluctant to help a stranger in need for fear of legal repercussions should they make some mistake in treatment. By contrast, a duty to rescue law requires people to offer assistance and holds those who fail to do so liable.
Question: do good samaritan laws protect those who help at an accident
Answer: Yes

Passage: Windows Movie Maker – Windows Movie Maker (formerly known as Windows Live Movie Maker in Windows 7) is a discontinued video editing software by Microsoft. It is a part of Windows Essentials software suite and offers the ability to create and edit videos as well as to publish them on OneDrive, Facebook, Vimeo, YouTube, and Flickr.
Question: is windows movie maker part of windows essentials
Answer: Yes

Passage: Epsom railway station – Epsom railway station serves the town of Epsom in Surrey. It is located off Waterloo Road and is less than two minutes' walk from the High Street. It is not in the London Oyster card zone unlike Epsom Downs or Tattenham Corner stations. The station building was replaced in 2012/2013 with a new building with apartments above the station (see end of article).
Question: can you use oyster card at epsom station
Answer: No

Passage: Da Vinci's Demons – The series premiered in the United States on Starz on 12 April 2013, and its second season premiered on 22 March 2014. The series was renewed for a third season, which premiered on 24 October 2015. On 23 July 2015, Starz announced that the third season would be the show's last. However Goyer has left it open for a miniseries return.
Question: will there be a season 4 of da vinci's demons
Answer: No

Passage: Powdered sugar – Powdered sugar, also called confectioners' sugar, icing sugar, and icing cake, is a finely ground sugar produced by milling granulated sugar into a powdered state. It usually contains a small amount of anti-caking agent to prevent clumping and improve flow. Although most often produced in a factory, powdered sugar can also be made by processing ordinary granulated sugar in a coffee grinder, or by crushing it by hand in a mortar and pestle.
Question: is confectionary sugar the same as powdered sugar
Answer: Yes

Table 15: Continued on the next page

Passage: Federal judiciary of the United States – The federal courts are composed of three levels of courts. The Supreme Court of the United States is the court of last resort. It is generally an appellate court that operates under discretionary review, which means that the Court can choose which cases to hear, by granting writs of certiorari. There is therefore generally no basic right of appeal that extends automatically all the way to the Supreme Court. In a few situations (like lawsuits between state governments or some cases between the federal government and a state) it sits as a court of original jurisdiction.
Question: is the federal court the same as the supreme court
Answer: No

Passage: Bixby letter – In the 1998 war film Saving Private Ryan, General George Marshall (played by Harve Presnell) reads the Bixby letter to his officers before giving the order to find and send home Private James Francis Ryan after Ryan's three brothers died in battle.
Question: did abraham lincoln write the letter in saving private ryan
Answer: Yes

Table 15: Few-shot exemplars prompt for the BoolQ task.

D   Few-shot CoT Prompts

Given a search engine query, first, consider what the user could have in mind when they type in the query and allow for misspellings or other ambiguity, then classify the relevance of keyword to the query into one of the following categories: "Not bad", or "Bad". The definitions of the categories are

"Not bad": the keyword is relevant to the user's search query. This can include: broader or narrower product selection, competitor or alternative products, accessories, products often purchased together and related topics as well as direct matches to the user's search.

"Bad": the keyword is not relevant to the user's search query. There is no relationship between the query and keyword. This includes but is not limited to: incorrect/unrelated product, unrelated topic, wrong location when location is important, cannot be used in place of query product nor are they commonly purchased together, etc.

Please predict whether the keyword is relevant to the query or not. The answer should be exact "Not bad" or "Bad".

Query: google data studio sharepoint
Keyword: sharepoint migration tool file share
Answer: **The relevance of the keyword "sharepoint migration tool file share" to the search query "google data studio sharepoint" is "Bad". This keyword is not relevant to the user's search query as it pertains to a different tool or technology (SharePoint) that is not mentioned in the search query. The keyword focuses on a migration tool for SharePoint file sharing, while the search query is focused on Google Data Studio. There is no relationship between the two, so the keyword is not relevant to the user's search query.** *Therefore, the relevance is "Bad".*

Query: motorhomes sale
Keyword: rv sale used class c
Answer: **The relevance of keyword "rv sale used class c" is "Not bad". The keyword "rv sale used class c" contains relevant information to the user's search query "motorhomes sale". It includes a specific type of motorhome, "class c", that is being sold used, which is a relevant product selection for the user's search. Additionally, the keyword "rv sale used class c" is a direct match to the user's search and is commonly purchased together with the products in the search query.** *Therefore, the relevance is "Not bad".*

Query: southern exposure seed exchange company
Keyword: uk poppy seeds
Answer: **The relevance of the keyword "uk poppy seeds" to the search query "southern exposure seed exchange company" is "Not bad". This is because the search query is about a seed exchange company, and the keyword "uk poppy seeds" is a specific type of seed that could be offered by such a company. Even though the company mentioned in the search query is based in the southern hemisphere, they may still offer seeds from other regions, including the UK, which could include poppy seeds.** *Therefore, the relevance is "Not bad".*

Query: nissan parts canada
Keyword: purchase tires
Answer: **The relevance of "purchase tires" to the query "nissan parts canada" is "Bad". The keyword "purchase tires" is not directly related to the user's search for Nissan parts in Canada. Tires are not a part of a Nissan car and therefore are not relevant to the search for Nissan parts. Additionally, the keyword does not relate to a common purchase with Nissan parts, and it is not a substitute for the search query.** *Therefore, the relevance is "Bad".*

Table 16: Few-shot exemplars for full chain-of-thought prompt for the QK task. The bold text indicates the explanations generated with the prompt in Table 1.

Task description: The task is to identify if the occurrences of w in s1 and s2 correspond to the same meaning or not. The target word w in s1 and s2 is enclosed in quotation marks. There is no need to consider the tense and form of w. If w in s1 and s2 corresponds to the same meaning, output "true", otherwise output "false".

w: "place" s1: Do you want to come over to my "place" later?
s2: A political system with no "place" for the less prominent groups.
Explanation: **The answer is "false". In the first sentence, "place" refers to a physical location or a person's home. In the second sentence, "place" refers to a position or status in a political system. Therefore, although the spelling and punctuation of "place" are the same in both sentences, the context and meaning of the word are different.** Therefore, the answer is "false".

w: "hold" s1: The general ordered the colonel to "hold" his position at all costs.
s2: "Hold" the taxi.
Explanation: **The answer is "true". In both sentences, "hold" means to maintain or stay in a specific position or location. In the first sentence, the general orders the colonel to hold his position at all costs, meaning the colonel should not retreat or move from his current location. In the second sentence, someone is asking the taxi driver to hold, meaning to wait and stay in the current location until the person returns. Therefore, in both sentences, "hold" is used to convey the idea of staying in a particular position or location.** Therefore, the answer is "true".

w: "summer"
s1: We like to "summer" in the Mediterranean.
s2: We "summered" in Kashmir.
Explanation: **The answer is "true". In both sentences, "summer" refers to the activity of spending a summer vacation in a certain location. Although the word form is different ("summer" in the first sentence and "summered" in the second), the meaning is the same. The use of quotation marks in both sentences implies that "summer" is being used as a verb, rather than a season. Therefore, both sentences convey the idea of enjoying a summer vacation in different locations, making the meaning the same.** Therefore, the answer is "true".

w: "approach"
s1: "Approach" a task.
s2: To "approach" the city.
Explanation: **The answer is "false". In the first sentence, "Approach" is used as a verb and means to begin or undertake a task. In the second sentence, "approach" is also used as a verb but means to move towards or get closer to the city. Although the spelling and the quotation marks around the word are the same in both sentences, the context and the meaning of the word are different. Therefore, the occurrences of "approach" in the two sentences do not correspond to the same meaning.** Therefore, the answer is "false".

w: "cover"
s1: "Cover" her face with a handkerchief.
s2: Count the cash in the drawer twice just to "cover" yourself.
Explanation: **The answer is "false". The word "cover" has different meanings in the two sentences. In the first sentence, "cover" means to hide or conceal something. In contrast, in the second sentence, "cover" means to protect oneself from potential problems or accusations. The context and usage of the word "cover" in the two sentences are different, indicating that the two instances of "cover" do not correspond to the same meaning.** Therefore, the answer is "false".

w: "head"
s1: His horse won by a "head".
s2: He is two "heads" taller than his little sister.
Explanation: **The answer is "true". In both sentences, "head" is used as a unit of measurement. In the first sentence, it means that the speaker's horse won by a small margin, specifically the length of the horse's head. In the second sentence, it means that the person is taller than their little sister by a specific amount, namely two head lengths. Therefore, in both cases, "head" is used as a measure of distance or height, indicating that the two occurrences correspond to the same meaning.** Therefore, the answer is "true".

w: "meet"
s1: The company agrees to "meet" the cost of any repairs.
s2: This proposal "meets" my requirements.
Explanation: **The answer is "true". Although the verb "meet" is used in different senses in the two sentences, in both cases, it conveys the idea of satisfying a requirement or obligation. In the first sentence, the company agrees to pay for any repairs, which is a requirement or obligation. In the second sentence, the proposal fulfills or satisfies the speaker's requirements. Therefore, even though the two sentences use "meet" in different senses, they convey the same general idea of satisfying an obligation or requirement, making the answer "true".** Therefore, the answer is "true".

Table 17: Continued on the next page

w: "development"
s1: The organism has reached a crucial stage in its "development".
s2: Our news team brings you the latest "developments".
Explanation: **The answer is "false". In the first sentence, "development" refers to the growth or progress of an organism. In contrast, in the second sentence, "developments" refer to the latest events or updates on a particular subject. The first sentence refers to a specific stage in the life of an organism, while the second sentence refers to recent events or news related to a particular topic. Therefore, the context and meaning of "development" in both sentences are different, and they cannot be considered the same.** Therefore, the answer is "false".

Table 17: Few-shot exemplars for full chain-of-thought prompt for the WiC task. The bold text indicates the explanations generated with the prompt in Table 7.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer. If there exists evidence in the passage that supports the facts in the question, the answer should be "Yes". If there exists evidence in the passage that denies the facts in the question, the answer should be "No". Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Question: is elder scrolls online the same as skyrim
Answer: **The answer is "No". Although The Elder Scrolls Online is set on the continent of Tamriel, like Skyrim and other games in The Elder Scrolls series, it takes place a millennium before Skyrim and around 800 years before Morrowind and Oblivion. While it has a similar structure to Skyrim, with two conflicts progressing at the same time, the primary struggle in The Elder Scrolls Online is against the Daedric Prince Molag Bal, who has stolen the player character's soul, and the second is to capture the vacant imperial throne. Therefore, The Elder Scrolls Online is not the same game as Skyrim.** Therefore, the answer is "No".

Passage: Good Samaritan law – Good Samaritan laws offer legal protection to people who give reasonable assistance to those who are, or who they believe to be, injured, ill, in peril, or otherwise incapacitated. The protection is intended to reduce bystanders' hesitation to assist, for fear of being sued or prosecuted for unintentional injury or wrongful death. An example of such a law in common-law areas of Canada: a good Samaritan doctrine is a legal principle that prevents a rescuer who has voluntarily helped a victim in distress from being successfully sued for wrongdoing. Its purpose is to keep people from being reluctant to help a stranger in need for fear of legal repercussions should they make some mistake in treatment. By contrast, a duty to rescue law requires people to offer assistance and holds those who fail to do so liable.
Question: do good samaritan laws protect those who help at an accident
Answer: **The answer is "Yes". Good Samaritan laws offer legal protection to people who provide reasonable assistance to those who are injured, ill, in peril, or otherwise incapacitated. The purpose of these laws is to encourage bystanders to provide help without fear of being sued or prosecuted for unintentional injury or wrongful death. In Canada, for example, the good Samaritan doctrine is a legal principle that prevents a rescuer who has voluntarily helped a victim in distress from being successfully sued for wrongdoing. Therefore, individuals who provide reasonable assistance at an accident can be protected by good Samaritan laws.** Therefore, the answer is "Yes".

Passage: Windows Movie Maker – Windows Movie Maker (formerly known as Windows Live Movie Maker in Windows 7) is a discontinued video editing software by Microsoft. It is a part of Windows Essentials software suite and offers the ability to create and edit videos as well as to publish them on OneDrive, Facebook, Vimeo, YouTube, and Flickr.
Question: is windows movie maker part of windows essentials
Answer: **The answer to the question "is Windows Movie Maker part of Windows Essentials?" is Yes. As stated in the passage, Windows Movie Maker is a discontinued video editing software by Microsoft and was a part of the Windows Essentials software suite. Windows Essentials was a collection of free software programs from Microsoft, including Photo Gallery, Movie Maker, Mail, Writer, and OneDrive, among others. Therefore, Windows Movie Maker was one of the programs included in the Windows Essentials suite, which was available for download from Microsoft's website.** Therefore, the answer is "Yes".

Table 18: Continued on the next page

Passage: Epsom railway station – Epsom railway station serves the town of Epsom in Surrey. It is located off Waterloo Road and is less than two minutes' walk from the High Street. It is not in the London Oyster card zone unlike Epsom Downs or Tattenham Corner stations. The station building was replaced in 2012/2013 with a new building with apartments above the station (see end of article).
Question: can you use oyster card at epsom station
Answer: **The answer is "No". The passage explicitly states that Epsom railway station is not in the London Oyster card zone, unlike Epsom Downs or Tattenham Corner stations. Therefore, you cannot use Oyster card at Epsom railway station.** Therefore, the answer is "No".

Passage: Da Vinci's Demons – The series premiered in the United States on Starz on 12 April 2013, and its second season premiered on 22 March 2014. The series was renewed for a third season, which premiered on 24 October 2015. On 23 July 2015, Starz announced that the third season would be the show's last. However Goyer has left it open for a miniseries return.
Question: will there be a season 4 of da vinci's demons
Answer: **The answer is "No". The passage states that the third season of "Da Vinci's Demons" was the show's last and that the series creator, David S. Goyer, left open the possibility of a miniseries return. However, there is no indication that a fourth season of the show was planned or produced. Therefore, based on the information provided, it can be concluded that there will not be a season 4 of "Da Vinci's Demons".** Therefore, the answer is "No".

Passage: Powdered sugar – Powdered sugar, also called confectioners' sugar, icing sugar, and icing cake, is a finely ground sugar produced by milling granulated sugar into a powdered state. It usually contains a small amount of anti-caking agent to prevent clumping and improve flow. Although most often produced in a factory, powdered sugar can also be made by processing ordinary granulated sugar in a coffee grinder, or by crushing it by hand in a mortar and pestle.
Question: is confectionary sugar the same as powdered sugar
Answer: **The answer is "Yes". Confectioners' sugar is another name for powdered sugar, as mentioned in the passage. It is produced by milling granulated sugar into a fine powder, and usually contains a small amount of anti-caking agent to prevent clumping. Both terms, confectioners' sugar and powdered sugar, can be used interchangeably to refer to the same product.** Therefore, the answer is "Yes".

Passage: Federal judiciary of the United States – The federal courts are composed of three levels of courts. The Supreme Court of the United States is the court of last resort. It is generally an appellate court that operates under discretionary review, which means that the Court can choose which cases to hear, by granting writs of certiorari. There is therefore generally no basic right of appeal that extends automatically all the way to the Supreme Court. In a few situations (like lawsuits between state governments or some cases between the federal government and a state) it sits as a court of original jurisdiction.
Question: is the federal court the same as the supreme court
Answer: **The answer is "No". The federal court is not the same as the Supreme Court. The federal courts consist of three levels of courts, while the Supreme Court is the highest level court of the federal judiciary. The Supreme Court operates under discretionary review and generally only hears cases on appeal. It also has original jurisdiction in certain specific situations, but it is not the same as the federal court.** Therefore, the answer is "No".

Passage: Batman & Robin (film) – With the box office success of Batman Forever in June 1995, Warner Bros. immediately commissioned a sequel. They hired director Joel Schumacher and writer Akiva Goldsman to reprise their duties the following August, and decided it was best to fast track production for a June 1997 target release date, which is a break from the usual 3-year gap between films. Schumacher wanted to homage both the broad camp style of the 1960s television series and the work of Dick Sprang. The storyline of Batman & Robin was conceived by Schumacher and Goldsman during pre-production on A Time to Kill. Portions of Mr. Freeze's back-story were based on the Batman: The Animated Series episode "Heart of Ice", written by Paul Dini.
Question: is batman and robin a sequel to batman forever
Answer: **The answer is "Yes". Based on the information provided in the passage, Batman & Robin is a sequel to Batman Forever. Warner Bros. commissioned the sequel after the box office success of Batman Forever, and hired director Joel Schumacher and writer Akiva Goldsman to continue their duties. The film was fast-tracked for a June 1997 release, with Schumacher and Goldsman conceiving the storyline during pre-production of another movie. Therefore, Batman & Robin is a direct sequel to Batman Forever, released just two years later.** Therefore, the answer is "Yes".

Table 18: Few-shot exemplars for full chain-of-thought prompt for the BoolQ task. The bold text indicates the explanations generated with the prompt in Table 8.

E Prompts Used to Test the Stability

We present the few-shot prompts p1, p2 and p3 in Tables 19, 20 and 21, respectively. The few-shot prompt p3 is obtained by swapping the order of the "Question" and "Passage" in Table 15. While few-shot prompts p1 and p2 have minor variations in their task description compared to p3, we have highlighted the differences in bold.

We show the few-shot CoT prompts p1, p2 and p3 in Tables 22, 23 and 24, respectively. The few-shot CoT prompt p3 is obtained by swapping the order of the "Question" and "Passage" in Table 18. While few-shot CoT prompts p1 and p2 have minor variations in their task description compared to p3, we have highlighted the differences in bold.

Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Question: is elder scrolls online the same as skyrim
Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Answer: No

......

Table 19: Few-shot prompt p1 for the BoolQ task.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer.
Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Question: is elder scrolls online the same as skyrim
Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Answer: No

......

Table 20: Few-shot prompt p2 for the BoolQ task.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer. If there exists evidence in the passage that supports the facts in the question, the answer should be "Yes". If there exists evidence in the passage that denies the facts in the question, the answer should be "No".
Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Question: is elder scrolls online the same as skyrim
Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Answer: No

......

Table 21: Few-shot prompt p3 for the BoolQ task.

Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Question: is elder scrolls online the same as skyrim
Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Answer: The answer is "No". Although The Elder Scrolls Online is set on the continent of Tamriel, like Skyrim and other games in The Elder Scrolls series, it takes place a millennium before Skyrim and around 800 years before Morrowind and Oblivion. While it has a similar structure to Skyrim, with two conflicts progressing at the same time, the primary struggle in The Elder Scrolls Online is against the Daedric Prince Molag Bal, who has stolen the player character's soul, and the second is to capture the vacant imperial throne. Therefore, The Elder Scrolls Online is not the same game as Skyrim. Therefore, the answer is "No".

......

Table 22: Few-shot CoT prompt p1 for the BoolQ task.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer.
Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Question: is elder scrolls online the same as skyrim
Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Answer: The answer is "No". Although The Elder Scrolls Online is set on the continent of Tamriel, like Skyrim and other games in The Elder Scrolls series, it takes place a millennium before Skyrim and around 800 years before Morrowind and Oblivion. While it has a similar structure to Skyrim, with two conflicts progressing at the same time, the primary struggle in The Elder Scrolls Online is against the Daedric Prince Molag Bal, who has stolen the player character's soul, and the second is to capture the vacant imperial throne. Therefore, The Elder Scrolls Online is not the same game as Skyrim. Therefore, the answer is "No".

......

Table 23: Few-shot CoT prompt p2 for the BoolQ task.

Yes/No question-answering consists of a short passage and a Yes/No question about the passage. The questions are provided anonymously and unsolicited by users of the Google search engine, and afterwards paired with a paragraph from a Wikipedia article containing the answer. If there exists evidence in the passage that supports the facts in the question, the answer should be "Yes". If there exists evidence in the passage that denies the facts in the question, the answer should be "No".
Your task is to read the passage and predict whether the answer to the question is "Yes" or "No".

Question: is elder scrolls online the same as skyrim
Passage: The Elder Scrolls Online – As with other games in The Elder Scrolls series, the game is set on the continent of Tamriel. The events of the game occur a millennium before those of The Elder Scrolls V: Skyrim and around 800 years before The Elder Scrolls III: Morrowind and The Elder Scrolls IV: Oblivion. It has a broadly similar structure to Skyrim, with two separate conflicts progressing at the same time, one with the fate of the world in the balance, and one where the prize is supreme power on Tamriel. In The Elder Scrolls Online, the first struggle is against the Daedric Prince Molag Bal, who is attempting to meld the plane of Mundus with his realm of Coldharbour, and the second is to capture the vacant imperial throne, contested by three alliances of the mortal races. The player character has been sacrificed to Molag Bal, and Molag Bal has stolen their soul, the recovery of which is the primary game objective.
Answer: The answer is "No". Although The Elder Scrolls Online is set on the continent of Tamriel, like Skyrim and other games in The Elder Scrolls series, it takes place a millennium before Skyrim and around 800 years before Morrowind and Oblivion. While it has a similar structure to Skyrim, with two conflicts progressing at the same time, the primary struggle in The Elder Scrolls Online is against the Daedric Prince Molag Bal, who has stolen the player character's soul, and the second is to capture the vacant imperial throne. Therefore, The Elder Scrolls Online is not the same game as Skyrim. Therefore, the answer is "No".

......

Table 24: Few-shot CoT prompt p3 for the BoolQ task.