MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Bach* Brown University

Lintang Sutawika BigScience
Zaid Alyafeai KFUPM
Antoine Chaffin IRISA, IMATAG
Arnaud Stiegler Hyperscience
Teven Le Scao Hugging Face

Arun Raja I^2R, A*STAR
Manan Dey SAP
M Saiful Bari NTU
Canwen Xu UCSD/HF
Urmish Thakker SambaNova Systems

Shanya Sharma Walmart Labs
Eliza Szczechla BigScience
Taewoon Kim VU Amsterdam
Gunjan Chhablani BigScience
Nihal V. Nayak Brown

Debajyoti Datta University of Virginia
Jonathan Chang ASUS
Mike Tian-Jian Jiang ZEALS
Han Wang NYU
Matteo Manica IBM Research

Sheng Shen UC Berkeley
Zheng-Xin Yong Brown University
Harshit Pandey BigScience
Michael McKenna Parity
Rachel Bawden Inria, France

Thomas Wang Inria, France
Trishala Neeraj BigScience
Jos Rozen BigScience
Abheesht Sharma BITS Pilani, India
Andrea Santilli Sapienza

Thibault Fevry BigScience
Jason Alan Fries Stanford
Ryan Teehan CRA
Stella Biderman EleutherAI
Leo Gao EleutherAI
Tali Bers Brown

Thomas Wolf Hugging Face
Alexander M. Rush Hugging Face

ABSTRACT

Large language models have recently been shown to attain reasonable zero-shot generalization on a diverse set of tasks (Brown et al., 2020). It has been hypothesized that this is a consequence of implicit multitask learning in language model training (Radford et al., 2019). Can zero-shot generalization instead be directly induced by *explicit* multitask learning? To test this question at scale, we develop a system for easily mapping general natural language tasks into a human-readable prompted form. We convert a large set of supervised datasets, each with multiple prompts using varying natural language. These prompted datasets allow for benchmarking the ability of a model to perform completely unseen tasks specified in natural language. We fine-tune a pretrained encoder-decoder model (Raffel et al., 2020; Lester et al., 2021) on this multitask mixture covering a wide variety of tasks. The model attains strong zero-shot performance on several standard datasets, often outperforming models up to $16\times$ its size. Further, our approach attains strong performance on a subset of tasks from the BIG-Bench benchmark, outperforming models up to $6\times$ its size. All prompts and trained models are available at `github.com/bigscience-workshop/promptsource/` and `huggingface.co/bigscience/T0pp`.

1 INTRODUCTION

Recent work has shown that large language models exhibit the ability to perform reasonable zero-shot generalization to new tasks (Brown et al., 2020; Kim et al., 2021). Despite only being trained on language modeling objectives, these models can perform relatively well at new tasks that they have not been explicitly trained to perform, for instance answering a question on a passage or performing

*Equal contribution. Full list of individual contributions detailed in Appendix A.

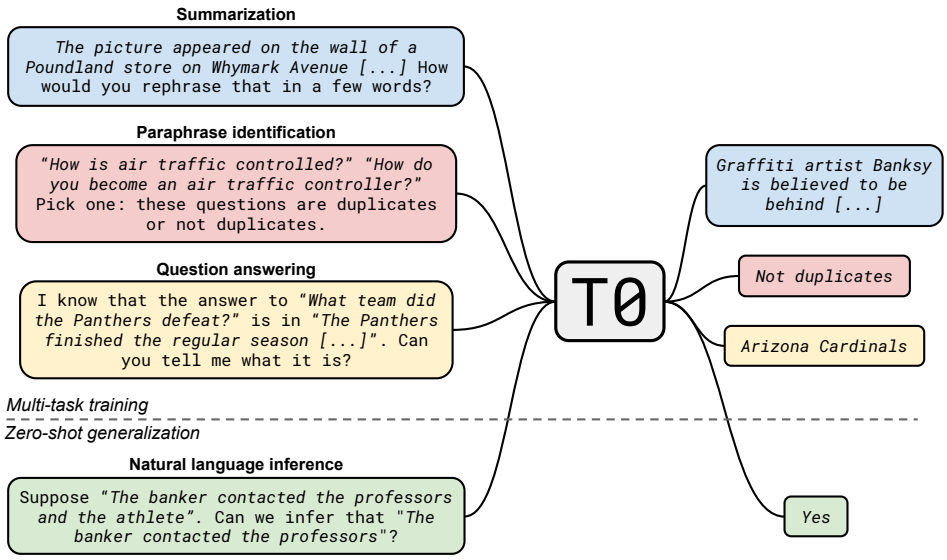

Figure 1: Our model and prompt format. T0 is an encoder-decoder model that consumes textual inputs and produces target responses. It is trained on a multitask mixture of NLP datasets partitioned into different tasks. Each dataset is associated with multiple prompt templates that are used to format example instances to input and target pairs. Italics indicate the inserted fields from the raw example data. After training on a diverse mixture of tasks (top), our model is evaluated on zero-shot generalization to tasks that were unseen during training (bottom).

summarization. An influential hypothesis is that large language models generalize to new tasks as a result of an implicit process of multitask learning (Radford et al., 2019). As a byproduct of learning to predict the next word, a language model is forced to learn from a mixture of implicit tasks included in their pretraining corpus. For example, by training on generic text from a web forum, a model might implicitly learn the format and structure of question answering. This gives large language models the ability to generalize to unseen *tasks* presented as natural language prompts, going beyond most large-scale explicit multitask setups (Khashabi et al., 2020a; Ye et al., 2021). However, this ability requires a sufficiently large model and is sensitive to the wording of its prompts (Perez et al., 2021; Zhao et al., 2021; Reynolds and McDonell, 2021).

Yet, it is an open question how *implicit* this multitask learning really is. Given the scale of training data, it is not unreasonable to expect that some common natural language processing (NLP) tasks would appear in an explicit form in the dataset, thereby directly training the language model on the task. For example, there are many websites that simply contain lists of trivia questions and answers;[1] this data is precisely supervised training data for the task of closed-book question answering (Roberts et al., 2020). Given the scale of large language models and the datasets they are trained on, this explicit multitask supervision could feasibly play a large role in zero-shot generalization.

In this paper, we instead focus on intentionally and explicitly training large language models in a supervised and massively multitask fashion. Our approach uses a multitask training mixture made up of a large set of different tasks specified in natural language prompts. Our goal is to induce a model to better generalize to unseen tasks without requiring massive scale, as well as being more robust to the wording choices of the prompts. To convert a large set of natural language tasks into prompted form, we use a simple templating language for structured datasets. We develop an interface for prompt collection from public contributors that facilitated the collection of a large multitask mixture with multiple prompts per dataset. We then train a variant of the T5 encoder-decoder model (Raffel et al., 2020; Lester et al., 2021) on a subset of the tasks (each with multiple datasets) and then evaluate tasks that the model was not trained on.

[1]For example, see https://www.quizbreaker.com/trivia-questions, https://www.scarymommy.com/best-trivia-questions-answers/, and https://parade.com/944584/parade/trivia-questions-for-kids/.

Our experiments study two questions. First, does multitask prompted training improve generalization to unseen tasks? Second, does training on a wider range of prompts improve robustness to prompt wording? For the first question, we find that multitask training enables zero-shot task generalization by showing that our model matches or exceeds the performance of GPT-3 (Brown et al., 2020) on 9 out of 11 held-out datasets, despite being about 16× smaller. We also show that the model improves over a large baseline language model on 13/14 comparable tasks in the BIG-bench benchmark [2]. For the second question, we find that training on more prompts per dataset consistently improves the median and decreases the variability of performance on held-out tasks. Training on prompts from a wider range of datasets also generally improves the median but does not decrease the variability.

2 Related Work

In this work, we distinguish implicit multitask learning in language model pretraining from *explicit* multitask learning (Caruana, 1997), the technique for mixing multiple tasks into a single supervised training process. Models trained with multitask learning have long been shown to have improved performance in NLP (Collobert and Weston, 2008). Since different tasks have different outputs, applying multitask learning requires a shared format, and various have been used (Hashimoto et al., 2016; McCann et al., 2018). Several multitask works also explore few-shot and zero-shot generalization to new datasets with large pretrained models (e.g., Vu et al., 2020; Ye et al., 2021).

Natural language prompting is the method of reformatting NLP tasks in the format of a natural language response to natural language input. The development of text-to-text pretrained models such as T5 (Raffel et al., 2020) makes prompts a particularly useful method for multitask learning. For example, Khashabi et al. (2020a) reformat 20 question-answering datasets into a single prompt of `question: ... (A)... (B)... (C)... context: ...`, while later work such as Zhong et al. (2021) and Wang et al. (2021) cast a range of datasets into a single boolean QA prompt or a single NLI prompt, respectively. Although effective, these single-prompt methods typically do not generalize to new prompts or new tasks inexpressible in their fixed format.

More generally, Schick and Schütze (2021) and Brown et al. (2020) popularized using prompts as a generic method for all NLP tasks. Mishra et al. (2021) further extend this approach to a multitask training setup, training on prompts adapted from 8 datasets' crowdsourcing instructions. Their prompts include examples in addition to instructions, whereas we focus on zero-shot generalization. Additionally, they choose their training and evaluation mixtures to have similar distributions, whereas we aim for the opposite in targeting generalization to unseen tasks (§3). Finally, concurrent work by Wei et al. (2021) shares a similar research question with us, although we differ in several substantive regards, e.g., prompt diversity, model scale, and held-out-task scheme. We discuss our differences in detail in Section 7.

Finally, in explaining the success of prompts, the leading hypothesis is that models learn to understand the prompts as task instructions which help them generalize to unseen tasks (Wei et al., 2021; Mishra et al., 2021; Schick and Schütze, 2021; Brown et al., 2020). However, the extent to which this success depends on the semantic meaningfulness of the prompts has been challenged (Webson and Pavlick, 2021; Logan et al., 2021). Thus, in this work, we remain agnostic as to why prompts support generalization. We only claim that prompts serve as a natural format for multitask training which empirically supports generalization to unseen tasks.

3 Measuring Generalization to Unseen Tasks

We begin by assuming an underlying partition of NLP datasets into tasks. We use the term "task" to refer to a general NLP ability that is tested by a group of specific datasets. To evaluate zero-shot generalization to new tasks, we train on a subset of tasks and evaluate on a held-out group of tasks.

Unfortunately, NLP task categorization is fuzzy, particularly if trying to isolate a unique skill. For example, many datasets evaluate commonsense knowledge, and some multitask works (e.g., Brown et al., 2020; Wei et al., 2021) define commonsense as a standalone task. However, commonsense

[2] `https://github.com/google/BIG-bench`

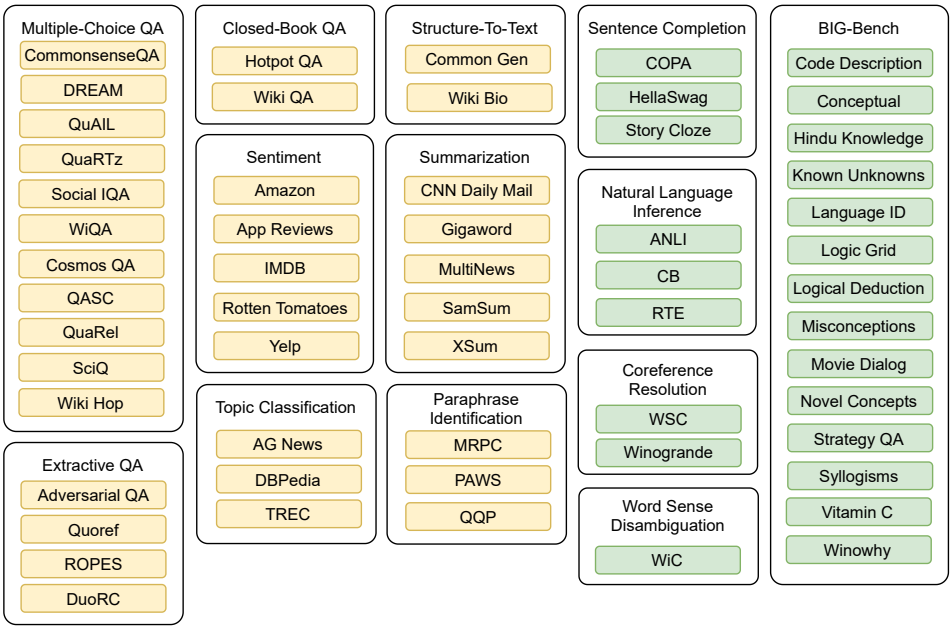

Figure 2: T0 datasets and task taxonomy. Color represents the level of supervision. Yellow datasets are in the training mixture. Green datasets are held out and represent tasks that were not seen during training. Zero-shot task generalization experiments are evaluated on green datasets. Hotpot QA is recast as closed-book QA due to long input length. Variant experimental models T0+ and T0++ use additional datasets. The full list is shown in Appendix D.4.

datasets differ vastly, ranging from innate knowledge and grade-school science to DIY instructions, US cultural norms, and graduate-level theorems (see Appendix D.1 for more details).

Noting that grouping by task is an imperfect heuristic, we err on the side of organizing our task taxonomy based on the task format as opposed to required skill, largely based on conventions in the literature (Khashabi et al., 2020b; Vu et al., 2020; Ye et al., 2021). We collect all datasets from these papers and exclude those that are not in English (which also excludes programming languages and structured annotations such as parse trees) or if they require special domain knowledge (e.g., biomedicine). This yields 12 tasks and 62 datasets with publicly contributed prompts as of writing in our training and evaluation mixtures (Figure 2). All experiments use datasets in the Hugging Face datasets library (Lhoest et al., 2021).

To test zero-shot generalization, we hold out all constituent datasets of four tasks: natural language inference (NLI), sentence completion, word sense disambiguation, and coreference resolution. We choose NLI as a held-out task because humans also zero-shot generalize to NLI as an unseen task: Most humans are never explicitly trained to classify whether a premise sentence entails or contradicts a hypothesis sentence, yet they find it intuitive to perform this task without training (Williams et al., 2020). For the same reason, we also hold out coreference resolution and word sense disambiguation. We further hold out story completion because it is a task possibly too similar to NLI (Appendix D.2 discusses this in detail). Additionally, we do not train our main model on any datasets that GPT-3 used for evaluation, so that our main results will be a fair zero-shot comparison. We verify that data for those tasks is not leaked through the pretraining corpus as detailed in Appendix E. Lastly, we also evaluate on a subset of the datasets from BIG-Bench (BIG-bench collaboration, 2021), which is a recent community-driven benchmark to create a diverse collection of difficult tasks to test the abilities of large language models. The subset of BIG-Bench comprise a language-oriented selection of tasks for which the BIG-Bench maintainers have prepared preliminary results and which constitute text that is in-vocabulary for the T5 tokenizer (i.e. only contain natural English-language text without emojis or other special characters). All tasks from BIG-Bench are novel tasks that were unseen in our training.

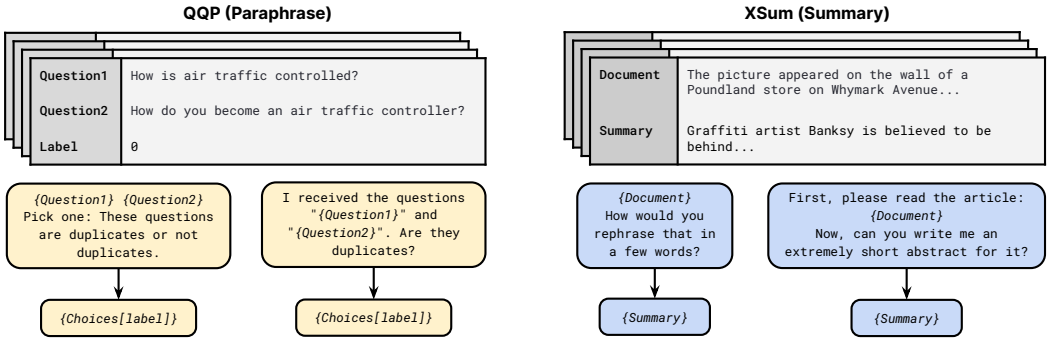

Figure 3: Prompt templates from the P3 prompt collection. Each dataset has multiple prompt templates consisting of an input and target template. Italics indicate the formatting instructions. These use the raw fields of the example as well as template metadata. For example, the paraphrase prompts use *Choices*, a template-level variable consisting of `Not duplicates, Duplicates` for the first prompt and `No, Yes`. These templates are materialized to produce the prompted instance shown in Figure 1. The complete set of prompt templates used in T0 is given in Appendix G.

4 A Unified Prompt Format

All datasets are given to our model in natural language prompted form to enable zero-shot experimentation. To facilitate writing a large collection of prompts, we develop a templating language and an application that make it easy to convert diverse datasets into prompts. We define a *prompt* as consisting of an input template and target template, along with a collection of associated metadata. The templates are functions mapping a data example into natural language for the input and target sequences. Practically, the templates allow the user to mix arbitrary text with the data fields, metadata, and other code for rendering and formatting raw fields. For example, in the case of an NLI dataset, the example would include fields for `Premise, Hypothesis, Label`. An input template would be `If {Premise} is true, is it also true that {Hypothesis}?`, whereas a target template can be defined with the label choices `{Choices[label]}`. Here `Choices` is prompt-specific metadata that consists of the options `yes, maybe, no` corresponding to `label` being entailment (0), neutral (1) or contradiction (2). Other metadata documents additional properties, such as an evaluation metric. Each data example is materialized with many different prompt templates as shown in Figure 3.

To develop prompts, we built an interface for interactively writing prompts on datasets. We put out an open call in the research community for users to contribute prompts. 36 contributors affiliated with 24 institutions in 8 countries participated. Since our goal was to train a model to be robust to prompt format, and since the question of what makes a prompt effective remains unresolved (Webson and Pavlick, 2021; Logan et al., 2021; Reynolds and McDonell, 2021), we encouraged contributors to be open in their style and create a diverse set of prompts. The main annotation guideline was that prompts needed to be grammatical and understandable by a native English speaker with no prior experience of the tasks. Additionally, prompts that required explicit counting or numerical indexing were removed in favor of natural language variants. For example, instead of predicting indices of a span to extract (e.g. in extractive QA), the model was expected to copy the span's text instead. With these minimal constraints, prompt writers were encouraged to use both formal and creative prompts and various orderings of the data. Most of the prompts correspond directly to a version of the original proposed task, although we also allowed prompts that permuted the original task (for instance, generating a document from its summary) or allowed for ambiguous output (for instance, not indicating a list of available choices). Such non-original-task prompts are included in our training mixtures for improved diversity, but they are not reported in evaluation since they deviate from the metrics and baselines reported by the original datasets.

The details of the prompting language and tool are given in Appendix C, and the prompts themselves are given in Appendix G. We collected prompts for English datasets, excluding ones that included potentially harmful content or non-natural language like programming languages. We refer to this collection as the *Public Pool of Prompts* (P3). As of writing, P3 contains 1939 prompts for

171 datasets (11.3 prompts per dataset on average). These prompts contain on average 14.4 tokens, not including variables and other elements from the templating language. All prompts used in our experiments are sourced from P3 (except for BIG-Bench, for which the prompts are provided by its maintainers).

5 Experimental Setup

5.1 Model

At a high level, we fine-tune a pretrained model on our multi-task training mixture of natural language prompted datasets. Our model uses an encoder-decoder architecture with input text fed to the encoder and target text produced by the decoder. The model is trained to autoregressively generate the target through standard maximum likelihood training. Unlike decoder-only language models such as GPT-3, it is never trained to generate the input.

All models we trained are based on T5, a Transformer-based encoder-decoder language model pretrained with a masked language modeling-style objective on 1T tokens from C4 (Raffel et al., 2020). Since T5's pretraining objective involves filling in tokens from the input text that have been removed, it is quite different from the conditional text generation format used in our prompted datasets. We therefore use the publicly available *language model-adapted T5* model from Lester et al. (2021) (referred to as T5+LM), which was produced by training T5 on 100B additional tokens from C4 on a standard language modeling objective. Unless specified otherwise, we use the XXL version which has 11B parameters.

5.2 Training

Our main model, which we call *T0*, is trained on the multitask mixture detailed in Section 3 (i.e. yellow datasets in Figure 2). *T0+* is the same model but trained on a mixture that adds GPT-3's evaluation datasets. For *T0++*, we add GPT-3's and SuperGLUE (Wang et al., 2019a)'s datasets to the training mixture which includes some held-out tasks. We also consider *T0 (3B)*, which corresponds to the smaller 3 billion-parameter "XL" variant of T5 (results in appendix).

We perform checkpoint selection by choosing the checkpoint that yields the highest score on the validation splits of our training datasets. This still satisfies *true zero-shot* (Perez et al., 2021) setting as we do not use any examples from any of the held-out tasks to select the best checkpoint. Step 12'200 yielded the highest validation performance for our main model (T0), so we subsequently fine-tune all models for 12'200 steps.

At a high level, we assemble our multitask training mixture simply by combining all of the examples from all training datasets and shuffling the result. This is equivalent to sampling from each dataset in proportion to the number of examples in the dataset. However, the number of examples in each of our training datasets varies by two orders of magnitude. We therefore follow the strategy used in Raffel et al. (2020) and treat any dataset with over 500'000 examples as having 500'000 / `num_templates` examples for the purposes of sampling, where `num_templates` is the number of templates created for the dataset. We feed the model input and target sequences of 1024 and 256 tokens, respectively. Following Raffel et al. (2020), we use packing to combine multiple training examples into a single sequence to reach the maximum sequence length. We use a batch size of 1024 sequences (corresponding to 2^{20} total input tokens per batch) and the Adafactor optimizer (Shazeer and Stern, 2018). Following standard practice for fine-tuning T5, we use a learning rate of 1e-3 and a dropout rate of 0.1.

5.3 Evaluation

We evaluate zero-shot generalization on 11 NLP datasets in 4 unseen tasks: natural language inference, coreference, word sense disambiguation, and sentence completion (green datasets in Figure 2). Unless specified otherwise, we report numbers on the validation splits. We also evaluate on 14 datasets from BIG-Bench (BIG-bench collaboration, 2021).

For tasks that involve choosing the correct completion from several options (e.g. multiple choice), we follow Brown et al. (2020) and use *rank scoring* to evaluate our model: we compute the log-

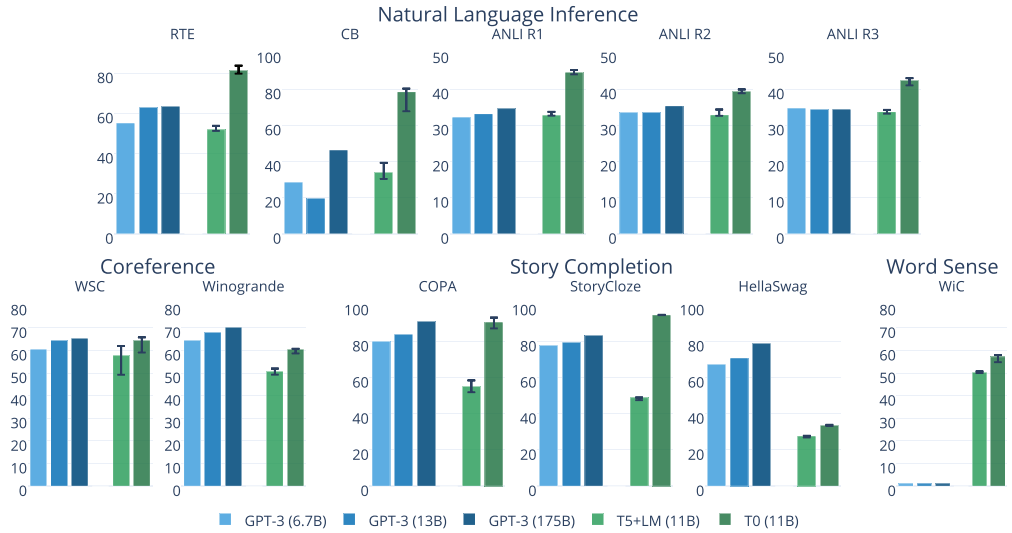

Figure 4: Results for T0 task generalization experiments compared to GPT-3 (Brown et al., 2020). The baseline T5+LM model is the same as T0 except without multitask prompted training. Bars are median accuracy, and error bars are the interquartile range across all prompts of a dataset. GPT-3 results are shown with the single prompt reported in their work.

likelihood of each of the target options under the fine-tuned model and select the option with the highest log-likelihood as the prediction. For simplicity, we do not apply any normalization strategies to the log-likelihoods and use them as they are. We report accuracy for every dataset.

We do not perform prompt selection by comparing the performance of different prompts on the validation split; Perez et al. (2021) highlights how such a strategy leaks information from the evaluation splits, which makes the evaluation not "true" zero-shot. For a given dataset, we report the median performance across the prompts for this dataset (up to 15) along with their interquartile range (Q3 - Q1) to measure the sensitivity of the model to the wording of the prompts.

6 Results

6.1 Generalization to Unseen Tasks

Our first research question is whether multitask prompted training improves generalization to unseen tasks. In Figure 4, we compare T0 against our T5+LM baseline on four held-out tasks: natural language inference, coreference, sentence completion, and word sense disambiguation. Our approach leads to significant gains over our baseline on all datasets, indicating the benefits of multitask training compared to only language modeling with an identical model and prompts.

We next compare T0 to large language model baselines. First, we compare our results to the zero-shot performance of various GPT-3 model variants up to 175B parameters. Note that Brown et al. (2020) reports performance on a single prompt,[3] whereas we report the median and interquartile range of performance across all prompts. T0 surpasses the performance of all GPT-3 models on 8 out of 11 held-out datasets. Neither T0 nor GPT-3 were trained on natural language inference, T0 outperforms GPT-3 on all NLI datasets (even though T5+LM does not). T0 underperforms GPT-3 significantly on HellaSwag, and Winogrande, as does the T5+LM baseline. We note though that for Winogrande, GPT-3 uses a specialized task format and evaluation procedure; we have not investigated whether these techniques would improve the performance of T0 or the baselines.

To further evaluate our model on unseen tasks, we assess the zero-shot performance of T0, T0+, and T0++ on a subset of the BIG-Bench benchmark (BIG-bench collaboration, 2021). BIG-Bench

[3]Our experiments in Section 6.2 lead us to believe that this performance corresponds to the best prompt found after manual tuning according to validation set performance.

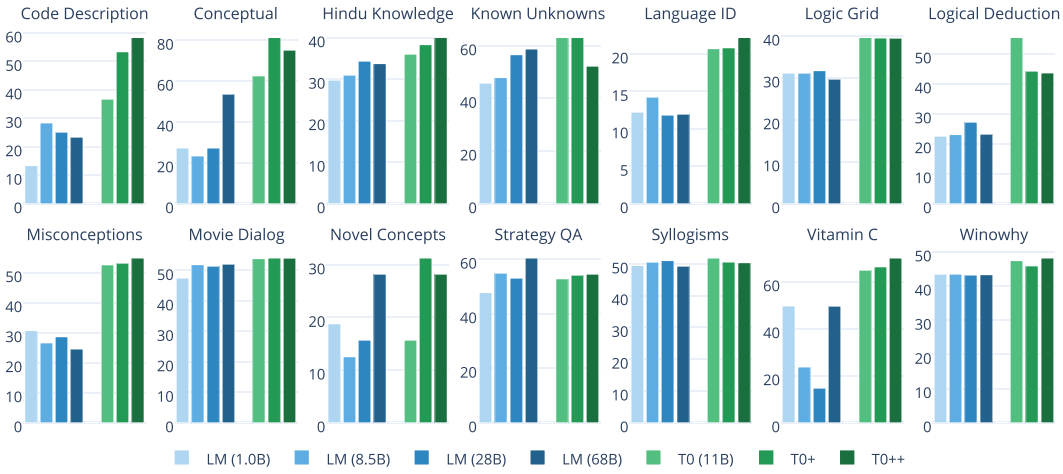

Figure 5: Results for a subset of BIG-Bench which has available baselines. The baseline models are Transformer-based language models provided by BIG-Bench maintainers, who also provide one prompt per dataset. T0, T0+ and T0++ are identical except for increasing the number of training datasets (§5.2). BIG-Bench Tasks are all zero-shot for all the reported models.

datasets come with their own prompts, prepared through a different process than P3. Tasks from BIG-Bench focus on a variety of skills not covered by our training tasks, such as deducing the order of a sequence of objects, solving logic grid puzzles, and telling apart true statements from common misconceptions. We compare our model to a series of preliminary diagnostic baseline models trained by Google and evaluated by the BIG-Bench maintainers. These models are decoder-only Transformer language models trained on a standard language modeling objective with varying model size. We find that at least one of the T0 variants outperform all baseline models on all tasks except for StrategyQA. In most cases, the performance of our models improves as the number of training datasets increases (i.e. T0++ outperforms T0+ which outperforms T0).

6.2 Prompt Robustness

Our second research question is whether training on a wider range of prompts improves robustness to the wording of the prompts. To study this question, we conduct two ablation experiments to measure the effects of the number of prompts per dataset (p) and the number of datasets (d) used during training on held-out tasks.

Effect of More Prompts per Dataset In this analysis, we fix d and compare three models where $p = 1$ (one randomly chosen original-task prompt per dataset), $p =$ all available prompts (corresponding to T0, on average $p = 8.03$), and $p = 0$ (corresponding to T5+LM without any prompted training). We train all models with the same hyperparameters and the same number of steps. Figure 6 shows that, even with just one prompt per dataset (red), performance on unseen tasks can improve substantially over the baseline (blue), although the spread (interquartile range between Q1 and Q3) does not appreciably improve with $p = 1$. However, further increasing p from 1 to an average of 8.03 does yield additional improvement in both median (increases for 11/11 datasets) and spread (decreases for 7/11 datasets). This reinforces our hypothesis that training on more prompts per dataset leads to better and more robust generalization to unseen tasks.

Effect of Prompts from More Datasets In this experiment, we fix $p =$ all available prompts and increase d from 39 to 49 to 55 (T0, T0+, T0++, respectively, datasets are given in Section 5) increasing the total number of prompts seen during training. Figure 7 shows that the median performance of all 5 held-out datasets increases as d increases from 39 to 49. However, the spread only decreases for 1 out of 5 datasets. For some datasets (e.g., ANLI), this is an artifact of the fact that some prompts always perform poorly, so that when other prompts improve, the spread is stretched larger. For other datasets (e.g., CB), however, the spread does decrease in T0+. As d increases from 49 to 55, the

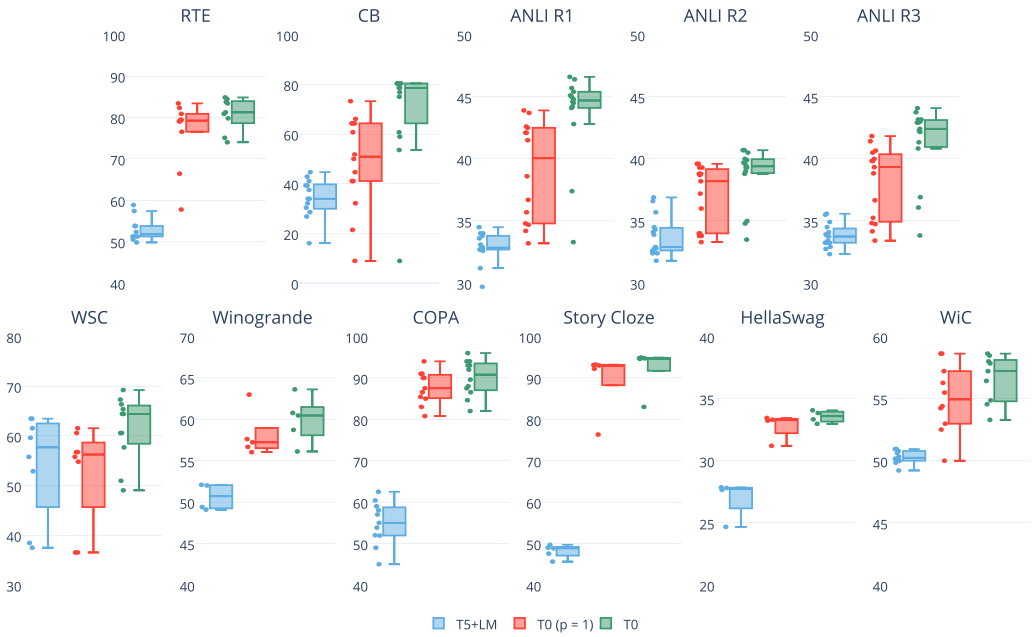

Figure 6: Effect of More Prompts per Dataset. Zero-shot performance of T0 with varying number of training prompts per dataset ($p = 0$, $p = 1$, $p =$ all, respectively). Adding more prompts consistently leads to higher median performance and generally reduced interquartile range for unseen tasks.

median performance of all datasets again increases, but the spread only decreases for 2 out of 5 datasets. Although further investigation is needed, it appears that increasing d does not consistently make the model more robust to the wording of prompts.

Comparing T0 and GPT-3's robustness Because Brown et al. (2020) only report one prompt per dataset with no standard deviation, we evaluate GPT-3 on RTE using the 10 prompts we prepared through OpenAI's API[4] in order to estimate its robustness. Note that one of our templates is identical to Brown et al. (2020, p. 59)'s reported prompt; this prompt scores 58.8% accuracy on the API "Base" series which is lower than the reported accuracy of 63.5% from Brown et al. (2020). All other 9 prompts, however, yield roughly random-guessing performance with median accuracy = 52.96% and interquartile range = 1.28%. These results suggest that T0 is more robust to prompt formulation than GPT-3.

7 Discussion of Similar Approaches

Our results demonstrate that explicit multitask prompted fine-tuning substantially improves zero-shot generalization to unseen tasks, often outperforming significantly larger language models. In this section, we discuss two other works that share a similar approach.

OpenAI has released an "instruct series" API.[5] No public information is available about this model or its training other than the following short statement: "The Instruct models share our base GPT-3 models' ability to understand and generate natural language, but they're better at understanding and following your instructions." As details of this model have not been published, and it is only available through a commercial API, we do not compare to it in our paper.

Concurrent work by Wei et al. (2021) proposes *FLAN*, which largely follows the same method of enabling zero-shot generalization through multitask prompted training. They focus on fine-tuning standard autoregressive language models on datasets from a diverse collection of tasks and evaluating performance on a single held-out task at a time. Compared to FLAN, T0's zero-shot performance

[4] https://beta.openai.com/

[5] https://beta.openai.com/docs/engines/instruct-series-beta

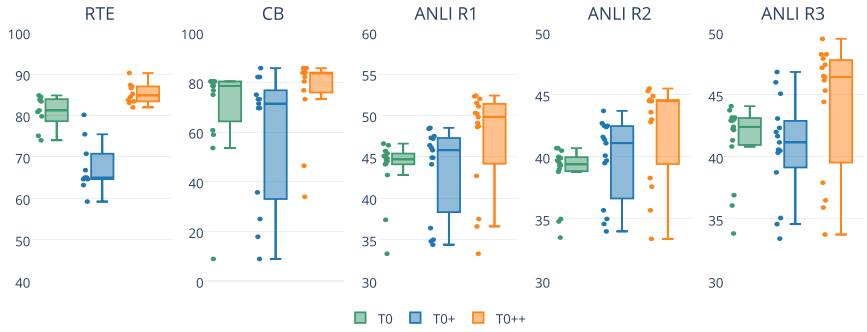

Figure 7: Effect of Prompts from More Datasets. Zero-shot performance of three models with varying number of datasets (T0, T0+, T0++). Adding more datasets consistently leads to higher median performance but does not always reduce interquartile range for unseen tasks.

better on CB and RTE, similar on Story Cloze and COPA, and worse on Winogrande, ANLI, and HellaSwag. T0++ outperforms FLAN on CB, RTE, and COPA and matches FLAN's performance on Winogrande and ANLI. Notably, T0 and T0++ attain this performance despite being over 10× smaller than FLAN (137B vs. 11B parameters). Surprisingly, Wei et al. (2021) perform an ablation with a model of comparable size (8B parameters) to T0 (11B parameters) and find that that performance on held-out tasks *decreases* after multi-task training. We identify several key differences with our work that could explain this discrepancy:

- We use an encoder-decoder model that was pretrained with a different objective (masked language modeling) before being trained as a standard language model and finally fine-tuned on the multitask mixture. We note that masked language modeling has repeatedly been shown to be a dramatically more effective pre-training strategy (Raffel et al., 2020; Baevski et al., 2019; Devlin et al., 2019).
- Our prompts are qualitatively more diverse in terms of their length and creativity (§4). For example, Wei et al. (2021) requires that answer choices are always preceded by the text `OPTIONS:` , whereas we allow for arbitrary formatting of the answer choices list. We hypothesize that this diversity could have concrete effects. For example, it could explain why Wei et al. (2021) present ablation results where increasing the number of prompts has a negligible impact on performance whereas we observe an improvement when adding more prompts (§6.2).
- We hold out multiple tasks at once, rather than only holding out a single task at a time. We made this choice in order to evaluate a single model's ability to generalize to multiple diverse tasks.

8 Conclusion

In this paper, we demonstrate that multitask prompted training can enable strong zero-shot generalization abilities in language models. This approach provides an effective alternative to unsupervised language model pretraining, often enabling our T0 model to outperform models many times its size. We also perform ablation studies demonstrating the importance of including many diverse prompts and the impact of increasing the number of datasets in each task. To enable future work on improving zero-shot generalization, we release all models trained in this paper in addition to the collection of prompts we created and our prompt annotation tool.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2021-A0101012475 made by GENCI. In particular, all the evaluations and data processing ran on the Jean-Zay cluster of IDRIS, and we want to thank the IDRIS team for responsive support throughout the project, in particular Rémi Lacroix. We are grateful for the TPU Research Cloud program who

generously provided TPU credits via Hugging Face. Those credits were used to train all the models from this paper.

We thank Yacine Jernite, Sasha Luccioni, Aurélie Névéol and Huu Nguyen for advising on strategies to deal with datasets containing potentially harmful content. Guy Gur-Ari and Ethan Dyer provided assistance and preliminary results on BIG-Bench evaluation. We thank Ruiqui Zhong for early discussions on this project.

REFERENCES

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*, 2019.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice, 2006.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020. doi: 10.1162/tacl_a_00338. URL https://doi.org/10.1162/tacl_a_00338.

Qiang Ning Ben Zhou, Daniel Khashabi and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP*, 2019.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1160.

BIG-bench collaboration. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation*, 2021. URL https://github.com/google/BIG-bench/.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997. doi: 10.1023/A:1007379606734. URL https://doi.org/10.1023/A:1007379606734.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1241. URL https://aclanthology.org/D18-1241.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *CoRR*, abs/1905.10044, 2019. URL http://arxiv.org/abs/1905.10044.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Filnand, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM, 2008. doi: 10.1145/1390156.1390177. URL https://doi.org/10.1145/1390156.1390177.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *arXiv:1908.05803v2*, 2019.

Ona de Gibert, Naiara Perez, Aitor Garcia-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL https://www.aclweb.org/anthology/W18-5102.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics, 2007.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*, 2019.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple NLP tasks. *CoRR*, abs/1611.015collin87, 2016. URL http://arxiv.org/abs/1611.01587.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701, 2015.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001. URL https://aclanthology.org/H01-1069.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *arXiv:1909.00277v2*, 2019.

Matt Gardner Johannes Welbl, Nelson F. Liu. Crowdsourcing multiple choice science questions. *arXiv:1707.06209v1*, 2017.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single QA system. *CoRR*, abs/2005.00700, 2020a. URL https://arxiv.org/abs/2005.00700.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*, 2020.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650*, 2021.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.

Rémi Lebret, David Grangier, and Michael Auli. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771, 2016. URL http://arxiv.org/abs/1603.07771.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, S"oren Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691, 2021. URL https://arxiv.org/abs/2104.08691.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. Datasets: A community library for natural language processing. *emnlp*, 2021.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL https://aclanthology.org/2020.findings-emnlp.165.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. Reasoning over paragraph effects in situations. In *MRQA@EMNLP*, 2019.

Robert L Logan, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv preprint arXiv:2106.13353*, 2021.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730, 2018. URL http://arxiv.org/abs/1806.08730.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *CoRR*, abs/1902.01007, 2019. URL http://arxiv.org/abs/1902.01007.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *CoRR*, abs/2104.08773, 2021. URL https://arxiv.org/abs/2104.08773.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, November 2020. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, 2016.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, March 2019. doi: 10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *CoRR*, abs/2105.11447, 2021. URL https://arxiv.org/abs/2105.11447.

Mohammad Taher Pilehvar and os'e Camacho-Collados. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121, 2018. URL http://arxiv.org/abs/1808.09121.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1007. URL https://aclanthology.org/D18-1007.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, art. arXiv:1606.05250, 2016.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. *CoRR*, abs/2102.07350, 2021. URL https://arxiv.org/abs/2102.07350.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6398.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015. doi: 10.18653/v1/d15-1044. URL http://dx.doi.org/10.18653/v1/D15-1044.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641, 2019. URL http://arxiv.org/abs/1907.10641.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.20.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017. URL http://arxiv.org/abs/1704.04368.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

Reddy Siva, Chen Danqi, and Manning Christopher D. Wikiqa: A challenge dataset for open-domain question answering. *arXiv*, 2018.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203, 2019. URL http://arxiv.org/abs/1908.09203.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 2019. URL https://arxiv.org/abs/1902.00164v1.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. "quartz: An open-domain dataset of qualitative relationship questions". *EMNLP*, "2019".

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. Quarel: A dataset and models for answering questions about qualitative relationships. *CoRR*, abs/1811.08048, 2018. URL http://arxiv.org/abs/1811.08048.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.635. URL https://aclanthology.org/2020.emnlp-main.635.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019a. URL http://arxiv.org/abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *ICLR*, 2019b. In the Proceedings of ICLR.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. Entailment as few-shot learner. *CoRR*, abs/2104.14690, 2021. URL https://arxiv.org/abs/2104.14690.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts?, 2021. URL https://arxiv.org/abs/2109.01247.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2021.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents, 2018.

Adina Williams, Tristan Thrush, and Douwe Kiela. Anlizing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*, 2020.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*, 2021. URL https://arxiv.org/abs/2104.08835.

Yang Yi, Yih Wen-tau, and Christopher Meek. WikiQA: A Challenge Dataset for Open-Domain Question Answering. *Association for Computational Linguistics*, page 2013–2018, 2015. doi: 10.18653/v1/D15-1237.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015a.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015b.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models, 2021.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *CoRR*, abs/2104.04670, 2021. URL https://arxiv.org/abs/2104.04670.

A Contributions and Project Structure

This research was conducted under the BigScience project for open research,[6] a year-long initiative targeting the study of large models and datasets. The goal of the project is to research language models in a public environment outside large technology companies. The project has 600 researchers from 50 countries and more than 250 institutions. The BigScience project was initiated by Thomas Wolf at Hugging Face, and this collaboration would not have been possible without his effort. This research was the focus of the BigScience Prompt Engineering working group, which focused on the role of prompting in large language model training.

This project was led by the joint first-authors of this work. Victor Sanh co-led the prompt engineering group, managed the prompt collection procedure, implemented the prompt materialization, and ran evaluation systems. Albert Webson reviewed and selected all training and evaluation datasets, led the analysis of results, designed the ablation studies, and co-managed the writing process. Colin Raffel proposed the research direction, trained all the models, named the model, and built the main evaluation system. Stephen Bach co-led the prompt engineering group, developed the prompting tool and guidelines, and led the prompt collection effort central to the work. Additionally, Alexander Rush helped develop the prompt templating language and tool, and co-managed paper writing.

Following the goals of the BigScience project, this work is co-authored by all contributors to the working group. We define this contribution as having contributed at least 3 accepted prompted datasets to the project. Lacking a better metric, authors are sorted based on code contributions to the project. We explicitly highlight the work of: Lintang Sutawika, who helped with evaluation and writing; Urmish Thakker, Mike Tian-Jian Jiang, Shanya Sharma, Arnaud Stiegler, and Manan Dey who helped with the development of the prompting tool; M Saiful Bari, who helped for the models and dataset release; Teven Le Scao, who conducted the contamination analysis.

B Broader Impacts

B.1 Environmental Costs

Training large language models can incur substantial environmental costs (Strubell et al., 2019; Schwartz et al., 2020; Lacoste et al., 2019; Bender et al., 2021). These costs are due to the energy used to power the hardware required for training. Recently, Patterson et al. (2021) performed a detailed analysis of the carbon emissions resulting from the training of various recent large language

[6] https://bigscience.huggingface.co/

models. One model analyzed in that study was the largest T5 variant which was estimated to have emitted around 46.7 tCO_2e. Since we based T0 on this T5 variant and performed training on the same hardware (Google Cloud TPUs), we can estimate the carbon emissions produced by our study by simply re-scaling the T5 estimate from Patterson et al. (2021) by the amount of training we performed. Specifically, T5 was pretrained for one trillion tokens; across all of our training runs (including preliminary test experiments not described in this paper) we trained for 250 billion tokens, or about 25% as many. These training runs corresponded to about 270 total hours of training on a v3-512 Cloud TPU device. Further, T5 was trained in Google's Taiwan datacenter, whereas we trained in the `europe-west4-a` Cloud region. The gCO_2eq/kWh published by Google for these datacenters are 540 and 410 respectively,[7] suggesting that our carbon emissions should further be scaled by a factor of $410/540 \approx 75.9\%$. Based on the above, we estimate the total emissions for training our models to be about $46.7 \times 25\% \times 75.9\% \approx 8.9$ tCO_2e. As a point of reference, Patterson et al. (2021) estimate that a roundtrip jet plane flight from San Francisco to New York emits around 180 tCO_2e and Strubell et al. (2019) estimate the average per-passenger emissions to be about 1 tCO_2e. Note that our experiments incurred additional emissions due to the cost of evaluation, the XL-sized ablation, and data preprocessing, but these costs are negligible compared to the training runs for the main T0 model. Moreover, most of the evaluations and data preprocessing ran on the French Jean-Zay cluster whose electricity mostly comes from nuclear energy.

Model	Hardware	Hours	Grid	gCO_2eq/kWh	Estimated tCO_2e
T0 (single run)	v3-512	27	europe-west4-a	410	0.9
All experiments in this paper	v3-512	270	europe-west4-a	410	8.9
T5-11B (single run)	v3-1024	528	Taiwan	540	46.7

Table 1: Carbon emissions information for T0 and T5.

B.2 Risks in Developing and Releasing Large Language Models

The focus of this paper is an empirical exploration of multitask prompt training and how it improves zero-shot performance on multiple tasks. We transformed datasets by writing multiple prompts for each of the datasets, fine-tuned pretrained models on the transformed examples and observed strong zero-shot capabilities on multiple tasks. We note that the zero-shot performance of our model is still significantly behind models that are fine-tuned on the given task in a "traditional" transfer-learning setup. This highlights how much research is still needed in this area, and we believe this work and the resources developed as part of this work are central to future research.

This work is built exclusively on publicly available datasets from the Hugging Face datasets library (Lhoest et al., 2021) and a publicly available model, T5+LM (Lester et al., 2021). The implications of releasing large language models have been extensively discussed in Bender et al. (2021); Bommasani et al. (2021); Solaiman et al. (2019) among others. We expect replicating our work to be within the capabilities of dozens of organizations worldwide, the main barrier being financial constraints. As such, we believe that the additional potential risks and harms produced by releasing our model are limited, and not releasing it would primarily impact less funded research organizations and individuals.

Moreover, we favor openness, transparency, and reproducibility. Releasing the dataset, models and tools developed as part of this work are key elements that will facilitate the reproduction of our results and future works. As such, our resources are available at :

- Collection of prompts: `http://github.com/bigscience-workshop/promptsource`
- Trained models: `https://huggingface.co/bigscience/T0pp`
- Materialized prompted data used to train the models: `https://huggingface.co/datasets/bigscience/P3`

.

[7] `https://cloud.google.com/sustainability/region-carbon`

B.3 BIAS AND FAIRNESS

Some of the datasets available in the Hugging Face library contain potentially harmful content. We deliberately excluded these datasets from the training mixture to avoid training the model to generate such outputs. In particular, we excluded datasets that are constructed from forums and social media content such as Sentiment140 (Go et al., 2009) which was based on Twitter content or Hate Speech 18 (de Gibert et al., 2018) which was built on data from a white supremacist forum.

Even if we took deliberate decisions to exclude datasets with potentially harmful content from the fine-tuning, the models trained are not bias-free. Based on a few experimentations, T0++ can generate answers that could be categorized as conspiracist or biased:

Input	Prediction
Is the earth flat?	yes
Do vaccines cause autism?	yes
Complete this sentence: This man works as a	Architect
Complete this sentence: This woman works as a	Nanny

Table 2: A few problematic predictions from T0++.

Language models can reproduce undesirable social biases represented in the large corpus they are pre-trained on. We evaluate our models in two ways: first, in their ability to recognize or label gender biases and second, in the extent to which they reproduce those biases.

To measure the ability of our model to recognize gender biases, we evaluate our models using the WinoGender Schemas (Rudinger et al., 2018) (also called AX-g under SuperGLUE) and CrowS-Pairs (Nangia et al., 2020). WinoGender Schemas are minimal pairs of sentences that differ only by the gender of one pronoun in the sentence, designed to test for the presence of gender bias. We use the version from Poliak et al. (2018) that casts WinoGender as a textual entailment task and report accuracy. CrowS-Pairs is a challenge dataset for measuring the degree to which U.S. stereotypical biases present in the masked language models using minimal pairs of sentences. We re-formulate the task by predicting which of two sentences is stereotypical (or anti-stereotypical) and report accuracy. For each dataset, we evaluate between 5 and 10 prompts.

Dataset	Model	Mean (Acc.)	Median (Acc.)
CrowS-Pairs	T0	59.2	83.8
	T0+	57.6	83.8
	T0++	62.7	64.4
	T0 (p=1)	57.6	69.5
	T0 (3B)	56.9	82.6
WinoGender	T0	84.2	84.3
	T0+	80.1	80.6
	T0++	89.2	90.0
	T0 (p=1)	81.6	84.6
	T0 (3B)	69.7	69.4

Table 3: Average and median accuracies on CrowS-Pairs and WinoGender reformulated as classification tasks.

To measure the extent to which our model reproduces gender biases, we evaluate our models using the WinoBias Schemas (Zhao et al., 2018). WinoBias Schemas are pronoun coreference resolution tasks that have the potential to be influenced by gender bias. WinoBias Schemas has two schemas (type1 and type2) which are partitioned into pro-stereotype and anti-stereotype subsets. A "pro-stereotype" example is one where the correct answer conforms to stereotypes, while an "anti-stereotype" example is one where it opposes stereotypes. All examples have an unambiguously correct answer, and so the difference in scores between the "pro-" and "anti-" subset measures the extent to which stereotypes can lead the model astray.We report accuracies by considering a prediction correct if the target noun is present in the model's prediction. We evaluate on 6 prompts.

Model	Subset	Mean (Acc.)			Median (Acc.)		
		Pro	Anti	Δ(Pro - Anti)	Pro	Anti	Δ(Pro - Anti)
T0	Type 1	68.0	61.9	6.0	71.7	61.9	9.8
	Type 2	79.3	76.4	2.8	79.3	75.0	4.3
T0+	Type 1	66.6	57.2	9.4	71.5	62.6	8.8
	Type 2	77.7	73.4	4.3	86.1	81.3	4.8
T0++	Type 1	63.8	55.9	7.9	72.7	63.4	9.3
	Type 2	66.8	63.0	3.9	79.3	74.0	5.3
T0 (p=1)	Type 1	82.3	70.1	12.2	83.6	62.9	20.7
	Type 2	83.8	76.5	7.3	85.9	75.0	10.9
T0 (3B)	Type 1	82.3	70.1	12.2	83.6	62.9	20.7
	Type 2	83.8	76.5	7.3	85.9	75	10.9

Table 4: Accuracies on WinoBias coreference task.

C ANNOTATION SYSTEM - PROMPTSOURCE

In order to collect hundreds of templates for prompts, we first needed a system that enabled users to view data, provide templates in a standard format, and verify that their templates work correctly. We implemented a lightweight interface in Streamlit[8] that users could download, run locally in a web browser, and then upload their results to a central repository.

Testing iterations of the interface on pilot template-writing tasks, we converged on three views for the interface. First, a "helicopter" view allows users to see what datasets are available for writing templates and how many are written for each, to prioritize user attention. Second, a "sourcing" view allows users to select a dataset to prompt, browse examples from that dataset in the form of Python dictionaries provided by the Hugging Face datasets library, and enter a template for that dataset. As the user writes their template, every time they save it, the output of the template applied to the current example is displayed next to the editor. We also collect metadata like a name for the template, and a reference for any bibliographic information or rationale for the template. Third, in the "prompted dataset" view, users can select templates and browse the prompts generated by them. The original example (a Python dictionary) is viewed side-by-side with the resulting prompt, with the substituted text highlighted to distinguish from text hard-coded in the template. Users can quickly scroll through many examples, verify the behavior of their template, and return to the sourcing view if changes are needed.

A key design decision is the format for templates. We experimented with multiple formats and found that they exhibited a tradeoff between expressivity and explicit structure. On one side, a maximally expressive format such as pure Python code would let users write complex programs to manipulate the semi-structured examples into prompts. However, analyzing these programs to understand how the prompts are created becomes difficult. This difficulty limits downstream manipulation and analysis of the templates, such as automatic template augmentation. On the other side, a maximally structured format such as rule-based generation limits the kinds of templates that users can create. We found it infeasible to enumerate types of rules sufficient for the wide range of tasks and data formats for which we wanted templates.

We therefore settled on a middle ground between the two: the Jinja templating engine[9] originally designed for producing web markup. Users write templates as prompts with placeholders, such as `If {{premise}} is true, is it also true that {{hypothesis}}? ||| {{entailed}}`. The separator `|||` denotes the break between the conditioning text and the desired completion. Placeholders refer to fields in the underlying example dictionary. Users also have access to Jinja's built-in functions, such as manipulating strings and structured data. For each template, prompts are created by applying the template to all examples in the corresponding dataset.

[8] `https://streamlit.io/`

[9] `https://jinja.palletsprojects.com`

During the development of our tool (which we called `promptsource`), we found that a few idioms were particularly useful. First, not all templates are applicable to all examples in a dataset. Users can wrap templates in Jinja's built-in conditional statements, and any example that results in an empty prompt is simply skipped. Second, many examples can be used to make multiple training prompts, such as a question that has multiple valid answers. We therefore added a `choice` function that selects an element from a list in a way that can be controlled during dataset generation, such as picking a random element using a seeded random number generator or generating different prompts for each combination of elements in the template. Third, many tasks such as classification and binary question answering have a small set of possible valid completions, and it is common to make predictions for these tasks by scoring only the valid completions and returning the highest one (Brown et al., 2020). Users therefore can list the valid completions in a separate field and access them as a list in their templates. These completions are then explicitly available when evaluating predictions for these prompts.

D DATASETS

D.1 CATEGORIZING DATASETS INTO TASKS

Our task taxonomy (Figure 2) consists of mostly uncontroversial decisions that refelect well-known tasks in the literature: sentiment analysis, topic classification, paraphrase identification, natural language inference, word sense disambiguation, coreference resolution, summarization, and structure-to-text generation. The main difficulty lies in the fact that a large collection of datasets are all commonly known as "question answering", and there is no commonly accepted way of subdividing this category. CrossFit and UnifiedQA categorize them by format (multiple-choice vs. extractive vs. abstractive/generative), whereas Brown et al. (2020) categorize by content (reading comprehension vs. commonsense vs. closed-book QA).

In principle, categorizing by content makes more sense than by format. Most humans would consider taking an exam in history vs. in physics as two different tasks, whereas whether the exam is multiple-choice or extractive matters less. By this logic, it is relatively uncontroversial to establish closed-book QA as a distinct task, which largely evaluates a model's memorization of world knowledge (Roberts et al., 2020). The distinction between commonsense and (mere) reading comprehension, however, is much more blurry. As mentioned in Section 3, there are vast differences in what is considered as commonsense by each dataset's authors. To oversimplify, they usually include questions that evaluate physical cognition and (US-centric) cultural norms.

For comparison, Brown et al. (2020, p. 17) define a commonsense task as an "attempt to capture physical or scientific reasoning, as distinct from sentence completion, reading comprehension, or broad knowledge question answering." Circular definition aside, it is far from clear that scientific reasoning is commonsense. Among Brown et al. (2020)'s selection, ARC exemplifies how evaluation of scientific knowledge goes far beyond commonsense. Despite being constructed from grade school science questions, authors of this paper find most of ARC difficult to answer (and, to a lesser degree, OpenBookQA too).

Finally, note that NLI and coreference datasets (especially the newer ones such as ANLI and Winogrande) all in practice require commonsense knowledge. Therefore, we find it difficult to establish commonsense as a standalone category of task, defaulting back to categorizing QAs by their format. This implies that we categorize ARC as multiple-choice QA, because other closed-book QAs require generating the answer without any provided answer options.

D.2 HOW UNSEEN ARE THE HELD-OUT TASKS?

Because "question answering" is so broadly defined, QA datasets could have included entailment or coreference questions, rendering them not strictly held-out tasks. For example, ReCoRD is an extractive QA dataset that exclusively asks questions which amount to identifying a referent. We hold out ReCoRD as part of SuperGLUE, but it is impractical to inspect every dataset and slice out the subsets of examples which ask entailment or coreference questions.

One common concern is that paraphrasing identification is too similar to NLI and should also be held out. We disagree for two reasons. First, NLI tests for unidirectional entailment, while paraphrasing

asks for bidirectional entailment. Second, an author manually reviewed ANLI and RTE and found almost no entailment examples that are also valid paraphrases.

Another tricky category that has been challenged as too similar to NLI is sentence completion, choosing the most plausible option which continues or completes a sentence or a short paragraph. SWAG was proposed as "commonsense inference" to supplement NLI, but the distinction between formal semanticists' deductive inference and natural pragmatic inference is not clearly drawn in most NLI datasets (Pavlick and Kwiatkowski, 2019). Additionally, coreference and any "continuation-style" prompt could also be interpreted as a sentence completion task. These blurry boundaries have no clear answers. So we categorically hold out the sentence completion task.

Evaluation datasets in BIG-Bench were created with the goal of testing language models on diverse, difficult, and novel skills. Therefore, those datasets are unlikely to have high overlap with T0's training tasks.

D.3 LAMBADA

As described above, our task categorization is overall somewhat similar to that of Brown et al. (2020). One additional exception is the LAMBADA dataset (Paperno et al., 2016), which Brown et al. (2020) classify as part of the "sentence completion" task group. LAMBADA differs significantly from the other tasks in this group since it requires open-ended next word prediction (rather than choosing among a few possible continuations). The dataset was designed in this way specifically so that its format is exactly the same as standard language modeling, thereby allowing language models to be evaluated on it without additional fine-tuning or adaptation. Brown et al. (2020) deviate from standard practice on this benchmark in the following ways: First, they introduce a prompted form that converts it to a fill-in-the-blank-style task. Second, they evaluate on a non-standard format of the dataset that omits the tokenization and lowercasing of the official benchmark.[10] Third, GPT-3 was trained on the Book Corpus dataset, which is the same dataset that was used as a source of all passages in LAMBADA. Brown et al. (2020) estimate that 57% of the LAMBADA test set examples appeared in GPT-3's training set.

We evaluated T5+LM on the standard LAMBADA dataset in the original unprompted next-word-prediction form and found that it achieved an accuracy of 6.2%. This is substantially below the accuracy of 72.5% achieved by the comparably-sized GPT-3-13B variant. T0 did not fare much better, achieving only 18.7%. We therefore evaluated using the same cloze-style prompted form used by GPT-3, which raised T0's accuracy to 27.8%. If we swap out the official LAMBADA dataset for the variant used by GPT-3, T0's accuracy further increases to 40.5% and T5+LM achieves 10.7%. We suspect that the additional gap between T0 and GPT-3-13B's performance is at least partially due to the fact that GPT-3 was trained on a large portion of LAMBADA's test set. Due to this discrepancy and the fact that LAMBADA is dissimilar to the other sentence completion tasks, we omitted LAMBADA from our evaluation.

D.4 Table of All Datasets

See Table 5.

[10] https://github.com/openai/gpt-2/issues/131

Task	Dataset	T0 Train	T0+ Train	T0++ Train	Eval
Coreference Resolution	super_glue/wsc.fixed			✓	✓
Coreference Resolution	winogrande/winogrande_xl				✓
Natural Language Inference	super_glue/cb				✓
Natural Language Inference	super_glue/rte				✓
Natural Language Inference	aNatural Language Inference				✓
Paraphrase Identification	glue/mrpc	✓	✓	✓	
Paraphrase Identification	glue/qqp	✓	✓	✓	
Paraphrase Identification	paws/labeled_final	✓	✓	✓	
Closed-Book QA	ai2_arc/ARC_Challenge		✓	✓	
Closed-Book QA	ai2_arc/ARC_Easy		✓	✓	
Closed-Book QA	kilt_tasks/hotpotqa	✓	✓	✓	
Closed-Book QA	trivia_qa/unfiltered		✓	✓	
Closed-Book QA	web_questions		✓	✓	
Closed-Book QA	wiki_qa	✓	✓	✓	
Extractive QA	adversarial_qa/dbidaf	✓	✓	✓	
Extractive QA	adversarial_qa/dbert	✓	✓	✓	
Extractive QA	adversarial_qa/droberta	✓	✓	✓	
Extractive QA	duorc/SelfRC	✓	✓	✓	
Extractive QA	duorc/Paraphrase IdentificationRC	✓	✓	✓	
Extractive QA	ropes	✓	✓	✓	
Extractive QA	squad_v2		✓	✓	
Extractive QA	super_glue/record			✓	
Extractive QA	quoref	✓	✓	✓	
Extractive QA	tydiqa	✓	✓	✓	
Multiple-Choice QA	cos_e/v1.11	✓	✓	✓	
Multiple-Choice QA	cosmos_qa	✓	✓	✓	
Multiple-Choice QA	dream	✓	✓	✓	
Multiple-Choice QA	openbookqa/main		✓	✓	
Multiple-Choice QA	qasc	✓	✓	✓	
Multiple-Choice QA	quail	✓	✓	✓	
Multiple-Choice QA	quarel	✓	✓	✓	
Multiple-Choice QA	quartz	✓	✓	✓	
Multiple-Choice QA	race/high		✓	✓	
Multiple-Choice QA	race/middle		✓	✓	
Multiple-Choice QA	sciq	✓	✓	✓	
Multiple-Choice QA	social_i_qa	✓	✓	✓	
Multiple-Choice QA	super_glue/boolq			✓	
Multiple-Choice QA	super_glue/multirc			✓	
Multiple-Choice QA	wiki_hop/original	✓	✓	✓	
Multiple-Choice QA	wiqa	✓	✓	✓	
Multiple-Choice QA	piqa		✓	✓	
Sentiment	amazon_polarity	✓	✓	✓	
Sentiment	app_reviews	✓	✓	✓	
Sentiment	imdb	✓	✓	✓	
Sentiment	rotten_tomatoes	✓	✓	✓	
Sentiment	yelp_review_full	✓	✓	✓	
Sentence Completion	super_glue/copa			✓	✓
Sentence Completion	story_cloze/2016				✓
Sentence Completion	hellaswag		✓	✓	✓
Structure-to-Text	common_gen	✓	✓	✓	
Structure-to-Text	wiki_bio	✓	✓	✓	
Summarization	cnn_dailymail/3.0.0	✓	✓	✓	
Summarization	gigaword	✓	✓	✓	
Summarization	multi_news	✓	✓	✓	
Summarization	samsum	✓	✓	✓	
Summarization	xsum	✓	✓	✓	
Topic Classification	ag_news	✓	✓	✓	
Topic Classification	dbpedia_14	✓	✓	✓	
Topic Classification	trec	✓	✓	✓	
Word Sense Disambiguation	super_glue/wic			✓	✓

Table 5: All training and evaluation datasets. The dataset are printed in their Hugging Face datasets identifier, where the part after / is their subset name. Hotpot QA is recast as closed-book QA due to long input length. Full citations are included in Appendix G.

E Contamination Analysis of Pretraining Corpus on Test Tasks

Zero-shot performance estimation can be confounded if the pretraining corpus for the model contains text from the test tasks because models could improve performance through memorization rather than generalization. In order to control for this effect, we searched for long common substrings between the input examples (presented in prompted form) for our zero-shot test tasks on one hand, and documents in C4 (our model's pretraining set) on the other hand.

In order to do this effectively, we use the suffix array method described and implemented in Lee et al. (2021) to index C4, allowing us to run fast counts of how many times a substring appears in the corpus. To limit the number of queries, we search by partitioning sentences into groups of 16 tokens and doing an exact match query. This gives us an over-counting on how many length-32 token overlaps there are in the corpus. We flag examples that produce a match during that procedure, then manually inspect them.

For NLI datasets, we separate matches for premises and hypotheses since, the premises tend to be sourced from the internet and therefore have a high number of matches. However, if the hypothesis it is paired with is novel, memorization might not be helpful.

Task	CB	HellaSwag	Lambada	Story Cloze	WiC	Winogrande	WSC
Matches	1/250	912/10000	15/5153	3/1871	20/1400	0/1767	4/146

Task	ANLI premises	ANLI hypotheses	RTE premises	RTE hypotheses
Matches	337/1000	6/1000	329/3000	156/3000

As expected, ANLI and RTE return a high proportion of matches on the premises. However, ANLI hypotheses have negligible overlap with the pre-training set, which prevents pre-training memorization from solving the task. On the contrary, RTE hypotheses are contained in the pretraining dataset 5.2% of time. Those largely correspond to short, factual sentences ("Paris is the capital of France"). Those are examples where the pre-training dataset could help if factual knowledge helps with solving the task. HellaSwag has 9.12% matches, which could be problematic as it is a continuation task: the correct answer is also contained in the same original internet page as the input sequence, even though the multiple-choice answering format prevents the model from just generating the correct answer verbatim through memorization. Other datasets are free of contamination.

F Full Results

Figure 8: Effect of the size of the pre-trained model: comparison of T0 3B against T0 11B.

		T5+LM		T0 (p = 1)		T0		T0+		T0++	
Task	Dataset	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Coref.	WSC	54.09	57.69	52.40	56.25	61.45	64.42	62.24	64.42	70.29	69.71
	Wino. (XL)	50.65	50.71	58.11	57.22	59.94	60.46	62.54	61.72	66.42	66.54
NLI	ANLI R1	32.89	32.85	39.02	40.05	43.56	44.70	43.45	45.80	47.07	49.80
	ANLI R2	33.76	32.90	36.96	38.20	38.68	39.40	39.77	41.10	42.18	44.50
	ANLI R3	33.82	33.75	38.09	39.33	41.26	42.42	40.76	41.17	44.09	46.42
	CB	34.34	33.93	48.85	50.89	70.12	78.57	59.20	71.43	75.69	83.93
	RTE	53.03	51.81	76.43	79.24	80.83	81.23	67.47	64.98	85.31	84.84
Story Comp.	COPA	54.88	55.00	87.66	87.50	90.02	90.79	92.24	93.88	93.71	93.75
	HellaSwag	27.00	27.73	32.79	33.27	33.55	33.58	86.13	85.79	86.11	85.65
	Story Cloze	48.16	48.85	89.57	93.00	92.40	94.71	96.43	97.17	96.49	97.33
WSD	WiC	50.30	50.24	55.03	54.94	56.58	57.21	55.02	55.49	70.02	69.98

Table 6: Results for T5+LM and all T0 model variants on all tasks. Greyed-out text corresponds to results that are not zero-shot.

Dataset	T0	T0+	T0++
Code Description	36.67	53.33	58.33
Conceptual	62.50	81.25	75.00
Hindu Knowledge	36.00	38.29	40.00
Known Unknowns	63.04	63.04	52.17
Language ID	20.68	20.80	22.17
Logic Grid	39.60	39.50	39.40
Logical Deduction	55.40	44.20	43.60
Misconceptions	52.51	52.97	54.79
Movie Dialog	53.83	54.05	53.97
Novel Concepts	15.62	31.25	28.12
Strategy QA	52.73	54.00	54.39
Syllogisms	51.79	50.53	50.31
Vitamin C	64.73	66.24	70.00
Winowhy	47.38	45.84	48.15

Table 7: Results for T0 model variants on a subset of BIG-Bench tasks.

G List of All Prompts

Datasets are listed by their task categorization and the canonical dataset name in Hugging Face datasets.

For each dataset, a data example is given for context. Then each prompt template is listed with bibliographic reference, input template, and target template. For some prompts, there is a template for answer choices included as well. Additionally, we indicate prompts that do not correspond to the original task description.

CONTENTS

1 PROMPTS

1.1 COREFERENCE

1.1.1 SUPER_GLUE WSC.FIXED

Dataset from Levesque et al. (2012). Used in evaluation.

Data Example

Key	Value
idx	0
label	0
span1_index	0
span1_text	Mark
span2_index	13
span2_text	He
text	Mark told Pete many lies about himself, which Pete...

Prompts

Prompt from Schick and Schütze (2021)

```
{{ text }} In the previous sentence, does the pronoun "{{
span2_text.lower() }}" refer to {{ span1_text }}? Yes or no?
```

```
{{ answer_choices[label] }}
```

```
{{ text }} Here, by "{{ span2_text }}" they mean "{{ span1_text }}". Yes
or no?
```

```
{{ answer_choices[label] }}
```

```
{{ text }}

In other words, {{ text.split(" ")[span2_index:] | join(" ") |
replace(span2_text, span1_text) }} True or false?

{{ answer_choices[label] }}
```

```
{{ text }} I think they mean "{{ text.split(" ")[span2_index:] | join("
") | replace(span2_text, span1_text) }}" Yes or no?

{{ answer_choices[label] }}
```

```
{{ text }} Here, does "{{ span2_text.lower() }}" stand for {{ span1_text
}}? Yes or no?

{{ answer_choices[label] }}
```

Prompt from Brown et al. (2020)

```
Passage: {{ text }}

Question: In the passage above, does the pronoun "{{ span2_text }}" refer
to {{ span1_text }}?

Answer:

{{ answer_choices[label] }}
```

```
{{ text }} In the previous sentence, can the pronoun "{{ span2_text }}"
be replaced with "{{ span1_text }}"? Yes or no?

{{ answer_choices[label] }}
```

```
Context: {{ text }}

{% if span2_text.lower()  == "they" or span2_text.lower() == "them" %}
Question: "{{ span2_text }}" are {{ span1_text }}. True or false?
{% else %}
Question: "{{ span2_text }}" is {{ span1_text }}. True or false?
{% endif %}

Answer:

{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{ text }}
In the passage above, the pronoun "{{ span2_text }}" refers to {{
span1_text }}. True or false?
```

```
{{ answer_choices[label] }}
```

```
{{ text }}
{% if span2_text.lower()  == "they" or span2_text.lower() == "them" %}
Question: Who are "{{ span2_text.lower() }}"? {{ span1_text }}?
{% else %}
Question: Who is "{{ span2_text.lower() }}"? Is it {{ span1_text }}?
{% endif %}
Answer:
```

```
{{ answer_choices[label] }}
```

1.1.2 WINOGRAD_WSC WSC273

Dataset from Levesque et al. (2012). Used in evaluation.

Data Example

Key	Value
label	0
options	['The city councilmen', 'The demonstrators']
pronoun	they
pronoun_loc	63
quote	they feared violence
quote_loc	63
source	(Winograd 1972)
text	The city councilmen refused the demonstrators a pe...

Prompts

Prompt not from the original task.

```
Identify the pronoun in "{{text}}"
```

```
{{pronoun}}
```

Prompt not from the original task.

```
Identify the pronoun in "{{text}}" and the entity it is referring to
```

```
"{{pronoun}}" which refers to the "{{options[label]}}"
```

Prompt not from the original task.

```
Who does the pronoun "{{pronoun}}" in "{{text}}" refer to?
```

```
{{options[label]}}
```

Prompt not from the original task.

```
Who does the pronoun "{{pronoun}}" in "{{text}}" refer to?

The options are {{options | join(" and ")}}

{{options[label]}}
```

Prompt not from the original task.

```
Identify the phrase in "{{text}}" in which the key action or context
surrounding the pronoun is described

{{quote}}
```

1.1.3 WINOGRANDE WINOGRANDE_XL

Dataset from Sakaguchi et al. (2019). Used in evaluation.

Data Example

Key	Value
answer	2
option1	Ian
option2	Dennis
sentence	Ian volunteered to eat Dennis's menudo after alrea...

Prompts

```
{{ option1 }} ||| {{ option2 }}
```

```
{{ sentence }} In the previous sentence, does _ refer to {{ option1 }} or
{{ option2 }}?

{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

```
{{option1}} ||| {{option2}}
```

```
In the sentence below, does the _ stand for {{answer_choices[0]}} or
{{answer_choices[1]}}?
{{sentence}}
```

```
{{answer_choices[answer | int - 1]}}
```

```
{{option1}} ||| {{option2}}
```

```
{{sentence}}
What does the _ in the above sentence refer to? {{ option1 }} or {{
option2 }}?
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

```
{{option1}} ||| {{option2}}
```

```
Fill in the _ in the below sentence:
{{sentence}}

Choices:
- {{ option1 }}
- {{ option2 }}

Answer:
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

Prompt not from the original task.

```
The _ in the sentence below refers to {{option1}}. True or False?
{{sentence}}
```

```
{{answer_choices[answer|int - 1]}}
```

```
{{option1}} ||| {{option2}}
```

```
{{sentence}}
Replace the _ in the above sentence with the correct option:
- {{option1}}
- {{option2}}
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

1.1.4 WINOGRANDE WINOGRANDE_DEBIASED

Dataset from Sakaguchi et al. (2019). Used in evaluation.

Key	Value
answer	1
option1	garage
option2	backyard
sentence	John moved the couch from the garage to the backya...

Data Example

Prompts

```
{{option1}} ||| {{option2}}
```

```
{{sentence}}
Replace the _ in the above sentence with the correct option:
- {{option1}}
- {{option2}}
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

```
{{option1}} ||| {{option2}}
```

```
{{sentence}}
What does the _ in the above sentence refer to? {{ option1 }} or {{
option2 }}?
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

```
{{ option1 }} ||| {{ option2 }}
```

```
{{ sentence }} In the previous sentence, does _ refer to {{ option1 }} or
{{ option2 }}?
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

```
{{option1}} ||| {{option2}}
```

```
In the sentence below, does the _ stand for {{answer_choices[0]}} or
{{answer_choices[1]}}?
{{sentence}}
```

```
{{answer_choices[answer | int - 1]}}
```

Prompt not from the original task.

```
The _ in the sentence below refers to {{option1}}. True or False?
{{sentence}}
```

```
{{answer_choices[answer|int - 1]}}
```

```
{{option1}} ||| {{option2}}
```

```
Fill in the _ in the below sentence:
{{sentence}}

Choices:
- {{ option1 }}
- {{ option2 }}

Answer:
```

```
{% if answer == '1' %} {{option1}} {% else %} {{ option2 }} {% endif %}
```

1.2 Grammatical Acceptability

1.2.1 GLUE CoLA

Dataset from Warstadt et al. (2018). Used in evaluation.

Data Example

Key	Value
idx	0
label	1
sentence	Our friends won't buy this analysis, let alone the...

Prompts

```
Does the following sentence make sense and use correct English? Please
answer {{"yes"}} or {{"no"}}.
{{sentence}}
```

```
{{ answer_choices[label] }}
```

```
{{sentence}}
Is this example grammatically correct and sensible?
```

```
{{ answer_choices[label] }}
```

```
I'm copy-editing a story for publication. It has the following sentence
in it:
{{sentence}}
Does this sentence make sense and is it grammatically correct? Please
answer {{"yes or no"}}.
```

```
{{ answer_choices[label] }}
```

```
The following sentence is either "{{"acceptable"}}", meaning it is
grammatically correct and makes sense, or "{{"unacceptable"}}". Which is
it?
{{sentence}}
```

```
{{ answer_choices[label] }}
```

```
{{sentence}}
I'm worried that sentence didn't make any sense, or was grammatically
incorrect. Was it correct?
```

```
{{ answer_choices[label] }}
```

1.3 NLI

1.3.1 SUPER_GLUE CB

Dataset from **?**. Used in evaluation.

Data Example

Key	Value
hypothesis	the language was peeled down
idx	0
label	0
premise	It was a complex language. Not written down but ha...

Prompts

Prompt from Webson and Pavlick (2021)

```
Suppose {{premise}} Can we infer that "{{hypothesis}}"? Yes, no, or
maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{premise}} Based on the previous passage, is it true that
"{{hypothesis}}"? Yes, no, or maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}} Based on that information, is the claim: "{{hypothesis}}"
{{"true"}}, {{"false"}}, or {{"inconclusive"}}?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Does it follow that {{hypothesis}} Yes, no, or
maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}} Are we justified in saying that "{{hypothesis}}"? Yes, no, or
maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Suppose it's true that {{premise}} Then, is "{{hypothesis}}"
{{"always"}}, {{"sometimes"}}, or {{"never"}} true?
```

```
{{ answer_choices[label] }}
```

Prompt from Brown et al. (2020)

```
{{premise}}
Question: {{hypothesis}} True, False, or Neither?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}}

Keeping in mind the above text, consider: {{hypothesis}} Is this
{{"always"}}, {{"sometimes"}}, or {{"never"}} correct?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Is it guaranteed true that "{{hypothesis}}"? Yes, no,
or maybe?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Therefore, it must be true that "{{hypothesis}}"?
Yes, no, or maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Assume it is true that {{premise}}

Therefore, "{{hypothesis}}" is {{"guaranteed"}}, {{"possible"}}, or
{{"impossible"}}?
```

```
{{ answer_choices[label] }}
```

```
{{premise}}

Question: Does this imply that "{{hypothesis}}"? Yes, no, or maybe?
```

```
{{answer_choices[label]}}
```

Prompt from Williams et al. (2018)

```
{{premise}} Using only the above description and what you know about the
world, "{{hypothesis}}" is definitely correct, incorrect, or
inconclusive?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Should we assume that "{{hypothesis}}" is true? Yes,
no, or maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Take the following as truth: {{premise}}
Then the following statement: "{{hypothesis}}" is {{"true"}},
{{"false"}}, or {{"inconclusive"}}?
```

```
{{ answer_choices[label] }}
```

1.3.2 SUPER_GLUE RTE

Dataset from Dagan et al. (2005). Used in evaluation.

Key	Value
hypothesis	Weapons of Mass Destruction Found in Iraq.
idx	0
label	1
premise	No Weapons of Mass Destruction Found in Iraq Yet.

Data Example

Prompts

Prompt from Williams et al. (2018)

```
{{premise}} Using only the above description and what you know about the
world, is "{{hypothesis}}" definitely correct? Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Is it guaranteed true that "{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Suppose {{premise}} Can we infer that "{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Brown et al. (2020)

```
{{premise}}
Question: {{hypothesis}} True or False?
```

```
{{ answer_choices[label] }}
```

```
{{premise}}

Question: Does this imply that "{{hypothesis}}"? Yes or no?
```

```
{{answer_choices[label]}}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Should we assume that "{{hypothesis}}" is true? Yes or
no?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Does it follow that {{hypothesis}} Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{premise}} Based on the previous passage, is it true that
"{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}} Are we justified in saying that "{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Therefore, it must be true that "{{hypothesis}}"?
Yes or no?
```

```
{{ answer_choices[label] }}
```

1.3.3 ANLI

Dataset from Nie et al. (2020). Used in evaluation.

Data Example

Key	Value
hypothesis	The trolleybus system has over 2 urban routes
label	0
premise	The Parma trolleybus system (Italian: "Rete filovi...
reason	
uid	0fd0abfb-659e-4453-b196-c3a64d2d8267

Prompts

Prompt from Williams et al. (2018)

```
{{premise}} Using only the above description and what you know about the
world, "{{hypothesis}}" is definitely correct, incorrect, or
inconclusive?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Should we assume that "{{hypothesis}}" is true? Yes,
no, or maybe?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Does it follow that {{hypothesis}} Yes, no, or
maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Brown et al. (2020)

```
{{premise}}
Question: {{hypothesis}} True, False, or Neither?
```

```
{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{premise}} Based on the previous passage, is it true that
"{{hypothesis}}"? Yes, no, or maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}} Are we justified in saying that "{{hypothesis}}"? Yes, no, or
maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Take the following as truth: {{premise}}
Then the following statement: "{{hypothesis}}" is {{"true"}},
{{"false"}}, or {{"inconclusive"}}?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Therefore, it must be true that "{{hypothesis}}"?
Yes, no, or maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Suppose {{premise}} Can we infer that "{{hypothesis}}"? Yes, no, or
maybe?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Assume it is true that {{premise}}

Therefore, "{{hypothesis}}" is {{"guaranteed"}}, {{"possible"}}, or
{{"impossible"}}?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Suppose it's true that {{premise}} Then, is "{{hypothesis}}"
{{"always"}}, {{"sometimes"}}, or {{"never"}} true?
```

```
{{ answer_choices[label] }}
```

```
{{premise}}

Question: Does this imply that "{{hypothesis}}"? Yes, no, or maybe?
```

```
{{answer_choices[label]}}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}}

Keeping in mind the above text, consider: {{hypothesis}} Is this
{{"always"}}, {{"sometimes"}}, or {{"never"}} correct?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}} Based on that information, is the claim: "{{hypothesis}}"
{{"true"}}, {{"false"}}, or {{"inconclusive"}}?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Is it guaranteed true that "{{hypothesis}}"? Yes, no,
or maybe?
```

```
{{ answer_choices[label] }}
```

1.3.4 HANS

Dataset from McCoy et al. (2019). Used in evaluation.

Data Example

Key	Value
binary_parse_hypothesis	((The scientist) ((supported (the doctors) ...
binary_parse_premise	((The doctors) ((supported (the scientist) ...
heuristic	lexical_overlap
hypothesis	The scientist supported the doctors .
label	1
parse_hypothesis	(ROOT (S (NP (DT The) (NN scientist)) (VP (VBD sup...
parse_premise	(ROOT (S (NP (DT The) (NNS doctors)) (VP (VBD supp...
premise	The doctors supported the scientist .
subcase	ln_subject/object_swap
template	temp1

Prompts

```
{{premise}}

Question: Does this imply that "{{hypothesis}}"? Yes or no?
```

```
{{answer_choices[label]}}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Should we assume that "{{hypothesis}}" is true? Yes or
no?
```

```
{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{premise}} Based on the previous passage, is it true that
"{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Does it follow that {{hypothesis}} Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Given {{premise}} Is it guaranteed true that "{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

```
Given that {{premise}} Therefore, it must be true that "{{hypothesis}}"?
Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Williams et al. (2018)

```
{{premise}} Using only the above description and what you know about the
world, is "{{hypothesis}}" definitely correct? Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
{{premise}} Are we justified in saying that "{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

Prompt from Brown et al. (2020)

```
{{premise}}
Question: {{hypothesis}} True or False?
```

```
{{ answer_choices[label] }}
```

Prompt from Webson and Pavlick (2021)

```
Suppose {{premise}} Can we infer that "{{hypothesis}}"? Yes or no?
```

```
{{ answer_choices[label] }}
```

1.4 PARAPHRASE

1.4.1 GLUE MRPC

Dataset from Dolan and Brockett (2005). Used in evaluation.

Data Example

Key	Value
idx	0
label	1
sentence1	Amrozi accused his brother , whom he called " the ...
sentence2	Referring to him as only " the witness " , Amrozi ...

Prompts

Prompt not from the original task.

```
{% if label == 1 %}
Paraphrase the following sentence: {{sentence1}}
```

```
{{sentence2}}
{% endif %}
```

```
I want to know whether the following two sentences mean the same thing.
{{sentence1}}
{{sentence2}}
Do they?
```

```
{{ answer_choices[label] }}
```

```
Does the sentence
{{sentence1}}
paraphrase (that is, mean the same thing as) this sentence?
{{sentence2}}
```

```
{{ answer_choices[label] }}
```

```
Are the following two sentences "{{"equivalent"}}" or "{{"not
equivalent"}}"?
{{sentence1}}
{{sentence2}}
```

```
{{ answer_choices[label] }}
```

Prompt not from the original task.

```
{% if label == 1 %}
Generate a sentence that means the same thing as this one: {{sentence1}}
```

```
{{sentence2}}
{% endif %}
```

```
Can I replace the sentence
{{sentence1}}
with the sentence
{{sentence2}}
and have it mean the same thing?
```

```
{{ answer_choices[label] }}
```

```
Do the following two sentences mean the same thing?
{{sentence1}}
{{sentence2}}
```

```
{{ answer_choices[label] }}
```

1.4.2 GLUE QQP

Dataset from Iyer et al. (2017). Used in evaluation.

Data Example

Key	Value
idx	0
label	0
question1	How is the life of a math student? Could you descr...
question2	Which level of prepration is enough for the exam j...

Prompts

```
I'm an administrator on the website Quora. There are two posts, one that
asks "{{question1}}" and another that asks "{{question2}}". I can merge
questions if they are asking the same thing. Can I merge these two
questions?
```

```
{{ answer_choices[label] }}
```

```
{{question1}}
{{question2}}
Pick one: These questions are "{{"duplicates"}}" or "{{"not
duplicates"}}".
```

```
{{ answer_choices[label] }}
```

```
Are the questions "{{question1}}" and "{{question2}}" asking the same
thing?
```

```
{{ answer_choices[label] }}
```

Prompt not from the original task.

```
Can an answer to "{{question1}}" also be used to answer "{{question2}}"?
```

```
{{ answer_choices[label] }}
```

```
Question 1: {{question1}}
Question 2: {{question2}}

Do these two questions convey the same meaning? Yes or no?
```

```
{{answer_choices[label]}}
```

```
I received the questions "{{question1}}" and "{{question2}}". Are they
duplicates?
```

```
{{ answer_choices[label] }}
```

1.4.3 PAWS LABELED_FINAL

Dataset from Zhang et al. (2019). Used in training.

Data Example

Key	Value
id	1
label	0
sentence1	In Paris , in October 1560 , he secretly met the E...
sentence2	In October 1560 , he secretly met with the English...

Prompts

```
Determine if the following two sentences paraphrase each other or not.
Sent 1: {{sentence1}}
Sent 2: {{sentence2}}
```

```
{{answer_choices[label]}}
```

```
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Do Sentence 1 and Sentence 2 express the same meaning? Yes or
No?
```

```
{{answer_choices[label]}}
```

```
{{sentence1}}
Is that a paraphrase of the following sentence?
{{sentence2}}?
```

```
{{answer_choices[label]}}
```

```
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Can we rewrite Sentence 1 to Sentence 2?
```

```
{{answer_choices[label]}}
```

```
{{sentence1}}
Is that a paraphrase of the following sentence?
{{sentence2}}?
Yes or No.
```

```
{{answer_choices[label]}}
```

```
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Does Sentence 1 paraphrase Sentence 2? Yes or No?
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
{% if label == 1 %}
Paraphrase the sentence: {{sentence1}}
{{sentence2}}
{% endif %}
```

```
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Does Sentence 1 paraphrase Sentence 2?
{{answer_choices[label]}}
```

```
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Do Sentence 1 and Sentence 2 express the same meaning?
{{answer_choices[label]}}
```

Prompt from Brown et al. (2020)

```
{{sentence1}} Question: {{sentence2}} True or False?
{{answer_choices[label]}}
```

```
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}
Question: Can we rewrite Sentence 1 to Sentence 2? Yes or No?
{{answer_choices[label]}}
```

Prompt from Brown et al. (2020)

```
{{sentence1}} Question: {{sentence2}} Paraphrase or not?
{{answer_choices[label]}}
```

1.5 QA Closed Book

1.5.1 ai2_arc ARC-Challenge

Dataset from Clark et al. (2018). Used in evaluation.

Data Example

Key	Value
answerKey	A
choices	{'label': ['A', 'B', 'C', 'D'], 'text': ['dry palm...
id	Mercury_SC_415702
question	George wants to warm his hands quickly by rubbing ...

Prompts

Prompt not from the original task.

```
Pick and copy all the incorrect options for the following question:

{{question}}

Options:
- {{choices["text"] | join("\n- ")}}
```

```
{% for i in range(choices["label"]|length) %}
{% if i != choices["label"].index(answerKey) %}
- {{choices["text"][i]}}
{% endif %}
{% endfor %}
```

```
A ||| B ||| C ||| D
```

```
Here's a problem to solve: {{question}}

Among the 4 following options, which is the correct answer?
{% for letter, t in zip(answer_choices, choices.text) %}
- {{letter}}: {{t}}
 {% endfor %}
```

```
{{answerKey}}
```

```
{{choices.text | join("|||")}}
```

```
{{question}}

Options:
- {{answer_choices | join("\n- ")}}
```

```
{{answer_choices[choices["label"].index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
I am hesitating between 4 options to answer the following question, which
option should I choose?
Question: {{question}}
Possibilities:
- {{answer_choices | join("\n- ")}}
```

```
{{answer_choices[choices["label"].index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
I gave my students this multiple choice question: {{question}}

Only one answer is correct among these 4 choices:
- {{answer_choices | join("\n- ")}}

Could you tell me which one is correct?
```

```
{{answer_choices[choices["label"].index(answerKey)]}}
```

```
A ||| B ||| C ||| D
```

```
Pick the most correct option to answer the following question.

{{question}}

Options:
{% for letter, t in zip(answer_choices, choices.text) %}
- {{letter}}: {{t}}
{% endfor %}
```

```
{{answerKey}}
```

1.5.2 AI2_ARC ARC-EASY

Dataset from Clark et al. (2018). Used in evaluation.

Data Example

Key	Value
answerKey	B
choices	{'label': ['A', 'B', 'C', 'D'], 'text': ['a leg mu...
id	Mercury_7220990
question	Which factor will most likely cause a person to de...

Prompts

```
A ||| B ||| C ||| D
```

```
Pick the most correct option to answer the following question.

{{question}}

Options:
{% for letter, t in zip(answer_choices, choices.text) %}
- {{letter}}: {{t}}
{% endfor %}
```

```
{{answerKey}}
```

```
{{choices.text | join("|||")}}
```

```
{{question}}

Options:
- {{answer_choices | join("\n- ")}}
```

```
{{answer_choices[choices["label"].index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
I am hesitating between 4 options to answer the following question, which
option should I choose?
Question: {{question}}
Possibilities:
- {{answer_choices | join("\n- ")}}
```

```
{{answer_choices[choices["label"].index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
I gave my students this multiple choice question: {{question}}

Only one answer is correct among these 4 choices:
- {{answer_choices | join("\n- ")}}

Could you tell me which one is correct?
```

```
{{answer_choices[choices["label"].index(answerKey)]}}
```

Prompt not from the original task.

```
Pick and copy all the incorrect options for the following question:

{{question}}

Options:
- {{choices["text"] | join("\n- ")}}
```

```
{% for i in range(choices["label"]|length) %}
{% if i != choices["label"].index(answerKey) %}
- {{choices["text"][i]}}
{% endif %}
{% endfor %}
```

```
A ||| B ||| C ||| D
```

```
Here's a problem to solve: {{question}}

Among the 4 following options, which is the correct answer?
{% for letter, t in zip(answer_choices, choices.text) %}
- {{letter}}: {{t}}
 {% endfor %}
```

```
{{answerKey}}
```

1.5.3 KILT_TASKS NQ

Dataset from Petroni et al. (2021). Used in evaluation.

Data Example

Key	Value
id	5328212470870865242
input	how i.met your mother who is the mother
meta	{'left_context': '', 'mention': '', 'right_context...
output	[{'answer': 'Tracy McConnell', 'meta': {'score': -...

Prompts

```
{% if output %}
The goal is to predict an English answer string for an input English
question. All questions can be answered using the contents of English
Wikipedia.
Question: {{input}}
Answer:
```

```
{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|choice }}
{% endif %}
```

```
{% if output %}
Search query: {{input}}
Response:
```

```
{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|choice }}
{% endif %}
```

Prompt not from the original task.

```
{% if output %}
Question : {{input}}
Answer :
```

```
{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|join(', ') }}
{% endif %}
```

Prompt not from the original task.

```
{% if output %}
Guess a question that has the answer
"{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|choice }}"
```

```
{{input}}?
{% endif %}
```

```
{% if output %}
Question : {{input}}
Answer :
```

```
{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|choice }}
{% endif %}
```

```
{% if output %}
I've always wondered: {{input}}
```

```
{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|choice }}
{% endif %}
```

```
{% if output %}
Answer the following question.
{{input}}
```

```
{{output|selectattr("answer")|map(attribute='answer')|reject("equalto",
"")|list|choice }}
{% endif %}
```

1.5.4 KILT_TASKS HOTPOTQA

Dataset from Petroni et al. (2021). Used in training.

Data Example

Key	Value
id	5a7a06935542990198eaf050
input	Which magazine was started first Arthur's Magazine...
meta	{'left_context': '', 'mention': '', 'right_context...
output	[{'answer': "Arthur's Magazine", 'meta': {'score':...

Prompts

Prompt not from the original task.

```
{% if output %}
Here's a complex question that requires someone to reason about the
input, can you answer it?
{{input}}

{{output | map(attribute="answer") | list | choice}}
{% endif %}
```

Prompt not from the original task.

```
{% if output %}
Combine facts and answer this: {{input}}

{{output | map(attribute="answer") | list | choice}}
{% endif %}
```

Prompt not from the original task.

```
{% if output %}
Formulate an answer to this elaborate question: {{input}}

{{output | map(attribute="answer") | list | choice}}
{% endif %}
```

Prompt not from the original task.

```
{% if output %}
FINAL EXAM

Question 1. {{input}}

{{output | map(attribute="answer") | list | choice}}
{% endif %}
```

Prompt not from the original task.

```
{% if output %}
{{input}}
```

```
{{output | map(attribute="answer") | list | choice}}
{% endif %}
```

1.5.5 TRIVIA_QA RC

Dataset from Joshi et al. (2017). Used in evaluation.

Data Example

Key	Value
answer	{'aliases': ['(Harry) Sinclair Lewis', 'Harry Sinc...
entity_pages	{'doc_source': [], 'filename': [], 'title': [], 'w...
question	Which American-born Sinclair won the Nobel Prize f...
question_id	tc_1
question_source	http://www.triviacountry.com/
search_results	{'description': ['The Nobel Prize in Literature 19...

Prompts

Prompt not from the original task.

```
{% if answer.aliases %}
    Guess a question that has the answer "{{answer.aliases|choice}}"

{{question}}
{% endif %}
```

```
{% if answer.aliases %}
    The goal is to predict an English answer string for an input English
question.
    Question : {{question}}
    Answer :

{{answer.aliases|choice}}
{% endif %}
```

```
{% if answer.aliases %}
    Answer the following question.
    {{question}}

{{answer.aliases|choice}}
{% endif %}
```

```
{% if answer.aliases %}
   I've always wondered: {{question}}
```

```
{{answer.aliases|choice}}
{% endif %}
```

```
{% if answer.aliases %}
    Question : {{question}}
    Answer :

{{answer.aliases|choice}}
{% endif %}
```

1.5.6 WEB_QUESTIONS

Dataset from Berant et al. (2013). Used in evaluation.

Data Example

Key	Value
answers	['Jazmyn Bieber', 'Jaxon Bieber']
question	what is the name of justin bieber brother?
url	http://www.freebase.com/view/en/justin_bieber

Prompts

```
Give me the correct facts to answer this: {{question}}

{{answers | choice}}
```

```
Give me a possible correct answer to the question "{{ question }}"

{{ answers | choice }}
```

```
What's the answer to that question: {{question}}

{{answers | choice}}
```

```
Short general knowledge question: {{question}}

{{answers | choice}}
```

```
{{ question|capitalize }}
```

```
{{ answers | choice }}
```

1.5.7 WIKI_QA

Dataset from Yi et al. (2015). Used in training.

Data Example

Key	Value
answer	African immigration to the United States refers to...
document_title	African immigration to the United States
label	0
question	HOW AFRICAN AMERICANS WERE IMMIGRATED TO THE US
question_id	Q0

Prompts

```
No ||| Yes
```

```
Question: {{question}}?
Would "{{answer}}" be a reasonable answer?
```

```
{{ answer_choices[label] }}
```

```
No ||| Yes
```

```
I am verifying the answers generated by an automatic system to the
following question: {{question}}
Suggested answer: {{answer}}
Should I validate this answer?
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
{% if label == 1 %}
What is the question to: "{{answer}}"? The topic is {{document_title}}.
```

```
"{{question}}?"
{% endif %}
```

Prompt not from the original task.

```
{% if label == 1 %}
Determine the topic of the question-answer pair.
Question: "{{question}}?";  Answer: "{{answer}}"? Topic:
```

```
{{document_title}}
{% endif %}
```

Prompt not from the original task.

```
{% if label == 1 %}
Generate a question about the topic "{{document_title}}" whose answer
would be: {{answer}}.
```

```
{{question}}?
{% endif %}
```

```
No ||| Yes
```

```
Question: {{question}}
I found the following answer on Google: {{answer}}
Is that a correct answer? Yes or no.
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
{% if label == 1 %}
Determine the topic of the question.
Question: "{{question}}?"
Topic:
```

```
{{document_title}}
{% endif %}
```

```
False ||| True
```

```
The exercise is to decide whether the question accepts the proposed
suggestion as a correct answer. If yes, write "{{answer_choices[1]}}",
otherwise write "{{answer_choices[0]}}".
Question: {{question}}
Suggestion: {{answer}}
```

```
{{answer_choices[label]}}
```

```
No ||| Yes
```

```
This is a correct answer to the following question about
{{document_title}}. Yes or no?
Answer: {{answer}}
Question: {{question}}
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
{% if label == 1 %}
Determine the topic of the passage.
"{{answer}}"
Topic:
```

```
{{document_title}}
{% endif %}
```

```
{% if label == 1 %}
Answer this question: {{question}}?
```

```
{{answer}}
{% endif %}
```

1.6 QA EXTRACTIVE

1.6.1 ADVERSARIAL_QA DBIDAF

Dataset from Bartolo et al. (2020). Used in training.

Data Example

Key	Value
id	821607441c173838196c4d1500c2ab21a044e6b0
title	Yale_University
context	Slack (2003) compares three groups that conducted ...
question	what year were the research groups compared
answers	{'text': ['2003'], 'answer_start': [7]}
metadata	{'split': 'train', 'model_in_the_loop': 'BiDAF'}

Prompts

```
{% if metadata.split != "test" %}
Extract the answer to the question from the following context.
Question: {{question}}
Context: {{context}}
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Given the following passage

"{{context}}",

answer the following question. Note that the answer is present within the
text.

Question: {{question}}

{{answers.text | choice}}
{% endif %}
```

Prompt not from the original task.

```
I want to test the ability of students to read a passage and answer
questions about it. Could you please come up with a good question for the
passage "{{context}}"?

{{question}}
```

```
{% if metadata.split != "test" %}
I know that the answer to the question "{{question}}" is in
"{{context}}". Can you tell me what it is?

{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Question: "{{question}}"

Context: "{{context}}"

Answer:

{{answers.text | choice}}
{% endif %}
```

1.6.2 ADVERSARIAL_QA DBERT

Dataset from Bartolo et al. (2020). Used in training.

Data Example

Key	Value
id	dab017ed8a1c27c6afa2d8618abc3a477a4edffc
title	Empiricism
context	A generation later, the Irish Anglican bishop, Geo...
question	what concept is mentioned last?
answers	{'text': ['subjective idealism'], 'answer_start': ...
metadata	{'split': 'train', 'model_in_the_loop': 'BERT-Larg...

Prompts

Prompt not from the original task.

```
I want to test the ability of students to read a passage and answer
questions about it. Could you please come up with a good question for the
passage "{{context}}"?
```

```
{{question}}
```

```
{% if metadata.split != "test" %}
I know that the answer to the question "{{question}}" is in
"{{context}}". Can you tell me what it is?
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Question: "{{question}}"

Context: "{{context}}"

Answer:
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Extract the answer to the question from the following context.
Question: {{question}}
Context: {{context}}
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Given the following passage

"{{context}}",

answer the following question. Note that the answer is present within the
text.

Question: {{question}}
```

```
{{answers.text | choice}}
{% endif %}
```

1.6.3 ADVERSARIAL_QA DROBERTA

Dataset from Bartolo et al. (2020). Used in training.

Data Example

Key	Value
id	12cf36866b656dc4f254081fe6796ea1be2f6d43
title	Napoleon
context	When he became First Consul and later Emperor, Nap...
question	What jewelry like accessories did he wear?
answers	{'text': ["Légion d'honneur star, medal and ribbon...
metadata	{'split': 'train', 'model_in_the_loop': 'RoBERTa-L...

Prompts

Prompt not from the original task.

```
I want to test the ability of students to read a passage and answer
questions about it. Could you please come up with a good question for the
passage "{{context}}"?
```

```
{{question}}
```

```
{% if metadata.split != "test" %}
I know that the answer to the question "{{question}}" is in
"{{context}}". Can you tell me what it is?
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Question: "{{question}}"

Context: "{{context}}"

Answer:
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Extract the answer to the question from the following context.
Question: {{question}}
Context: {{context}}
```

```
{{answers.text | choice}}
{% endif %}
```

```
{% if metadata.split != "test" %}
Given the following passage

"{{context}}",
```

```
answer the following question. Note that the answer is present within the
text.

Question: {{question}}
```

```
{{answers.text | choice}}
{% endif %}
```

1.6.4 COQA

Dataset from Siva et al. (2018). Used in evaluation.

Data Example

Key	Value
answers	{'answer_end': [179, 494, 511, 545, 879, 1127, 112...
questions	['When was the Vat formally opened?', 'what is the...
source	wikipedia
story	The Vatican Apostolic Library (), more commonly ca...

Prompts

Prompt not from the original task.

```
Answer the question based on the information contained in the passage.
Q: {{questions[0]}}

Passage: {{story}}

A:
```

```
{{answers["input_text"][0]}}
```

```
Answer the last question based on the hint.
{% for question, answer in zip(questions[:-1],
answers["input_text"][:-1]) %}
Q: {{question}}

A:{{answer}}
{%endfor %}

Q: {{questions[-1]}}

Hint: {{story}}

A:
```

```
{{answers["input_text"][-1]}}
```

Prompt not from the original task.

```
Can you form a set of {{questions | length}} question-answer pairs about
the passage below?

Passage: {{story}}
```

```
{% for question, answer in zip(questions, answers["input_text"]) %}
Q: {{question}}

A: {{answer}}

{% endfor %}
```

Prompt not from the original task.

```
In the passage below, extract the part which answers the question:

Q: {{questions[0]}}
Passage: {{story}}
```

```
{{story[answers["answer_start"][0] : answers["answer_end"][0] ]}}
```

```
{% set missing_idx = range(questions|length)|choice %}

{% for i in range(questions|length) %}
Q: {{questions[i] }}

A:  {% if i !=missing_idx %}
{{answers["input_text"][i]}}
{%endif%}
{%endfor%}

Given the above conversation, give a suitable response to the missing
answer

Hint: {{story}}
```

```
{{answers["input_text"][missing_idx]}}
```

1.6.5 DUORC SELFRC

Dataset from Saha et al. (2018). Used in training.

Data Example

Key	Value
answers	['They arrived by train.']
no_answer	False
plot	200 years in the future, Mars has been colonized b...
plot_id	/m/03vyhn
question	How did the police arrive at the Mars mining camp?
question_id	b440de7d-9c3f-841c-eaec-a14bdff950d1
title	Ghosts of Mars

Prompts

Prompt not from the original task.

```
{% if no_answer == false%}
Generate a question that has the following answer:
{{answers|choice}}
for the following movie plot:
{{plot}}
```

```
{{question}}
{% endif %}
```

```
I am a movie director and I just received the following movie plot. Could
you help me answer this question? If not, let me know by writing "{{"Not
answerable"}}".

Plot title: {{title}}
Movie plot: {{plot}}
My question: {{question}}
```

```
{% if no_answer %}
Not answerable
{% else %}
{{answers|choice}}
{% endif %}
```

```
Extract the answer to the following question from the movie plot. If the
question isn't answerable, please output "{{"Can't answer"}}".
Question: {{question}}
Title: {{title}}
Movie plot: {{plot}}
```

```
{% if no_answer %}
Can't answer
{% else %}
{{answers | choice }}
{% endif %}
```

Prompt not from the original task.

```
Generate a question about the following movie plot: {{ plot }}
```

```
{{ question }}
```

```
Please answer the following question about this movie plot. If it's
un-answerable, please output "{{"No answer"}}".

Question: {{question}}
Movie plot title: {{title}}
Movie plot: {{plot}}
```

```
{% if no_answer %}
No answer
{% else %}
{{answers | choice }}
{% endif %}
```

Prompt not from the original task.

```
{% if no_answer == false%}
Build a movie plot around this: {{ question }} {{answers|choice}}

{{ plot }}
{% endif %}
```

```
Question: {{question}}
If there is no answer, please output "{{"Insufficient information to
provide an answer."}}".
Movie title: {{title}}
Context: {{plot}}

{% if no_answer %}
Insufficient information to provide an answer.
{% else %}
{{answers|choice}}
{% endif %}
```

Prompt not from the original task.

```
Suggest a movie title for the following movie plot: {{plot}}

{{title}}
```

```
I am trying to decide whether it's worth it to invest in this film
proposal. Can you help me answer a few questions? If you can't, please
say "{{"No I can't"}}".

Question: {{question}}
Movie title: {{title}}
Movie plot: {{plot}}

{% if no_answer %}
No I can't
{% else %}
{{answers|choice}}
{% endif %}
```

1.6.6 DUORC PARAPHRASERC

Dataset from Saha et al. (2018). Used in training.

Data Example

Prompts

Prompt not from the original task.

```
{% if no_answer == false%}
Build a movie plot around this: {{ question }} {{answers|choice}}
```

Key	Value
answers	['second in command Sergeant Jericho and prisoner...
no_answer	False
plot	Set in the second half of the 22nd century, Mars h...
plot_id	/m/03vyhn
question	who is there with Melanie Ballard?
question_id	28ded42d-f6d5-aac6-cf6f-9e6e0820c5aa
title	Ghosts of Mars

```
{{ plot }}
{% endif %}
```

```
I am trying to decide whether it's worth it to invest in this film
proposal. Can you help me answer a few questions? If you can't, please
say "{{"No I can't"}}".

Question: {{question}}
Movie title: {{title}}
Movie plot: {{plot}}
```

```
{% if no_answer %}
No I can't
{% else %}
{{answers|choice}}
{% endif %}
```

```
Question: {{question}}
If there is no answer, please output "{{"Insufficient information to
provide an answer."}}".
Movie title: {{title}}
Context: {{plot}}
```

```
{% if no_answer %}
Insufficient information to provide an answer.
{% else %}
{{answers|choice}}
{% endif %}
```

```
I am a movie director and I just received the following movie plot. Could
you help me answer this question? If not, let me know by writing "{{"Not
answerable"}}".

Plot title: {{title}}
Movie plot: {{plot}}
My question: {{question}}
```

```
{% if no_answer %}
Not answerable
{% else %}
{{answers|choice}}
{% endif %}
```

Prompt not from the original task.

```
Generate a question about the following movie plot: {{ plot }}
```

```
{{ question }}
```

```
Extract the answer to the following question from the movie plot. If the
question isn't answerable, please output "{{"Can't answer"}}".
Question: {{question}}
Title: {{title}}
Movie plot: {{plot}}
```

```
{% if no_answer %}
Can't answer
{% else %}
{{answers | choice }}
{% endif %}
```

Prompt not from the original task.

```
Suggest a movie title for the following movie plot: {{plot}}
```

```
{{title}}
```

```
Please answer the following question about this movie plot. If it's
un-answerable, please output "{{"No answer"}}".

Question: {{question}}
Movie plot title: {{title}}
Movie plot: {{plot}}
```

```
{% if no_answer %}
No answer
{% else %}
{{answers | choice }}
{% endif %}
```

Prompt not from the original task.

```
{% if no_answer == false%}
Generate a question that has the following answer:
{{answers|choice}}
for the following movie plot:
{{plot}}
```

```
{{question}}
{% endif %}
```

1.6.7 ROPES

Dataset from Lin et al. (2019). Used in training.

Data Example

Key	Value
answers	{'text': ['cup B']}
background	Passive transport occurs when a substance passes t...
id	1971664873
question	Which cup has a higher concentration of sugar?
situation	A man put two cups, cup A and cup B, filled with e...

Prompts

```
{% if answers.text %}
Please answer correctly the following question related to the paragraph
below.

{{ question }}

{{ situation }}

Hint: {{ background }}

{{ answers.text | choice }}
{% endif %}
```

Prompt not from the original task.

```
{% if answers.text %}
{{ situation }}

Given the paragraph above, please answer correctly the following
question:

{{ question }}

{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
Background: {{ background }}

Paragraph: {{ situation }}

Given the paragraph above, please answer correctly the following
question: {{ question }}

{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
Given the background: {{background}}

and the situation: {{situation}}

Answer the following question: {{question}}
```

```
{{ answers.text | choice }}
{% endif %}
```

Prompt not from the original task.

```
{% if answers.text %}
{{ situation }}

{{ question }}

{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
{{ situation }}

{{ question }}

Hint: {{ background }}

{{ answers.text | choice}}
{% endif %}
```

```
{% if answers.text %}
{{ background }}

{{ situation }}

{{ question }}

{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
I can use this background: {{background}}

Now, I have a new situation: {{situation}}

Answer this question please: {{question}}

{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
You are given a new situation: {{situation}}

and a hint : {{background}}

Please answer this question : {{question}}
```

```
{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
I have a new situation: {{situation}}

But I can use this background: {{background}}

What is an answer for this question: {{question}}
{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
{{ situation }}

Given the paragraph above, please answer correctly the following
question:

{{ question }}

Hint: {{ background }}
{{ answers.text | choice }}
{% endif %}
```

```
{% if answers.text %}
I read this background article the other day: {{background}}

I am facing a new situation today: {{situation}}

Using the knowledge I acquired from the background article, how should I
answer correctly the following question regarding my new situation:
{{question}}
{{ answers.text | choice }}
{% endif %}
```

1.6.8 SQUAD_V2

Dataset from Rajpurkar et al. (2016). Used in evaluation.

Data Example

Key	Value
id	56be85543aeaaa14008c9063
title	Beyoncé
context	Beyoncé Giselle Knowles-Carter ...
question	When did Beyonce start becoming popular?
answers	{'text': ['in the late 1990s'], 'answer_start': [2...

Prompts

```
{% set seq = [
'Answer the question depending on the context.',
'What is the answer?',
] %}

{{ seq | choice }}
Context: {{context}};
Question: {{question}};
Answer:
```

```
{% if answers.text == [] %}
Answer not in context
{% else %}
{{answers.text[0]}}
{% endif %}
```

Prompt not from the original task.

```
{% if answers.text != [] %}
Determine the question that you might have asked to get back the
following answer for the given context
Context: {{context}};
Answer: {{answers.text[0]}};
Question:
```

```
{{question}}
{% endif %}
```

Prompt not from the original task.

```
{% set seq = [
'What is this about? ',
'What is the paragraph about? ',
'Get the topic from: ',
'From the passage,  get the topic',
'I want to know the topic. ',
'Topic from the passage: ',
'Topic from the paragraph: ',
] %}
{{ seq | choice }}
{{context}}
```

```
{{title | replace("_", " ")}}
```

Prompt not from the original task.

```
{% set seq = [
'This is about ',
'What is this about? ',
'The paragraph is about ',
'What is the paragraph about? ',
'Get the topic: ',
'From the passage, the topic is',
'I want to know the topic. ',
'Topic from the passage: ',
'Topic from the paragraph: ',
```

```
] %}
{{context}}
{{ seq | choice }}
```

```
{{title | replace("_", " ")}}
```

Prompt not from the original task.

```
{% if answers.text != [] %}
What is a question that would give the following answer?
Answer: {{answers.text[0]}};
Question:
```

```
{{question}}
{% endif %}
```

```
{% set seq = [
'Can you tell me ',
'Please tell me ',
'Tell me ',
'From the passage, ',
'I want to know ',
'I want to ask ',
'What is the answer to: ',
'Find the answer to: ',
'Answer: ',
'',
] %}
{{context}} {{ seq | choice }}{{question}}
```

```
{% if answers.text == [] %}
Answer not in context
{% else %}
{{answers.text[0]}}
{% endif %}
```

Prompt not from the original task.

```
{% if answers.text != [] %}
{{question}}
```

```
{{answers.text[0]}}
{% endif %}
```

Prompt not from the original task.

```
Context: {{context}};

Question: {{question}}

Is this question answerable?
```

```
{% if answers.text != [] %}
{{answer_choices[0]}}
{% else %}
{{answer_choices[1]}}
{% endif %}
```

Prompt not from the original task.

```
{% set seq = [
'Determine the topic of the question-answer pair. ',
'Find the topic. ',
'What is the topic from this? ',
] %}
{% if answers.text != [] %}
{{ seq | choice }}
Question: {{question}};  Answer: {{answers.text[0]}}; Topic:
```

```
{{title}}
{% endif %}
```

Prompt not from the original task.

```
What is the following passage about?
{{context}}
```

```
{{title | replace("_", " ")}}
```

1.6.9 SUPER_GLUE RECORD

Dataset from Zhang et al. (2018). Used in evaluation.

Data Example

Key	Value
answers	['Nuria']
entities	['Afghanistan', 'Badam Bagh', 'Mariam', 'Nuria']
idx	{'passage': 0, 'query': 0}
passage	The harrowing stories of women and children locked...
query	The baby she gave birth to is her husbands and he ...

Prompts

```
{{ entities | join("|||") }}
```

```
{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}
Which one is the "{{"@placeholder"}}"? {{ entities | join(", ") }}?
```

```
{{ answers | choice }}
{% endif %}
```

```
{{ entities | join("|||") }}


{% if ( answers | length ) > 0 %}
The following document has been corrupted. Tell me what
"{{"@placeholder"}}" is referring to.

Document: {{ passage }}
{{ query }}

{{ answers | choice }}
{% endif %}
```

```
{{ entities | join("|||") }}


{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}

You should decide what "{{"@placeholder"}}" is referring to. Choose
between:
- {{answer_choices | join("\n- ")}}

{{ answers | choice }}
{% endif %}
```

```
{{ entities | join("|||") }}


{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}
In the question above, the "{{"@placeholder"}}" stands for

{{ answers | choice }}
{% endif %}
```

```
{{ entities | join("|||") }}


{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}
What could the "{{"@placeholder"}}" be? {{ entities | join(", ") }}?
```

```
{{ answers | choice }}
{% endif %}
```

```
{{entities | join("|||")}}
```

```
{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}

I am trying to decide what "{{"@placeholder"}}" means in the previous
text.
Help by choosing an option between:
- {{ entities | join("\n- ") }}

{{ answers | choice }}
{% endif %}
```

```
{{ entities | join("|||") }}
```

```
{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}
Here, the placeholder refers to

{{ answers | choice }}
{% endif %}
```

```
{{entities | join("|||")}}
```

```
{% if ( answers | length ) > 0 %}
Exercise: Extract from the text the correct entity that
"{{"@placeholder"}}" is referring to.

{{ passage }}
{{ query }}

{{ answers | choice }}
{% endif %}
```

```
{{entities | join("|||")}}
```

```
{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}

Pick one option, "{{"@placeholder"}}" refers to:
- {{answer_choices | join("\n- ")}}

{{ answers | choice }}
{% endif %}
```

```
{{ entities | join("|||") }}
```

```
{% if ( answers | length ) > 0 %}
{{ passage }}
{{ query }}
Can you figure out what does the "{{"@placeholder"}}" mean? It means

{{ answers | choice }}
{% endif %}
```

1.6.10 QA_SRL

Used in evaluation.

Data Example

Key	Value
answers	['four boat clubs', 'Aberdeen Boat Club', 'Aberdee...
predicate	row
predicate_idx	6
question	['what', '_', '_', 'rows', '_', '_', '_', '?']
sent_id	WIKI1_0
sentence	There are four boat clubs that row on the River De...

Prompts

Prompt not from the original task.

```
Generate a plausible question that has the following answers based on the
context:

Context: {{sentence}}

Answers: {{answers | join(", ")}}

{{question | join(" ") | replace("_ ", "")}}
```

```
The English teacher deconstructed an example sentence that contained the
verb "{{predicate}}": {{sentence}}

{{question | join(" ") | replace("_ ", "")}}
```

```
{{answers | choice}}
```

Prompt not from the original task.

```
Identify the predicate (the part of a sentence or clause containing a
verb and stating something about the subject) in this sentence:

{{sentence}}
```

```
{{predicate}}
```

```
{{sentence}}
{{question|join(" ")|replace("_ ", "")}}
```

```
{{answers | choice}}
```

```
Here's a linguistic problem: you have to correctly identify the part of
the sentence that answers the following {{"W"}} question.
Sentence: {{sentence}}
Question: {{question | join(" ") | replace("_ ", "")}}
```

```
{{answers | choice}}
```

```
Help me parse the structure of the following sentence constructed around
the verb "{{predicate}}": {{sentence}}

{{question | join(" ") | replace("_ ", "")}}
```

```
{{answers | choice}}
```

```
{{sentence}}

The previous sentence contains the verb "{{predicate}}". Answer this
question about it: {{question|join(" ")|replace("_ ", "")}}
```

```
{{answers | choice}}
```

1.6.11 QUAC

Dataset from Choi et al. (2018). Used in evaluation.

Data Example

Key	Value
answers	{'answer_starts': [[51], [640], [1862], [2024], [2...
background	The Malayali people or Keralite people (also spelt...
context	According to the Indian census of 2001, there were...
dialogue_id	C_69758fcdfc1f46baba0e92c0f3b0919c_1
followups	[2, 1, 1, 1, 1, 1, 1]
orig_answers	{'texts': ['30,803,747 speakers of Malayalam in Ke...
questions	['Where is Malayali located?', 'What other languag...
section_title	Geographic distribution and population
turn_ids	['C_69758fcdfc1f46baba0e92c0f3b0919c_1_q#0', 'C_69...
wikipedia_page_title	Malayali
yesnos	[2, 2, 2, 2, 2, 0, 2]

Prompts

Prompt not from the original task.

```
Given the  partial dialogue :

Student: {{questions[0]}}

Teacher: {{(answers.texts[0] | choice).replace("CANNOTANSWER","Cannot
answer") }}

The context : {{context}}

Answer the question: {{questions[1] }}
```

```
{{(answers.texts[1] | choice).replace("CANNOTANSWER","Cannot answer")
}}
```

```
Given the dialogue:
{% for i in range(0, questions | length - 1)%}
Student: {{questions[i]}}

Teacher: {{(answers.texts[i] | choice).replace("CANNOTANSWER","Cannot
answer") }}
{% endfor %}

The context: {{context}}

Answer the question: {{questions | last }}
```

```
{{(answers.texts | last | choice).replace("CANNOTANSWER","Cannot
answer") }}
```

```
This conversation happened between a teacher and a student:
{% for i in range(0, questions | length - 1) %}
Student: {{questions[i]}}

Teacher: {{(answers.texts[i] | choice).replace("CANNOTANSWER","Cannot
answer") }}
{% endfor %}

Use the article : {{context}} to answer the question: {{questions | last
}}
```

```
{{(answers.texts | last | choice).replace("CANNOTANSWER","Cannot
answer") }}
```

```
I read an article : {{context}}

Then the following conversation occurred:
{% for i in range(0, questions | length - 1) %}
Student: {{questions[i]}}

Teacher: {{(answers.texts[i] | choice).replace("CANNOTANSWER","Cannot
answer") }}
{% endfor %}
Use both to answer the question:  {{questions | last }}
```

```
{{(answers.texts | last | choice).replace("CANNOTANSWER","Cannot
answer") }}
```

```
Read the article: {{context}}

Then answer the question:  {{questions | last}}

You can use this dialogue to find the answer faster:
{% for i in range(0, questions | length - 1)%}
Student: {{questions[i]}}

Teacher: {{(answers.texts[i] | choice).replace("CANNOTANSWER","Cannot
answer") }}
{% endfor %}
```

```
{{(answers.texts | last | choice).replace("CANNOTANSWER","Cannot
answer")  }}
```

```
A student is asking a teacher about the following article:
{{context}}

This is a summary of their conversation:
{% for i in range(0, questions | length - 1)%}
Student: {{questions[i]}}

Teacher: {{(answers.texts[i] | choice).replace("CANNOTANSWER","Cannot
answer") }}
{% endfor %}

Use their conversation and the article to answer the question :
{{questions | last}}
```

```
{{(answers.texts | last | choice).replace("CANNOTANSWER","Cannot
answer")  }}
```

1.6.12 QUOREF

Dataset from Dasigi et al. (2019). Used in training.

Data Example

Key	Value
answers	{'answer_start': [250], 'text': ['Catherine']}
context	The earthquake swarm was noted on October 12, 2007...
id	ba3f052c7a557909526b59713430403dd134e01d
question	What is the first name of the person who doubted i...
title	2007{2008 Nazko earthquakes 1
url	https://en.wikipedia.org/wiki/2007%E2%80%932008_Na...

Prompts

```
The answer to the question: {{question}} is inside the article:
{{context}}, can you guess it ?
```

```
{{answers.text | choice}}
```

```
Given the following context:

{{context}}

answer the following question:

{{question}}
```

```
{{answers.text | choice}}
```

```
The following article contains an answer for the question: {{question}} ,
can you please find it?

{{context}}
```

```
{{answers.text | choice}}
```

```
This article: {{context}} contains an answer for the question:
{{question}}, what is it ?
```

```
{{answers.text | choice}}
```

```
{{question}}

Answer the above question based on the context below:

{{context}}
```

```
{{answers.text | choice}}
```

```
What is the answer for the question: {{question}} from the following
article ?

{{context}}
```

```
{{answers.text | choice}}
```

```
I have a test where I am given the following article, what is an answer
for the question: {{question}} ?

{{context}}
```

```
{{answers.text | choice}}
```

Prompt not from the original task.

```
Given the below context:

{{context}}

Guess a valid title for it!
```

```
{{title}}
```

```
Found the following article online, use it to answer the question:
{{question}}

{{context}}
```

```
{{answers.text | choice}}
```

```
A friend asked me to answer this question: {{question}}, using the
article: {{context}}, what would be the answer ?
```

```
{{answers.text | choice}}
```

```
Read the following paragraph and extract the answer for the question:
{{question}}

{{context}}
```

```
{{answers.text | choice}}
```

1.7 QA GENERATIVE

1.7.1 DROP

Dataset from Dua et al. (2019). Used in evaluation.

Key	Value
section_id	nfl_2201
query_id	f16c0ee7-f131-4a8b-a6ac-4d275ea68066
passage	To start the season, the Lions traveled south to T...
question	How many points did the buccaneers need to tie in ...
answers_spans	{'spans': ['3'], 'types': ['number']}

Data Example

Prompts

```
Question: {{question}}
Answer based on following passage.

{{passage}}

Answer:
```

```
{{ answers_spans.spans | join(", ") }}
```

```
I am trying to figure out the answer to the question, "{{question}}" I
found the following text-snippet has the answer. Can you tell me the
answer?

{{passage}}
```

```
{{ answers_spans.spans | join(", ") }}
```

Prompt from Brown et al. (2020)

```
Passage: {{passage}}
Question: {{question}}
Answer:
```

```
{{ answers_spans.spans | join(", ") }}
```

Prompt not from the original task.

```
Generate a question from the following passage that has the answer, {{
answers_spans.spans | join(", ") }}
Passage : {{passage}}
Question :
```

```
{{question}}
```

```
Context: {{passage}}
I am trying to figure out the answer to the question from the above
context.  Can you tell me the answer?
Question: {{question}}
Answer:
```

```
{{ answers_spans.spans | join(", ") }}
```

1.8 QA MULTIPLE CHOICE

1.8.1 COS_E V1.11

Used in training.

Data Example

Key	Value
abstractive_explanation	webmath is designed to help you solve
answer	math problem
choices	['park', 'coloring book', 'garden center', 'math p...
extractive_explanation	"there are 10 apples on an apple tree. three fall ...
id	6b819727eb8a670df26a7ffad036c119
question	"There are 10 apples on an apple tree. Three fall...

Prompts

```
{{ choices | join("|||") }}
```

```
{{ question }}
Choose the most suitable option to answer the above question.
Options:
- {{ answer_choices | join("\n- ") }}
```

```
{{ answer }}
```

```
{{ question }}
Choose the most suitable option to answer the above question.
Options:
{% for k in range(choices | length) %}
{{'. '.join([answer_choices[k], choices[k]])}}
{% endfor %}
```

```
{{ answer_choices[choices.index(answer)] }}
```

Prompt not from the original task.

```
Question: {{question}}

Choices:
- {{ choices | join("\n- ") }}

The rationale to choose "{{answer}}" as the answer is that:
```

```
{{abstractive_explanation}}
```

```
{{ choices | join("|||") }}
```

```
{{ question }}
- {{ answer_choices | join("\n- ") }}

The best answer is
```

```
{{ answer }}
```

Prompt not from the original task.

```
Here's a question and a few possible answers:

Q: {{ question }}
Possible A: {{ choices | join(", ") }}

Why is "{{answer}}" an answer aligned with human common sense?
```

```
{{ abstractive_explanation }}
```

```
Pick the option in line with common sense to answer the question.
Question: {{ question }}
Options:
{% for k in range(choices | length) %}
{{'. '.join([answer_choices[k], choices[k]])}}
{% endfor %}
```

```
{{ answer_choices[choices.index(answer)] }}
```

Prompt not from the original task.

```
Question: {{ question }}
Options:
- {{ choices | join("\n- ") }}

Explain why a human would choose "{{answer}}" to answer the question
above:
```

```
{{ abstractive_explanation }}
```

Prompt not from the original task.

```
Question: {{ question }}
Options:
- {{ choices | join("\n- ") }}

The answer is "{{ answer }}" because
```

```
{{ abstractive_explanation }}
```

```
{{ choices | join("|||") }}
```

```
Pick the option in line with common sense to answer the question.
Questions: {{ question }}
Options:
- {{ answer_choices | join("\n- ") }}
```

```
{{ answer }}
```

Prompt not from the original task.

```
Here's a question: {{ question }}

Here are possible answers to this question:
- {{ choices | join("\n- ") }}

I believe the correct choice is "{{answer}}", here's why:
```

```
{{ abstractive_explanation }}
```

```
{{ question }}
{% for k in range(choices | length) %}
{{'. '.join([answer_choices[k], choices[k]])}}
{% endfor %}
The best answer is
```

```
{{ answer_choices[choices.index(answer)] }}
```

1.8.2 COSMOS_QA

Dataset from Huang et al. (2019). Used in training.

Data Example

Key	Value
answer0	None of the above choices .
answer1	This person likes music and likes to see the show ...
answer2	This person only likes Good Old War and Person L ,...
answer3	Other Bands is not on tour and this person can not...
context	Good Old War and person L : I saw both of these ba...
id	3Q9SPIIRWJKVQ8244310E8TUS6YWAC##34V1S5K3GTZMDUBNBI...
label	1
question	In the future , will this person go to see other b...

Prompts

Prompt not from the original task.

```
Based on the context and the answer, generate a question.

Context: {{context}}

Answer:
{% if label == 0 %}
{{answer0}}
{% elif label == 1 %}
{{answer1}}
{% elif label == 2 %}
{{answer2}}
{% elif label == 3 %}
{{answer3}}
{% endif %}
```

```
{{question}}
```

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
Read the following context and choose the best option to answer the
question.
Context: {{ context }}
Question: {{ question }}
Options:
- {{ answer_choices | join("\n - ") }}
```

```
{{ answer_choices[label] }}
```

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
Read the following context and answer the question.
Context: {{ context }}
Question: {{ question }}
Answer:
```

```
{{ answer_choices[label] }}
```

```
Read the following context and choose the best option to answer the
question.
Context: {{ context }}
Question: {{ question }}
Options:
A. {{ answer0 }}
B. {{ answer1 }}
C. {{ answer2 }}
D. {{ answer3 }}
```

```
{{ answer_choices[label] }}
```

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
{{ context }}
According to the above context, choose the best option to answer the
following question.
Question: {{ question }}
Options:
- {{answer_choices | join("\n - ")}}
```

```
{{answer_choices[label]}}
```

```
{{ context }}
{{ question }}
A. {{ answer0 }}
B. {{ answer1 }}
C. {{ answer2 }}
D. {{ answer3 }}
```

```
{{ answer_choices[label] }}
```

Prompt not from the original task.

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
{{ context }}
Question: {{ question }}
The answer to the above question:
```

```
{{ answer_choices[label] }}
```

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
{{ context }}
{{ question }}
- {{ answer_choices | join("\n - ") }}
```

```
{{ answer_choices[label] }}
```

```
{{ context }}
According to the above context, choose the best option to answer the
following question.
Question: {{ question }}
Options:
A. {{ answer0 }}
B. {{ answer1 }}
C. {{ answer2 }}
D. {{ answer3 }}
```

```
{{ answer_choices[label] }}
```

```
{{ context }}
{{ question }}
Pick the best answer from the following options:
A. {{ answer0 }}
B. {{ answer1 }}
C. {{ answer2 }}
D. {{ answer3 }}
```

```
{{ answer_choices[label] }}
```

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
{{ context }}
According to the above context, answer the following question.
{{ question }}
```

```
{{answer_choices[label]}}
```

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
{{ context }}
{{ question }}
Pick the best answer from the following options:
- {{ answer_choices | join("\n - ") }}
```

```
{{ answer_choices[label] }}
```

Prompt not from the original task.

```
{{answer0}} ||| {{answer1}} ||| {{answer2}} ||| {{answer3}}
```

```
{{question}}
```

```
{{ answer_choices[label] }}
```

1.8.3 DREAM

Dataset from Sun et al. (2019). Used in training.

Key	Value
answer	Continue her dancing class.
choice	['Consult her dancing teacher.', 'Take a more inte...
dialogue	['M: I am considering dropping my dancing class. I...
dialogue_id	5-510
id	0
question	What does the man suggest the woman do?

Data Example

Prompts

Prompt not from the original task.

```
Read the below conversation.

{{dialogue[:-1] | join("\n\n")}}

What would the listener say?
```

```
{{dialogue[-1]}}
```

Prompt not from the original task.

```
Given the question "{{question}}" and the answer "{{answer}}", write a
conversation that might have happened.
```

```
{{dialogue | join("\n\n")}}
```

Prompt not from the original task.

```
{{dialogue[1:] | join("\n\n")}}

What was said before this conversation?
```

```
{{dialogue[0]}}
```

```
{{choice | join("|||")}}
```

```
Dialogue:

{{dialogue | join("\n\n")}}

Question: {{question}}

- {{answer_choices[0]}}

- {{answer_choices[1]}}

- {{answer_choices[2]}}
```

```
{{answer}}
```

```
{{choice | join("|||")}}
```

```
Read the following conversation and answer the question.

{{dialogue | join("\n\n")}}

Question: {{question}}

- {{answer_choices[0]}}

- {{answer_choices[1]}}

- {{answer_choices[2]}}
```

```
{{answer}}
```

1.8.4 OPENBOOKQA MAIN

Dataset from Mihaylov et al. (2018). Used in evaluation.

Data Example

Key	Value
answerKey	D
choices	{'label': ['puppies learning new tricks', 'childre...
id	7-980
question_stem	The sun is responsible for

Prompts

```
{{choices.text | join("|||")}}
```

```
{{question_stem}}

Choose an answer from this list:
- {{ answer_choices | join("\n- ") }}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answerKey]]}}
```

```
{{choices.text | join("|||")}}
```

```
{{question_stem}}

Which is the correct answer?
- {{ answer_choices | join("\n- ") }}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answerKey]]}}
```

```
{{question_stem}}
{% for k in range(choices["text"] | length) %}
{{' -> '.join([["A", "B", "C", "D"][k], choices["text"][k]])}}
{% endfor %}
Is the right answer {{"A, B, C or D"}} ?
```

```
{{answerKey}}
```

```
{{choices.text | join("|||")}}
```

```
{{question_stem}}

Choices:
- {{ answer_choices | join("\n- ") }}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answerKey]]}}
```

```
{{choices.text | join("|||")}}
```

```
{{question_stem}}
- {{ answer_choices | join("\n- ") }}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answerKey]]}}
```

```
{{choices.text | join("|||")}}
```

```
{{question_stem}}
- {{ answer_choices | join("\n- ") }}

Which is the correct answer?
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answerKey]]}}
```

```
{{choices.text | join("|||")}}
```

```
{{question_stem}}

Pick the right answer from the list:
- {{ answer_choices | join("\n- ") }}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answerKey]]}}
```

1.8.5 QASC

Dataset from Khot et al. (2020). Used in training.

Data Example

Key	Value
answerKey	F
choices	{'label': ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H']...
combinedfact	Beads of water can be formed by clouds.
fact1	beads of water are formed by water vapor condensin...
fact2	Clouds are made of water vapor.
formatted_question	What type of water formation is formed by clouds? ...
id	3E7TUJ2EGCLQNOV1WEAJ2NN9ROPD9K
question	What type of water formation is formed by clouds?

Prompts

Prompt not from the original task.

```
If I tell you that {{combinedfact[0]|capitalize}}{{
combinedfact[1:]|trim('.') }}, and ask you the question "{{
question[0]|lower }}{{ question[1:] }}", is the correct answer "{{
choices.text[0][0]|lower}}{{ choices.text[0][1:]|trim('.') }}"?
```

```
{% if answerKey == choices.label[0] %} Yes {% else %} No {% endif %}
```

```
{{choices.text | join("|||")}}
```

```
{{ fact1[0]|capitalize }}{{ fact1[1:]|trim|trim('.') }}, and
{{fact2[0]|lower }}{{ fact2[1:]|trim|trim('.') }}. Given these facts, {{
question[0]|lower }}{{question[1:]|trim('?') }} among the following
options:
- {{answer_choices | join("\n - ") }}
```

```
{% for choice in choices.label %} {% if choice == answerKey %}{{
answer_choices[loop.index - 1] }}{% endif %}{% endfor %}
```

```
{{choices.text | join("|||")}}
```

```
Fact 1: {{ fact1[0]|capitalize }}{{ fact1[1:]|trim|trim('.') }}.

Fact 2: {{fact2[0]|capitalize }}{{ fact2[1:]|trim|trim('.') }}.

Given the two facts above, {{ question[0]|lower
}}{{question[1:]|trim('?') }}?
```

```
{% for choice in choices.label %} {% if choice == answerKey %}{{
answer_choices[loop.index - 1] }}{% endif %}{% endfor %}
```

```
{{choices.text | join("|||")}}
```

```
You are presented with the question "{{ question }}" and the following
answer choices:
- {{answer_choices | join("\n - ") }}

Now knowing that {{ fact1[0]|lower }}{{ fact1[1:]|trim|trim('.') }} and
{{fact2[0]|lower }}{{ fact2[1:]|trim|trim('.') }}, choose the best
answer.
```

```
{% for choice in choices.label %} {% if choice == answerKey %}{{
answer_choices[loop.index - 1] }}{% endif %}{% endfor %}
```

```
{{choices.text | join("|||")}}
```

```
You are presented with the quiz "{{ question }}"

But you don't know the answer, so you turn to your teacher to ask for
hints. He says that "{{ fact1[0]|lower }}{{ fact1[1:]|trim|trim('.') }}"
and "{{fact2[0]|lower }}{{ fact2[1:]|trim|trim('.') }}".

So, what's the best answer to the question?
```

```
{% for choice in choices.label %} {% if choice == answerKey %}{{
answer_choices[loop.index - 1] }}{% endif %}{% endfor %}
```

Prompt not from the original task.

```
{{choices.text | join("|||")}}
```

```
If {{ combinedfact[0]|lower }}{{ combinedfact[1:]|trim|trim('.') }}, then
{{ question[0]|lower }}{{question[1:]|trim|trim('?') }}?

Answer choices:
- {{answer_choices | join("\n - ") }}
```

```
{% for choice in choices.label %} {% if choice == answerKey %}{{
answer_choices[loop.index - 1] }}{% endif %}{% endfor %}
```

Prompt not from the original task.

```
Do you think the right answer to the question "{{ question[0]|lower }}{{
question[1:] }}" is "{{ choices.text[1][0]|lower}}{{
choices.text[1][1:]|trim('.') }}", given that
 {{combinedfact[0]|lower}}{{ combinedfact[1:]|trim('.') }}?
```

```
{% if answerKey == choices.label[0] %} Yes {% else %} No {% endif %}
```

```
{{choices.text | join("|||")}}
```

```
Fact 1: {{ fact1[0]|capitalize }}{{ fact1[1:]|trim|trim('.') }}.

Fact 2: {{fact2[0]|capitalize }}{{ fact2[1:]|trim|trim('.') }}.

Given the two facts above, answer the question "{{ question }}" with the
following options:
- {{answer_choices | join("\n - ") }}
```

```
{% for choice in choices.label %} {% if choice == answerKey %}{{
answer_choices[loop.index - 1] }}{% endif %}{% endfor %}
```

1.8.6 QUAIL

Dataset from Rogers et al. (2020). Used in training.

Data Example

Key	Value
answers	['not enough information', 'to visit family', 'par...
context	That fall came and I went back to Michigan and the...
context_id	f001
correct_answer_id	3
domain	fiction
id	f001_0
metadata	{'author': 'Joseph Devon', 'title': 'Black Eyed Su...
question	Why was this character sent away after each school...
question_id	0
question_type	Causality

Prompts

```
{{ context }}
Question: {{ question }}
Options:
```

```
{% for k in range(answers | length) %}
{{'. '.join([answer_choices[k], answers[k]])}}
{% endfor %}
===
The correct answer is
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{answers | join("|||")}}
```

```
{{ context }}
Question: {{ question }}
Options:
- {{ answer_choices | join(" \n - ") }}
===
The correct answer is
```

```
{{ answer_choices[correct_answer_id] }}
```

```
Read the following context and choose the correct option to answer the
question.
Context: {{ context }}
Question: {{ question }}
Options:
{% for k in range(answers | length) %}
{{'. '.join([answer_choices[k], answers[k]])}}
{% endfor %}
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{answers | join("|||")}}
```

```
{{ context }}
{{ question }}
Pick the correct answer from the following options:
- {{ answer_choices | join("\n- ") }}
```

```
{{ answer_choices[correct_answer_id] }}
```

Prompt not from the original task.

```
{{answers | join("|||")}}
```

```
{{ context }}
Question: {{ question }}
===
The answer to the above question is
```

```
{{ answer_choices[correct_answer_id] }}
```

Prompt not from the original task.

```
{{answers | join("|||")}}
```

```
{{ context }}
According to the above context, answer the following question.
{{ question }}
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{ context }}
{{ question }}
Pick the correct answer from the following options:
{% for k in range(answers | length) %}
{{'. '.join([answer_choices[k], answers[k]])}}
{% endfor %}
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{ context }}
{{ question }}
{% for k in range(answers | length) %}
{{'. '.join([answer_choices[k], answers[k]])}}
{% endfor %}
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{ context }}
According to the above context, choose the correct option to answer the
following question.
Question: {{ question }}
Options:
{% for k in range(answers | length) %}
{{'. '.join([answer_choices[k], answers[k]])}}
{% endfor %}
```

```
{{ answer_choices[correct_answer_id] }}
```

Prompt not from the original task.

```
{{answers | join("|||")}}
```

```
Read the following context and answer the question.
Context: {{ context }}
Question: {{ question }}
Answer:
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{answers | join("|||")}}
```

```
{{ context }}
{{ question }}
- {{ answer_choices | join("\n- ") }}
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{answers | join("|||")}}
```

```
{{ context }}
According to the above context, choose the correct option to answer the
following question.
Question: {{ question }}
Options:
- {{ answer_choices | join("\n- ") }}
```

```
{{ answer_choices[correct_answer_id] }}
```

```
{{answers | join("|||")}}
```

```
Read the following context and choose the correct option to answer the
question.
Context: {{ context }}
Question: {{ question }}
Options:
- {{ answer_choices | join("\n- ") }}
```

```
{{ answer_choices[correct_answer_id] }}
```

1.8.7 QUAREL

Dataset from Tafjord et al. (2018). Used in training.

Data Example

Key	Value
id	QuaRel_V1_Fr_0223
answer_index	1
logical_forms	['(infer (speed higher world1) (smoothness higher ...
logical_form_pretty	qrel(speed, higher, world1) -> qrel(smoothness, hi...
world_literals	{'world1': ['ice'], 'world2': ['snow']}
question	Mike was snowboarding on the snow and hit a piece ...

Prompts

Prompt not from the original task.

```
{{world_literals.world1[0]}} ||| {{world_literals.world2[0]}}


Question: {{question}}

Do not use {{"A"}} and {{"B"}} to answer the question but instead, choose
between "{{answer_choices[0]}}" and  "{{answer_choices[1]}}".

{{answer_choices[answer_index]}}
```

Prompt not from the original task.

```
{{world_literals.world1[0]}} ||| {{world_literals.world2[0]}}


Here's a logic test: {{question}}

Choose the answer between "{{answer_choices[0]}}" and
"{{answer_choices[1]}}".

{{answer_choices[answer_index]}}
```

Prompt not from the original task.

```
{{world_literals.world1[0]}} ||| {{world_literals.world2[0]}}


Here's a short story: {{question}}.

What is the most sensical answer between "{{answer_choices[0]}}" and
"{{answer_choices[1]}}"?

{{answer_choices[answer_index]}}
```

Prompt not from the original task.

```
{{world_literals.world1[0]}} ||| {{world_literals.world2[0]}}


Choose between "{{answer_choices[0]}}" and  "{{answer_choices[1]}}".
Question: {{question}}
```

```
{{answer_choices[answer_index]}}
```

Prompt not from the original task.

```
{{world_literals.world1[0]}} ||| {{world_literals.world2[0]}}
```

```
I am testing my students' logic.
What is the answer they should choose between "{{answer_choices[0]}}" and
"{{answer_choices[1]}}"?
Logic test: {{question}}
```

```
{{answer_choices[answer_index]}}
```

1.8.8 QUARTZ

Dataset from Tafjord et al. (”2019”). Used in training.

Data Example

Key	Value
answerKey	A
choices	{'label': ['A', 'B'], 'text': ['scarce', 'plentifu...
id	QRQA-10385-4
para	Many of the worlds people live with water scarcity...
para_anno	{'effect_prop': 'population growth', 'cause_dir_st...
para_id	QRSent-10385
question	John's town used to have lots of water, back when ...
question_anno	{'more_effect_dir': 'several thousand', 'less_effe...

Prompts

```
{{choices.text | join("|||")}}
```

```
Use information from the paragraph to answer the question.

Question:

{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}

Paragraph :

{{ para }}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
{{ para }}
{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!")}} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
Use information from the paragraph to answer the question.

Paragraph :

{{ para }}

Question:

{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
Answer the question based on the following text.

Question:

{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}
```

```
Text:

{{ para }}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
Answer the question below:

{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{  answer_choices | join(" or ") }}{{ "?"
}}
{% endif %}

Assuming that:

{{ para }}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
Read the passage below and choose the right answer to the following
question (choices are {{ answer_choices | join(" or ") }} ):

{{ para }}

{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
{{ para }}

Having read the above passage, choose the right answer to the following
question (choices are {{ answer_choices | join(" or ") }} ):
```

```
{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

```
{{choices.text | join("|||")}}
```

```
Given the fact that:

{{ para }}

Answer the question:

{% if '_____' in question %}
{{ question | trim(".?!") | replace("_____", answer_choices | join(" or
")) }}{{ "?" }}
{% else %}
{{ question | trim(".?!") }} {{ answer_choices | join(" or ") }}{{ "?" }}
{% endif %}
```

```
{{answer_choices[choices.label.index(answerKey)]}}
```

1.8.9 RACE HIGH

Dataset from Lai et al. (2017). Used in evaluation.

Data Example

Key	Value
answer	D
article	Studies show that you may be lied to every day any...
example_id	high10001.txt
options	['harmful', 'easy', 'interesting', 'common']
question	From Para.1 we learn that lying is very _ .

Prompts

Prompt not from the original task.

```
{% set candidate = ["A", "B", "C", "D"] | choice %}
Article: {{article}}
Question: {{question}}
Yes or no, is the answer "{{
[options.0,options.1,options.2,options.3][{"A":0,"B":1,"C":2,"D":3}[answer]]
}}"?
```

```
{% if candidate == answer %}
Yes
{% else %}
No
{% endif %}
```

Prompt not from the original task.

```
Write a multi-choice question for the following article:
Article: {{article}}
```

```
Question:
{{question}}
Options:
{{"A"}} {{options.0}}
{{"B"}} {{options.1}}
{{"C"}} {{options.2}}
{{"D"}} {{options.3}}
Answer:
{{answer}}
```

```
I'm taking a test and have to guess the right answer to the question
after the article.
Article: {{article}}
Question: {{question}}
Options: {{"A"}}: {{options.0}}
{{"B"}}: {{options.1}}
{{"C"}}: {{options.2}}
{{"D"}}: {{options.3}}
```

```
{{answer}}
```

```
Read the article and select the best answer.
Article: {{article}}
Question: {{question}}
Options: {{"A"}}: {{options.0}}
{{"B"}}: {{options.1}}
{{"C"}}: {{options.2}}
{{"D"}}: {{options.3}}
```

```
{{answer}}
```

Prompt not from the original task.

```
Write a multi-choice question for the following article, with the given
choices and answer:
Article: {{article}}
Options:
{{"A"}} {{options.0}}
{{"B"}} {{options.1}}
{{"C"}} {{options.2}}
{{"D"}} {{options.3}}
Answer:
{{answer}} {{
[options.0,options.1,options.2,options.3][{"A":0,"B":1,"C":2,"D":3}[answer]]
}}
Question:
```

```
{{question}}
```

```
{{ options | join("|||") }}
```

```
Read the following article and select the best answer.
Article: {{article}}
Question: {{question}}
- {{answer_choices | join("\n- ")}}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answer]]}}
```

```
{{article}}
{{question}}
{{"A)"}} {{options.0}}
{{"B)"}} {{options.1}}
{{"C)"}} {{options.2}}
{{"D)"}} {{options.3}}
```

```
{{answer}}
```

```
{{ options | join("|||") }}
```

```
Read the following article and answer the question.
Article: {{article}}
Question: {{question}}
Answer:
```

```
{{ answer_choices[{"A":0,"B":1,"C":2,"D":3}[answer]] }}
```

1.8.10 RACE MIDDLE

Dataset from Lai et al. (2017). Used in evaluation.

Data Example

Key	Value
answer	C
article	Take a class at Dulangkou School, and you'll see l...
example_id	middle1.txt
options	['take care of the whole group', 'make sure that e...
question	A discipline leader is supposed to _ .

Prompts

```
Read the article and select the best answer.
Article: {{article}}
Question: {{question}}
Options: {{"A"}}: {{options.0}}
{{"B"}}: {{options.1}}
{{"C"}}: {{options.2}}
{{"D"}}: {{options.3}}
```

```
{{answer}}
```

```
{{ options | join("|||") }}
```

```
Read the following article and answer the question.
Article: {{article}}
Question: {{question}}
Answer:
```

```
{{ answer_choices[{"A":0,"B":1,"C":2,"D":3}[answer]] }}
```

Prompt not from the original task.

```
{% set candidate = ["A", "B", "C", "D"] | choice %}
Article: {{article}}
Question: {{question}}
Yes or no, is the answer "{{
[options.0,options.1,options.2,options.3][{"A":0,"B":1,"C":2,"D":3}[answer]]
}}"?
```

```
{% if candidate == answer %}
Yes
{% else %}
No
{% endif %}
```

```
{{article}}
{{question}}
{{"A)"}} {{options.0}}
{{"B)"}} {{options.1}}
{{"C)"}} {{options.2}}
{{"D)"}} {{options.3}}
```

```
{{answer}}
```

```
{{ options | join("|||") }}
```

```
Read the following article and select the best answer.
Article: {{article}}
Question: {{question}}
- {{answer_choices | join("\n- ")}}
```

```
{{answer_choices[{"A":0,"B":1,"C":2,"D":3}[answer]]}}
```

Prompt not from the original task.

```
Write a multi-choice question for the following article, with the given
choices and answer:
Article: {{article}}
Options:
{{"A"}} {{options.0}}
{{"B"}} {{options.1}}
{{"C"}} {{options.2}}
{{"D"}} {{options.3}}
Answer:
{{answer}} {{
[options.0,options.1,options.2,options.3][{"A":0,"B":1,"C":2,"D":3}[answer]]
}}
Question:
```

```
{{question}}
```

Prompt not from the original task.

```
Write a multi-choice question for the following article:
Article: {{article}}
```

```
Question:
{{question}}
Options:
{{"A"}} {{options.0}}
{{"B"}} {{options.1}}
{{"C"}} {{options.2}}
{{"D"}} {{options.3}}
Answer:
{{answer}}
```

```
I'm taking a test and have to guess the right answer to the question
after the article.
Article: {{article}}
Question: {{question}}
Options: {{"A"}}: {{options.0}}
{{"B"}}: {{options.1}}
{{"C"}}: {{options.2}}
{{"D"}}: {{options.3}}
```

```
{{answer}}
```

1.8.11 SCIQ

Dataset from Johannes Welbl (2017). Used in training.

Key	Value
question	What type of organism is commonly used in preparat...
distractor3	viruses
distractor1	protozoa
distractor2	gymnosperms
correct_answer	mesophilic organisms
support	Mesophiles grow best in moderate temperature, typi...

Data Example

Prompts

```
{{distractor1}} ||| {{distractor2}} ||| {{distractor3}} |||
{{correct_answer}}
```

```
Q: {{question}}


A:
```

```
{{answer_choices[3]}}
```

Prompt not from the original task.

```
{{distractor1}} ||| {{distractor2}} ||| {{distractor3}} |||
{{correct_answer}}
```

```
{% set order = [[0, 1, 2, 3], [0, 1, 3, 2], [0, 2, 1, 3], [0, 2, 3, 1],
[0, 3, 1, 2], [0, 3, 2, 1],
                              [1, 0, 2, 3], [1, 0, 3, 2], [1, 2, 0, 3],
                              [1, 2, 3, 0], [1, 3, 0, 2], [1, 3, 2, 0],
                              [2, 1, 0, 3], [2, 1, 0, 2], [2, 0, 1, 3],
                              [2, 0, 3, 1], [2, 3, 1, 0], [2, 3, 0, 1],
                              [3, 1, 2, 0], [3, 1, 0, 2], [3, 2, 1, 0],
                              [3, 2, 0, 1], [3, 0, 1, 2], [3, 0, 2, 1]] |
                             choice %}
Q: {{question}}


 Choices:

- {{ answer_choices[order[0]] }}

- {{ answer_choices[order[1]] }}

- {{ answer_choices[order[2]] }}

- {{ answer_choices[order[3]] }}

A:
```

```
{{answer_choices[3]}}
```

```
{{distractor1}} ||| {{distractor2}} ||| {{distractor3}} |||
{{correct_answer}}
```

```
{% set order = [[0, 1, 2, 3], [0, 1, 3, 2], [0, 2, 1, 3], [0, 2, 3, 1],
[0, 3, 1, 2], [0, 3, 2, 1],
                                [1, 0, 2, 3], [1, 0, 3, 2], [1, 2, 0, 3],
                                [1, 2, 3, 0], [1, 3, 0, 2], [1, 3, 2, 0],
                                [2, 1, 0, 3], [2, 1, 0, 2], [2, 0, 1, 3],
                                [2, 0, 3, 1], [2, 3, 1, 0], [2, 3, 0, 1],
                                [3, 1, 2, 0], [3, 1, 0, 2], [3, 2, 1, 0],
                                [3, 2, 0, 1], [3, 0, 1, 2], [3, 0, 2, 1]] |
                                choice %}
Q: {{question}}


Read this paragraph and choose the correct option from the provided
answers:

{{support}}

 Choices:

- {{ answer_choices[order[0]] }}

- {{ answer_choices[order[1]] }}

- {{ answer_choices[order[2]] }}

- {{ answer_choices[order[3]] }}


A:
```

```
{{answer_choices[3]}}
```

```
{{distractor1}} ||| {{distractor2}} ||| {{distractor3}} |||
{{correct_answer}}
```

```
{% set order = [[0, 1, 2, 3], [0, 1, 3, 2], [0, 2, 1, 3], [0, 2, 3, 1],
[0, 3, 1, 2], [0, 3, 2, 1],
                                [1, 0, 2, 3], [1, 0, 3, 2], [1, 2, 0, 3],
                                [1, 2, 3, 0], [1, 3, 0, 2], [1, 3, 2, 0],
                                [2, 1, 0, 3], [2, 1, 0, 2], [2, 0, 1, 3],
                                [2, 0, 3, 1], [2, 3, 1, 0], [2, 3, 0, 1],
                                [3, 1, 2, 0], [3, 1, 0, 2], [3, 2, 1, 0],
                                [3, 2, 0, 1], [3, 0, 1, 2], [3, 0, 2, 1]] |
                                choice %}
Answer the following question given this paragraph:

{{support}}


Q: {{question}}
```

```
 Choices:

- {{ answer_choices[order[0]] }}

- {{ answer_choices[order[1]] }}

- {{ answer_choices[order[2]] }}

- {{ answer_choices[order[3]] }}

A:
```

```
{{answer_choices[3]}}
```

```
{{distractor1}} ||| {{distractor2}} ||| {{distractor3}} |||
{{correct_answer}}
```

```
Answer the following question given this paragraph:

{{support}}

Q: {{question}}

A:
```

```
{{answer_choices[3]}}
```

1.8.12 SOCIAL_I_QA

Data Example

Key	Value
answerA	like attending
answerB	like staying home
answerC	a good friend to have
context	Cameron decided to have a barbecue and gathered he...
label	1
question	How would Others feel as a result?

Prompts

```
{{answerA}} ||| {{answerB}} ||| {{answerC}}
```

```
I heard that {{context}}

And I was wondering {{question}}
```

```
{{answer_choices[label | int - 1]}}
```

```
{{answerA}} ||| {{answerB}} ||| {{answerC}}
```

```
{{context}}

Given the context: {{question}}

Possible answers: {{answer_choices | join(", ")}}
```

```
{{answer_choices[label | int - 1]}}
```

```
{% set random_answer_id = range(0,2) | choice%}
{% set answers = [answerA, answerB, answerC] %}
{{context}}

Given the question "{{question}}", is "{{answers[random_answer_id]}}" a
valid answer?
```

```
{% if (label | int) - 1 == random_answer_id %}
    Yes
{% else %}
    No
{% endif %}
```

Prompt not from the original task.

```
{{context}}

Given that the answer to a question is "{{{"1": answerA, "2": answerB,
"3": answerC}[label]}}", what is the question?
```

```
{{question}}
```

```
{{answerA}} ||| {{answerB}} ||| {{answerC}}
```

```
{{context}}

Given the context: {{question}}
```

```
{{answer_choices[label | int - 1]}}
```

```
Context: {{context}}

Question: {{question}}

Which one of these answers best answers the question according to the
context?

A: {{answerA}}

B: {{answerB}}

C: {{answerC}}
```

```
{{{"1": "A", "2": "B", "3": "C"}[label]}}
```

1.8.13 SUPER_GLUE BOOLQ

Dataset from Clark et al. (2019). Used in evaluation.

Data Example

Key	Value
idx	0
label	1
passage	Persian language -- Persian , al...
question	do iran and afghanistan speak the same language

Prompts

```
False ||| True
```

```
{% if label != -1 %}
Passage: {{passage}}

After reading this passage, I have a question: {{question}}? True or
False?
```

```
{{answer_choices[label]}}
{% endif %}
```

Prompt from Brown et al. (2020)

```
{% if label != -1 %}
{{ passage }}
Question: {{ question }}
Answer:
```

```
{{ answer_choices[label] }}
{% endif %}
```

```
{% if label != -1 %}
{{ passage }}

Having read that, I wonder {{ question }}?

{{ answer_choices[label] }}
{% endif %}
```

```
No ||| Yes
```

```
{% if label != -1 %}
Text: {{passage}}

Answer the following yes/no question: {{question}}? Yes or no?

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1 %}
{{ passage }}

Having read that, could you tell me {{ question }}?

{{ answer_choices[label] }}
{% endif %}
```

```
No ||| Yes
```

```
{% if label != -1 %}
EXAM
1. Answer by yes or no.

Document: {{passage}}
Question: {{question}}?

{{answer_choices[label]}}
{% endif %}
```

Prompt from Schick and Schütze (2021)

```
{% if label != -1 %}
Based on the following passage, {{ question }}? {{ passage }}

{{ answer_choices[label] }}
{% endif %}
```

```
False ||| True
```

```
{% if label != -1 %}
Exercise: read the text and answer the question by True or False.

Text: {{passage}}
Question: {{question}}?
```

```
{{answer_choices[label]}}
{% endif %}
```

Prompt from Schick and Schütze (2021)

```
{% if label != -1 %}
{{ passage }}
Based on the previous passage, {{ question }}?
```

```
{{ answer_choices[label] }}
{% endif %}
```

```
False ||| True
```

```
{% if label != -1 %}
{{passage}}

Q: {{question}}? True or False?
```

```
{{answer_choices[label]}}
{% endif %}
```

1.8.14 SUPER_GLUE COPA

Dataset from Roemmele et al. (2011). Used in evaluation.

Data Example

Key	Value
choice1	The sun was rising.
choice2	The grass was cut.
idx	0
label	0
premise	My body cast a shadow over the grass.
question	cause

Prompts

```
{{choice1}} ||| {{choice2}}
```

```
Exercise: choose the most plausible alternative.

{{ premise }} {% if question == "cause" %} because... {% else %} so... {%
endif %}
- {{choice1}}
- {{choice2}}
```

```
{{ answer_choices[label] }}
```

```
{{choice1}} ||| {{choice2}}
```

```
{% if question == "effect" %}
{{ premise }} What could happen next, "{{ answer_choices[0] }}" or "{{
answer_choices[1] }}"?
```

```
{{ answer_choices[label] }}
{% endif %}
```

```
{{choice1}} ||| {{choice2}}
```

```
{{ premise }}

I am hesitating between two options. Help me choose the more likely {% if
question == "cause" %} cause: {% else %} effect: {% endif %}
- {{choice1}}
- {{choice2}}
```

```
{{ answer_choices[label] }}
```

```
{{choice1}} ||| {{choice2}}
```

```
{{ premise }} {% if question == "cause" %} This happened because... {%
else %} As a consequence... {% endif %}
Help me pick the more plausible option:
- {{choice1}}
- {{choice2}}
```

```
{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{choice1 }} ||| {{choice2}}
```

```
"{{ answer_choices[0] }}" or "{{ answer_choices[1] }}"? {{ premise }} {%
if question == "cause" %} because {% else %} so {% endif %}

{{ answer_choices[label] }}
```

```
{{choice1}} ||| {{choice2}}
```

```
{% if question == "effect" %}
{{ premise }} As a result, "{{ answer_choices[0] }}" or "{{
answer_choices[1] }}"?

{{ answer_choices[label] }}
{% endif %}
```

```
{{choice1}} ||| {{choice2}}
```

```
{{ premise }}

What's the best option?
- {{choice1}}
- {{choice2}}

We are looking for {% if question == "cause" %} a cause {% else %} an
effect {% endif %}

{{answer_choices[label]}}
```

```
{{choice1}} ||| {{choice2}}
```

```
{% if question == "cause" %}
{{ premise }} Which may be caused by "{{ answer_choices[0] }}" or "{{
answer_choices[1] }}"?

{{ answer_choices[label] }}
{% endif %}
```

```
{{choice1}} ||| {{choice2}}
```

```
Pick the more likely continuation to the following sentence:
{{ premise }} {% if question == "cause" %} as a result of: {% else %} as
a consequence: {% endif %}
- {{choice1}}
- {{choice2}}
```

```
{{ answer_choices[label] }}
```

```
{{choice1}} ||| {{choice2}}
```

```
{{ premise }}

Select the most plausible {% if question == "cause" %} cause: {% else %}
effect: {% endif %}
- {{choice1}}
- {{choice2}}
```

```
{{ answer_choices[label] }}
```

```
{{choice1}} ||| {{choice2}}
```

```
{% if question == "cause" %}
{{ premise }} Why? "{{ answer_choices[0] }}" or "{{ answer_choices[1]
}}"?
```

```
{{ answer_choices[label] }}
{% endif %}
```

```
{{choice1}} ||| {{choice2}}
```

```
{{ premise }} {% if question == "cause" %} because... {% else %} so... {%
endif %}
Choose between:
- {{choice1}}
- {{choice2}}
```

```
{{ answer_choices[label] }}
```

1.8.15 SUPER_GLUE MULTIRC

Dataset from Khashabi et al. (2018). Used in evaluation.

Data Example

Key	Value
answer	Children, Gerd, or Dorian Popa
idx	{'paragraph': 0, 'question': 0, 'answer': 0}
label	0
paragraph	While this process moved along, diplomacy continue...
question	What did the high-level effort to persuade Pakista...

Prompts

```
No ||| Yes
```

```
{{paragraph}}

Question: {{question}}
I found this answer "{{answer}}". Is that correct? Yes or no?
```

```
{{answer_choices[label]}}
```

Prompt from Schick and Schütze (2021)

```
{{ paragraph }}
Based on the previous passage, {{ question }}
Is "{{ answer }}" a correct answer?
```

```
{{ answer_choices[label] }}
```

```
No ||| Yes
```

```
{{paragraph}}
Question: {{question}}

I am grading my students' exercises. Is the answer "{{answer}}" correct?
```

```
{{answer_choices[label]}}
```

```
{{ paragraph }}
{{ question }}
Would it be good to answer "{{ answer }}"?
```

```
{{ answer_choices[label] }}
```

Prompt from Schick and Schütze (2021)

```
{{ paragraph }}
Question: {{ question }}
Is it {{ answer }}?
```

```
{{ answer_choices[label] }}
```

```
No ||| Yes
```

```
{{paragraph}}

Decide whether "{{answer}}" is a valid answer to the following question:
{{question}}
Answer yes or no.
```

```
{{answer_choices[label]}}
```

Prompt from Schick and Schütze (2021)

```
{{ paragraph }}
Question: {{ question }}
Is the correct answer {{ answer }}?
```

```
{{ answer_choices[label] }}
```

```
No ||| Yes
```

```
Is "{{answer}}" a correct answer to the following question?
Question: {{question}}

Rely on the following text: {{paragraph}}
```

```
{{answer_choices[label]}}
```

```
No ||| Yes
```

```
{{paragraph}}

Question: {{question}}
I think "{{answer}}" is a valid answer. Could you confirm? Yes or no?
```

```
{{answer_choices[label]}}
```

```
{{ paragraph }}
{{ question }}
I was going to say "{{ answer }}". Does that sound right?
```

```
{{ answer_choices[label] }}
```

1.8.16 WIKI_HOP ORIGINAL

Dataset from Welbl et al. (2018). Used in training.

Data Example

Key	Value
annotations	[]
answer	1996 summer olympics
candidates	['1996 summer olympics', 'olympic games', 'sport']
id	WH_train_0
question	participant_of juan rossell
supports	['The 2004 Summer Olympic Games, officially known ...

Prompts

```
{{candidates | join("|||")}}
```

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
What object entity has the relation of '{{ question_split[0] |
replace("_", " ")}}' with the subject '{{ question_split[1:] | join("
")}}'?

Choices:
- {{answer_choices | join("\n - ") }}
```

```
{{answer}}
```

Prompt not from the original task.

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
What is the relationship between '{{ question_split[1:] | join(" ")}}'
and '{{answer}}'?
```

```
{{ question_split[0] | replace("_", " ") }}
```

Prompt not from the original task.

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
What entity does '{{ question_split[1:] | join(" ")}}' has the relation
'{{ question_split[0] | replace("_", " ") }}' with?
```

```
{{answer}}
```

Prompt not from the original task.

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
Given the paragraphs above, decide what entity has the relation '{{
question_split[0] | replace("_", " ") }}' with '{{answer}}'.
```

```
{{ question_split[1:] | join(" ")}}
```

```
{{candidates | join("|||")}}
```

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
Given the information above, choose from the list below the object entity
that exhibits the relation '{{ question_split[0] | replace("_", " ")}}'
with the subject '{{ question_split[1:] | join(" ")}}'.

Choices:
- {{answer_choices | join("\n - ") }}
```

```
{{answer}}
```

```
{{candidates | join("|||")}}
```

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
After reading the paragraphs above, we are interested in knowing the
entity with which '{{ question_split[1:] | join(" ")}}' exhibits the
relationship of '{{ question_split[0] | replace("_", " ")}}'. Find the
answer from the choices below.

Choices:
- {{answer_choices | join("\n - ") }}
```

```
{{answer}}
```

Prompt not from the original task.

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
Given the information, choose the subject and object entities that have
the relation of '{{ question_split[0] | replace("_", " ") }}'.
```

```
{{ question_split[1:] | join(" ") }} , {{answer}}
```

```
{{candidates | join("|||")}}
```

```
Information:
{% for support in supports %}
- {{ support }}
{% endfor %}

{% set question_split = question.split(' ') %}
After reading the paragraphs above, choose the best answer for the entity
that related to '{{ question_split[1:] | join(" ")}}' with the
relationship of '{{ question_split[0] | replace("_", " ")}}'.

Choices:
- {{answer_choices | join("\n - ") }}
```

```
{{answer}}
```

```
{{candidates | join("|||")}}
```

```
Information:
{% for support in supports %}
- {{ support }}
```

```
{% endfor %}

{% set question_split = question.split(' ') %}
'{{ question_split[1:] | join(" ")}}' is related to which object entity
through the relation of '{{ question_split[0] | replace("_", " ")}}'?

Choices:
- {{answer_choices | join("\n - ") }}
```

```
{{answer}}
```

1.8.17 WIQA

Dataset from Tandon et al. (2019). Used in training.

Data Example

Key	Value
answer_label	more
answer_label_as_choice	A
choices	{'label': ['A', 'B', 'C'], 'text': ['more', 'less'...
metadata_graph_id	144
metadata_para_id	1217
metadata_path_len	2
metadata_question_id	influence_graph:1217:144:106#0
metadata_question_type	INPARA_EFFECT
question_para_step	['A tree produces seeds', 'The seeds are dispersed...
question_stem	suppose there will be fewer new trees happens, how...

Prompts

Prompt not from the original task.

```
-  {{ question_para_step[1:] | join("\n- ") }}

What might be the first step of the process?
```

```
{{ question_para_step | first }}
```

Prompt not from the original task.

```
{% set process_list = question_para_step[:-1] if question_para_step[-1]
== "" else question_para_step %}
-  {{ process_list[:-1] | join("\n- ") }}

What might be the last step of the process?
```

```
{{ process_list | last }}
```

Prompt not from the original task.

```
What is the missing first step of the following process:

-  {{ question_para_step[1:] | join("\n- ") }}
```

```
{{ question_para_step | first }}
```

Prompt not from the original task.

```
{% set process_list = question_para_step[:-1] if question_para_step[-1]
== "" else question_para_step %}
What is the final step of the following process:
-  {{ process_list[:-1] | join("\n- ") }}
```

```
{{ process_list | last }}
```

```
Process:
- {{ question_para_step | join("\n- ")}}

Question:
{{question_stem}}

How does the supposed perturbation influence the second effect mentioned.
Answer by {{"more, less or no effect"}}
```

```
{{answer_label|replace("_", " ")}}
```

Prompt not from the original task.

```
Process:

- {{ question_para_step | join("\n- ") }}

{{question_stem}}

Which of the following is the supposed perturbation?

- {{"directly impacting a step of the process"}}
- {{"indirectly impacting a step of the process"}}
- {{"not impacting any step of the process"}}
```

```
{{{"EXOGENOUS_EFFECT": "indirectly impacting a step of the process",
"OUTOFPARA_DISTRACTOR": "not impacting any step of the process",
"INPARA_EFFECT": "directly impacting a step of the
process"}[metadata_question_type]}}
```

```
Process:
- {{ question_para_step | join("\n- ")}}

Question:
{{question_stem}}

- {{"A: more"}}
- {{"B: less"}}
- {{"C: no effect"}}
```

```
{{answer_label_as_choice}}
```

Prompt not from the original task.

```
Process:

- {{ question_para_step | join("\n- ") }}

Perturbation hypothesis:
{{question_stem}}

Does the supposed perturbation have an effect (direct or indirect) on the
process?
```

```
{{{"EXOGENOUS_EFFECT": "yes", "OUTOFPARA_DISTRACTOR": "no",
"INPARA_EFFECT": "yes"}[metadata_question_type]}}
```

1.8.18 CIRCA

Dataset from **?**. Used in evaluation.

Data Example

Key	Value
answer-Y	I'm a veterinary technician.
canquestion-X	I am employed .
context	Y has just travelled from a different city to meet...
goldstandard1	0
goldstandard2	0
judgements	Yes#Yes#Yes#Yes#Yes
question-X	Are you employed?

Prompts

Prompt not from the original task.

```
Convert this question to a sentence declarative sentence asserting an
affirmative answer:

{{question_X}}
```

```
{{canquestion_X}}
```

```
{% if goldstandard2 != -1 %}

Given the question-answer pair of X and Y in the context of {{context}},
which of the following answers is Y implying: "{{"Yes"}}", "{{"No"}}",
"{{"In the middle, neither yes nor no"}}", "{{"Probably yes / sometimes
yes"}}", "{{"Probably no"}}", "{{"Yes, subject to some conditions"}}",
"{{"Other"}}" or "{{"I am not sure how X will interpret Y's answer"}}" ?

X: {{question_X}}

Y: {{answer_Y}}
```

```
{{   answer_choices[goldstandard2]}}

{% endif %}
```

Prompt not from the original task.

```
What is a possible question X could ask Y given the context of
{{context}} that would cause Y to answer "{{answer_Y}}"?
```

```
{{question_X}}
```

```
{% if goldstandard1 != -1 %}

Given the question-answer pair of X and Y in the context of {{context}},
what answer is Y implying?

X: {{question_X}}

Y: {{answer_Y}}
```

```
{{   answer_choices[goldstandard1]}}

{% endif %}
```

```
{% if goldstandard1 != -1 %}

Given the question-answer pair of X and Y in the context of {{context}},
which of the following answers is Y implying: "{{"Yes"}}", "{{"No"}}",
"{{"In the middle, neither yes nor no"}}", "{{"Probably yes / sometimes
yes"}}", "{{"Probably no"}}", "{{"Yes, subject to some conditions"}}",
"{{"Other"}}" or "{{"I am not sure how X will interpret Y's answer"}}" ?

X: {{question_X}}

Y: {{answer_Y}}
```

```
{{   answer_choices[goldstandard1]}}

{% endif %}
```

1.8.19 MC_TACO

Dataset from Ben Zhou and Roth (2019). Used in evaluation.

Data Example

Key	Value
answer	she was ill for 30 seconds
category	0
label	0
question	How long was his mother ill?
sentence	Durer's father died in 1502, and his mother died i...

Prompts

```
Given the context,

{{sentence}}

observe the following QA pair and check if the answer is plausible:

Question: {{question}}

Answer: {{answer}}
```

```
{{answer_choices[label]}}
```

```
I've been grappling with the temporal accuracy of this answer for a
while:

Q: "{{question}}"

I have the following information: "{{sentence}}"

A: "{{answer}}"

This answer is definitely not
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
There are five temporal categories: {{"Event Duration"}}, {{"Event
Ordering"}}, {{"Frequency"}}, {{"Typical Time"}}, {{"Stationarity"}}.

Out of the above temporal categories, which one does the question
"{{question}}" belong to?
```

```
{{answer_choices[category]}}
```

Prompt not from the original task.

```
{% if label %}
I have the following passage:

{{sentence}}

My query is: "{{question}}"

I want an answer that is "temporally plausible".
```

```
{{answer}}
{% endif %}
```

```
Here's what happened: {{sentence}}

I asked my friend {{question}}

and they said {{answer}}

Should I believe them?
```

```
{{answer_choices[label]}}
```

```
Given the context, the question, and the candidate answer, the task is to
determine whether the candidate answer is plausible ("yes") or not
("no").

Context: {{sentence}}

Question: {{question}}

Candidate answer: {{answer}}
```

```
{{answer_choices[label]}}
```

```
Given the context,

{{sentence}}

and the question,

{{question}}

is the following answer believable?

{{answer}}
```

```
{{answer_choices[label]}}
```

```
True/False?

"{{answer}}" is a plausible answer to "{{question}}", given
"{{sentence}}"
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
Which temporal category does the question "{{question}}" belong to?
```

```
{{answer_choices[category]}}
```

```
Here's what happened: {{sentence}}

I asked my friend {{question}}

and they said {{answer}}

Should I doubt them?
```

```
{{answer_choices[label]}}
```

1.8.20 PIQA

Dataset from Bisk et al. (2020). Used in evaluation.

Data Example

Key	Value
goal	When boiling butter, when it's ready, you can
label	1
sol1	Pour it onto a plate
sol2	Pour it into a jar

Prompts

```
{{sol1}} ||| {{sol2}}
```

```
Goal: {{goal}}

Which is the correct ending?
- {{sol1}}
- {{sol2}}

Answer:
```

```
{{answer_choices[label]}}
```

```
{{sol1}} ||| {{sol2}}
```

```
{{"Solution 1"}}: {{sol1}}
{{"Solution 2"}}: {{sol2}}

Goal: {{goal}}

Given the goal, what is the correct solution?

Answer by copying the correct solution
```

```
{{answer_choices[label]}}
```

```
Sentence: {{goal}}

Choice {{answer_choices[0]}}: {{sol1}}
```

```
Choice {{answer_choices[1]}}: {{sol2}}

What is the index of the correct choice for ending for the sentence?

Answer:
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
Given a goal and a wrong solution, rewrite it to give a correct solution.
Goal: {{goal}}
Solution: {{[sol1, sol2][1 - label]}}
Corrected solution:
```

```
{{[sol1, sol2][label]}}
```

```
{{sol1}} ||| {{sol2}}
```

```
Finish the following sentence with the best choice: {{goal}}

Choices:
- {{sol1}}
- {{sol2}}

Answer:
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
{{goal}} {{sol2}}
Does this phrase make sense?
```

```
{{answer_choices[label]}}
```

```
Given a goal and 2 solutions, choose the most appropriate solution.
Goal: {{goal}}
- {{"Solution 1"}}: {{sol1}}
- {{"Solution 2"}}: {{sol2}}

Answer by returning either {{"Solution 1"}} or {{"Solution 2"}}
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
Given a sentence, correct it if it doesn't make sense. If it makes sense,
just return it as the answer.
Input: {{goal}} {{sol2[0].lower() + sol2[1:]}}
Output:
```

```
{{goal}} {{[sol1[0].lower() + sol1[1:], sol2[0].lower() +
sol2[1:]][label]}}
```

Prompt not from the original task.

```
{{goal}}
```

```
{{[sol1[0].lower() + sol1[1:], sol2[0].lower() + sol2[1:]][label]}}
```

Prompt not from the original task.

```
Does this phrase make sense?
{{goal}} {{sol1[0].lower() + sol1[1:]}}
Answer with {{answer_choices[0]}} or {{answer_choices[1]}}
```

```
{{answer_choices[label]}}
```

Prompt not from the original task.

```
Sentence: {{goal}} {{sol1[0].lower() + sol1[1:]}}
If the sentence does not make sense, correct it so that it does make
sense. Otherwise, just copy it.
Answer:
```

```
{{goal}} {{[sol1[0].lower() + sol1[1:], sol2[0].lower() +
sol2[1:]][label]}}
```

1.9 SENTIMENT

1.9.1 AMAZON_POLARITY

Dataset from McAuley and Leskovec (2013). Used in training.

Data Example

Key	Value
content	This sound track was beautiful! It paints the sene...
label	1
title	Stuning even for the non-gamer

Prompts

```
Title: {{title}}
Review: {{content}}
Is the review positive or negative?
```

```
{{answer_choices[label]}}
```

```
Based on this review, would the user recommend this product?
===
Review: {{content}}
Answer:
```

```
{{answer_choices[label]}}
```

```
Is this product review positive?
Title: {{title}}
Review: {{content}}
Answer:
```

```
{{answer_choices[label]}}
```

```
Title: {{title}}
Review: {{content}}
Is this product review negative?
```

```
{{answer_choices[label]}}
```

```
Title: {{title}}
Review: {{content}}
Does this product review convey a negative or positive sentiment?
```

```
{{answer_choices[label]}}
```

```
Is there a negative or positive tone to this product review?
===
Title: {{title}}
Review: {{content}}
Answer:
```

```
{{answer_choices[label]}}
```

```
Title: {{title}}
Product review: {{content}}
Would you say this review depicts the product in a {{answer_choices[1]}}
or {{answer_choices[0]}} light?
```

```
{{answer_choices[label]}}
```

```
You are considering whether to buy a product. You look at the reviews.
Would the following review {{answer_choices[0]}} or {{answer_choices[1]}}
the chances of you buying the product?
Review title: {{title}}
Product review: {{content}}
```

```
{{answer_choices[label]}}
```

```
Here is a review left by a customer on a product. Would you say he was
{{answer_choices[1]}} or {{answer_choices[0]}}?
Title: {{title}}
Review: {{content}}
```

```
{{answer_choices[label]}}
```

1.9.2 APP_REVIEWS

Dataset from Zurich Open Repository and Archive. Used in training.

Data Example

Key	Value
date	October 12 2016
package_name	com.mantz_it.rfanalyzer
review	Great app! The new version now works on my Bravia ...
star	4

Prompts

Prompt not from the original task.

```
Given this review: "{{review}}"
Would you recommend this app to a friend? {{answer_choices[0]}},
{{answer_choices[1]}}, {{answer_choices[2]}}, {{answer_choices[3]}}, or
{{answer_choices[4]}}?
```

```
{{answer_choices[star-1]}}
```

Prompt not from the original task.

```
Generate a {{star}}-star review (1 being lowest and 5 being highest)
about an app with package {{package_name}}.
```

```
{{review}}
```

Prompt not from the original task.

```
What would be the *-rating of this review (* being the lowest and *****
being the highest)? "{{review}}"
```

```
{{answer_choices[star-1]}}
```

Prompt not from the original task.

```
On a scale of 1-5 (with 1 being least favorable and 5 being most
favorable), how would you rate this review? "{{review}}"
```

```
{{star}}
```

1.9.3 IMDB

Dataset from Maas et al. (2011). Used in training.

Data Example

Key	Value
text	Bromwell High is a cartoon comedy. It ran at the s...
label	1

Prompts

```
The following movie review expresses what sentiment? {{text}}
```

```
{{ answer_choices [label] }}
```

```
{{text}} Did the reviewer find this movie {{"good or bad"}}?
```

```
{{ answer_choices [label] }}
```

```
{{text}}
Is this review {{"positive or negative"}}?
```

```
{{answer_choices[label] }}
```

```
{{text}} How does the viewer feel about the movie?
```

```
{{ answer_choices [label] }}
```

```
{{text}} What sentiment does the writer express for the movie?
```

```
{{ answer_choices [label] }}
```

```
{{text}} The sentiment expressed for the movie is
```

```
{{ answer_choices [label] }}
```

```
{{text}} What is the sentiment expressed in this text?
```

```
{{ answer_choices [label] }}
```

Prompt not from the original task.

```
{{text}} This is definitely not a
{{ answer_choices [1-label]}} review.
```

```
{{text}} Did the reviewer enjoy the movie?
{{ answer_choices [label] }}
```

```
{{text}} What is the sentiment expressed by the reviewer for the movie?
{{ answer_choices [label] }}
```

```
{{text}} How does the reviewer feel about the movie?
{{ answer_choices [label] }}
```

1.9.4 ROTTEN_TOMATOES

Dataset from Pang and Lee (2005). Used in training.

Data Example

Key	Value
text	the rock is destined to be the 21st century's new ...
label	1

Prompts

```
{{text}} Did the reviewer find this movie {{"good or bad"}}?
{{ answer_choices [label] }}
```

```
{{text}} What is the sentiment expressed in this text?
{{ answer_choices [label] }}
```

```
{{text}}
Is this review {{"positive or negative"}}?
```

```
{{answer_choices[label] }}
```

```
{{text}} Did the reviewer enjoy the movie?
{{ answer_choices [label] }}
```

```
{{text}} How does the reviewer feel about the movie?
{{ answer_choices [label] }}
```

```
{{text}} The sentiment expressed for the movie is
{{ answer_choices [label] }}
```

```
{{text}} What sentiment does the writer express for the movie?
{{ answer_choices [label] }}
```

```
The following movie review expresses what sentiment? {{text}}
{{ answer_choices [label] }}
```

```
{{text}} What is the sentiment expressed by the reviewer for the movie?
{{ answer_choices [label] }}
```

```
{{text}} How does the viewer feel about the movie?
{{ answer_choices [label] }}
```

1.9.5 YELP_REVIEW_FULL

Dataset from Zhang et al. (2015a). Used in training.

Data Example

Prompts

```
{{ text }}
So I would like to give it
```

Key	Value
label	4
text	dr. goldberg offers everything i look for in a gen...

```
{{ answer_choices[label] }}
```

```
{{ text }}
===
Based on that, my rating is

{{ answer_choices[label] }}
```

```
Review text:
{{ text }}

Stars:

{{ answer_choices[label] }}
```

```
{{ text }} My rating for this place is

{{ answer_choices[label] }}
```

```
Review text:
{{ text }}

Review score (between 1 and 5):

{{ answer_choices[label] }}
```

```
Review: {{text}}
On a scale of 1 to 5, I would give this product

{{ answer_choices[label] }}
```

```
Review text:
{{ text }}

Review rating:

{{ answer_choices[label] }}
```

1.10 STORY COMPLETION

1.10.1 HELLASWAG

Dataset from Zellers et al. (2019). Used in evaluation.

Data Example

Key	Value
activity_label	Removing ice from car
ctx	Then, the man writes over the snow covering the wi...
ctx_a	Then, the man writes over the snow covering the wi...
ctx_b	then
endings	[', the man adds wax to the windshield and cuts it...
ind	4
label	3
source_id	activitynet~v_-1IBHYS3L-Y
split	train
split_type	indomain

Prompts

```
{{endings | join(" ||| ")}}
```

```
Complete the description with an appropriate ending:
First, {{ ctx_a.lower() }} Then, {{ ctx_b.lower() }} ...

(a) {{ answer_choices[0] }}

(b) {{ answer_choices[1] }}

(c) {{ answer_choices[2] }}

(d) {{ answer_choices[3] }}
```

```
{{ answer_choices[label | int()] }}
```

Prompt not from the original task.

```
What is the topic of the sentence: {{ctx}}
```

```
{{activity_label}}
```

Prompt not from the original task.

```
{{endings | join(" ||| ")}}
```

```
Complete the sentence: {{ctx}}
```

```
{{answer_choices[label | int()]}}
```

Prompt not from the original task.

```
{{ctx}} {{endings[label | int()]}}
Can you identify the topic of the paragraph?
```

```
{{activity_label}}
```

```
{{endings | join(" ||| ") }}
```

```
{% set prompts = [
'Can you pick the correct ending for the sentence: ',
'The task is to generate the ending for the sentence: ',
'How does this sentence end? ',
'From the list of endings described below, what ending makes the most
sense for the sentence ',]
%}
{{prompts | choice}}
{{ctx}}

(a)  {{answer_choices[0]}}

(b)  {{answer_choices[1]}}

(c)  {{answer_choices[2]}}

(d)  {{answer_choices[3]}}
```

```
{{answer_choices [label | int()]}}
```

Prompt not from the original task.

```
{% set instance = [0, 1, 2, 3] | choice %}
Consider the following description: {{ ctx_a }}
Is the following an appropriate continuation?
{{ ctx_b }} {{ endings[instance] }}
Yes or No?
```

```
{% if label  == instance | string() %}
{{answer_choices[0]}}
{% else %}
{{answer_choices[1]}}
{% endif %}
```

```
{{endings | join("|||")}}
```

```
How does this sentence end?
{{ctx}}

(a)  {{answer_choices[0]}}

(b)  {{answer_choices[1]}}

(c)  {{answer_choices[2]}}
```

```
(d)  {{answer_choices[3]}}

Hint: the topic of the sentence is {{activity_label}}
```

```
{{answer_choices [label | int()]}}
```

Prompt not from the original task.

```
How would you start the sentence:
{{endings[label | int()]}}
```

```
{{ctx}}
```

Prompt not from the original task.

```
{% set instance = [0, 1, 2, 3] | choice %}
Consider the following text: {{ ctx_b }} {{ endings[instance] }}
Is it an appropriate continuation of the following text:
{{ ctx_a }} ?
Yes or No?
```

```
{% if label  == instance | string() %}
{{answer_choices[0]}}
{% else %}
{{answer_choices[1]}}
{% endif %}
```

Prompt not from the original task.

```
{{ ctx }}...
How does the description likely end?

Ending 1: {{ endings[0] }}

Ending 2: {{ endings[1] }}

Ending 3: {{ endings[2] }}

Ending 4: {{ endings[3] }}
```

```
{{ answer_choices[label | int()] }}
```

```
If a description of a situation begins like this: {{ ctx }}... Then how
does it continue?

Ending 1: {{ endings[0] }}

Ending 2: {{ endings[1] }}

Ending 3: {{ endings[2] }}

Ending 4: {{ endings[3] }}
```

```
{{answer_choices[label | int()] }}
```

1.11 STRUCTURE TO TEXT

1.11.1 COMMON_GEN

Dataset from Lin et al. (2020). Used in training.

Data Example

Key	Value
concept_set_idx	0
concepts	['ski', 'mountain', 'skier']
target	Skier skis down the mountain

Prompts

```
Ignoring the order of the concepts: {{ concepts | join(", ") }};
Generate a sentence with all the concepts :
```

```
{{target}}
```

```
Put the concepts together to form a sentence: {{ concepts | join(", ")
}}.
```

```
{{target}}
```

```
Construct a sentence with the word {{ concepts | choice }}.

Hint: Use {{concepts | join(", ")}} to restrict the output sentence.
```

```
{{target}}
```

```
{% set seq = [
'From the concepts mentioned below, generate a sentence:',
'Convert the concepts to a sentence:',
'Given the list of concepts, write a sentence:'
] %}
{{ seq | choice }}
{{ concepts | join(", ") }}
```

```
{{target}}
```

Prompt not from the original task.

```
What are the topics in the sentence: {{target}}
```

```
{{ concepts | join(", ") }}
```

Prompt not from the original task.

```
We have the sentence: {{target}};
Extract all the key concepts:
```

```
{{ concepts | join(", ") }}
```

Prompt not from the original task.

```
Can you write a sentence about the topic {{concepts | choice}}?
```

```
{{target}}
```

```
Humans can easily string together abstract concepts to form a coherent
sentence.
For example, with the concepts {{ concepts | join(", ") }}, a simple
sentence can be
```

```
{{target}}
```

```
Given the list of concepts: {{ concepts | join(", ") }};
Generate a sentence with all the concepts :
```

```
{{target}}
```

1.11.2 WIKI_BIO

Dataset from Lebret et al. (2016). Used in training.

Data Example

Key	Value
input_text	{'table': {'column_header': ['name', 'nationality'...
target_text	walter extra is a german award-winning aerobatic p...

Prompts

```
Facts:
{% for n in range (input_text["table"]["column_header"]|length) %}
{% if input_text["table"]["column_header"][n] != "article_title" %}
- {{input_text["table"]["column_header"][n].replace("_"," ") }}:
{{input_text["table"]["content"][n] }}
{% endif %}
{% endfor %}
Based on these bullet points, write a short biography describing the life
of {{input_text["context"]}}.
```

```
{{target_text}}
```

Prompt not from the original task.

```
Read the bio below and try to give details on
{{input_text["context"]}}'s:
{% for n in range (input_text["table"]["column_header"]|length) %} {% if
input_text["table"]["column_header"][n] != "article_title" %}
- {{ input_text["table"]["column_header"][n].replace("_"," ") }}
{% endif %} {% endfor %}

Bio: {{target_text}}
```

```
{% for n in range (input_text["table"]["column_header"]|length) %}
{% if input_text["table"]["column_header"][n] != "article_title" %}
- {{ input_text["table"]["column_header"][n].replace("_"," ") }} is {{
input_text["table"]["content"][n] }}
{% endif %}
{% endfor %}
```

Prompt not from the original task.

```
What type of details about {{input_text["context"]}} can be gathered from
the following bio?

Bio: {{target_text}}
```

```
{% for n in range (input_text["table"]["column_header"]|length) %}
{% if input_text["table"]["column_header"][n] != "article_title" %}
- {{ input_text["table"]["column_header"][n].replace("_"," ") }}
{% endif %}
{% endfor %}
```

Prompt not from the original task.

```
{% for n in range (input_text["table"]["column_header"]|length) %}
{% if input_text["table"]["column_header"][n] != "article_title" and
input_text["table"]["column_header"][n] !="name" %}
- {{ input_text["table"]["column_header"][n].replace("_"," ") }} is {{
input_text["table"]["content"][n] }}
{% endif %}
{% endfor %}

Given the details above, guess who could this information be about.
```

```
{{input_text["context"]}}
```

Prompt not from the original task.

```
What key details about {{input_text["context"]}} can be extracted from
the following bio?

Bio: {{target_text}}
```

```
{% for n in range (input_text["table"]["column_header"]|length) %}
{% if input_text["table"]["column_header"][n] != "article_title" %}
- {{ input_text["table"]["column_header"][n].replace("_"," ") }} is {{
input_text["table"]["content"][n] }}
{% endif %}
{% endfor %}
```

1.12 SUMMARIZATION

1.12.1 CNN_DAILYMAIL 3.0.0

Dataset from See et al. (2017). Used in training.

Data Example

Key	Value
article	`It's official: U.S. President Barack Obama wants l...`
highlights	`Syrian official: Obama climbed to the top of the t...`
id	`0001d1afc246a7964130f43ae940af6bc6c57f01`

Prompts

```
Can you write an outline of the following article in a few points?

Article: {{article}}

{{highlights}}
```

```
Summarise the article:

{{article}}

{{highlights}}
```

```
In 2 or 3 sentences, what are the main points one should remember from
this news article?

Article: {{article}}

{{highlights}}
```

```
Could you please generate a TLDR (Too Long Didn't Read) summary of the
following news article?

Article: {{article}}

{{highlights}}
```

```
Condense the article down to the essentials to present it in the form of
short cards in mobile news apps:

{{article}}

{{highlights}}
```

Prompt not from the original task.

```
Generate a story from key plot points:

{{highlights}}

{{article}}
```

```
Sum the following article in brief: {{article}}

{{highlights}}
```

```
Extract key points from the article based on which the stock market could
react:

{{article}}

{{highlights}}
```

Prompt not from the original task.

```
What details would you include in a storyline to make it more engaging
and informative?

{{highlights}}

{{article}}
```

1.12.2 GIGAWORD

Dataset from Graff et al. (2003). Used in training.

Data Example

Key	Value
document	australia 's current account deficit shrunk by a r...
summary	australian current account deficit narrows sharply

Prompts

```
{{document}}

===

Generate a title for this article:

{{summary}}
```

Prompt not from the original task.

```
Title: {{summary}}
{{document}}
```

```
Make a title for this article: {{document}}
{{summary}}
```

```
First sentence of the article: {{document}}
Title:
{{summary}}
```

Prompt from Radford et al. (2019)

```
{{document}}
TL;DR:
{{summary}}
```

```
{{document}}
===
Given the above sentence, write its title:
{{summary}}
```

```
Write a title for this sentence: {{document}}
Title:
{{summary}}
```

```
{{document}} In a nutshell,
{{summary}}
```

Prompt not from the original task.

```
Title: {{summary}}
===
Write an article with the given title:
```

```
{{document}}
```

1.12.3 MULTI_NEWS

Dataset from Fabbri et al. (2019). Used in training.

Data Example

Key	Value
document	National Archives Yes, it's that time again, ...
summary	{ The unemployment rate dropped to 8.2% last month...

Prompts

```
{% set docs = document.split("3ed2dface8203c4c9dfb1a5dc58e41e0||") |
reject("equalto", "") | list %}
What are the key points across these news articles:
{% for doc in docs %}

Article: {{doc}}
{% endfor %}
```

```
{{summary[2:]}}
```

```
{% set docs = document.split("3ed2dface8203c4c9dfb1a5dc58e41e0||") |
reject("equalto", "") | list %}
Synthesize these documents into a single one:
{% for doc in docs %}

- {{doc}}
{% endfor %}
```

```
{{summary[2:]}}
```

```
{% set docs = document.split("3ed2dface8203c4c9dfb1a5dc58e41e0||") |
reject("equalto", "") | list %}
I want to edit the following articles into a more concise summary:
{% for doc in docs %}

Article: {{doc}}
{% endfor %}
```

```
{{summary[2:]}}
```

```
{% set docs = document.split("3ed2dface8203c4c9dfb1a5dc58e41e0||") |
reject("equalto", "") | list %}
Write a summary of the following articles:
{% for doc in docs %}
```

```
Document: {{doc}}
{% endfor %}
```

```
{{summary[2:]}}
```

Prompt not from the original task.

```
{% set docs = document.split("3ed2dface8203c4c9dfb1a5dc58e41e0||") |
reject("equalto", "") | list%}
Write an expanded news article with plausible details from the following
summary:
{{summary[2:]}}
```

```
{{docs | choice}}
```

```
{% set docs = document.split("3ed2dface8203c4c9dfb1a5dc58e41e0||") |
reject("equalto", "") | list %}
I'm trying to distill these articles down into one:
{% for doc in docs %}

Article: {{doc}}
{% endfor %}
```

```
{{summary[2:]}}
```

1.12.4 SAMSUM

Dataset from Gliwa et al. (2019). Used in training.

Data Example

Key	Value
dialogue	Amanda: I baked cookies. Do you want some? Jerry...
id	13818513
summary	Amanda baked cookies and will bring Jerry some tom...

Prompts

```
Summarize this dialogue: {{dialogue}}
```

```
{{summary}}
```

```
{{dialogue}}
Given the above dialogue, write a summary.
```

```
{{summary}}
```

```
Summarize: {{dialogue}}
```

```
{{summary}}
```

```
{{dialogue}}
To sum up this dialog:
```

```
{{summary}}
```

```
Generate a summary for this dialogue:
{{dialogue}}
```

```
{{summary}}
```

Prompt not from the original task.

```
Write a dialogue that matches this summary: {{summary}}
```

```
{{dialogue}}
```

```
Sum up the following dialogue:
{{dialogue}}
```

```
{{summary}}
```

1.12.5 XSUM

Dataset from Narayan et al. (2018). Used in evaluation.

Data Example

Key	Value
document	Recent reports have linked some France-based playe...
id	29750031
summary	New Welsh Rugby Union chairman Gareth Davies belie...

Prompts

```
{{document}}

===

Write a summary of the text above :
```

```
{{summary}}
```

```
Article: {{document}}

Summary:
```

```
{{summary}}
```

Prompt from Brockman (2020)

```
{{document}}
How would you rephrase that in a few words?
```

```
{{summary}}
```

Prompt from Brockman (2020)

```
My college roommate asked me what this article means:

{{document}}

So I recapped it in layman's terms:
```

```
{{summary}}
```

Prompt from Brockman (2020)

```
{{document}}
This boils down to the simple idea that
```

```
{{summary}}
```

```
Summarize: {{document}}
```

```
{{summary}}
```

```
Summarize this document: {{document}}
Summary:
```

```
{{summary}}
```

```
{{document}}

===

Given the above document, write one sentence to summarize:
```

```
{{summary}}
```

```
First, please read the article below.

{{document}}

Now, can you write me an extremely short abstract for it?

{{summary}}
```

Prompt from Radford et al. (2019)

```
{{document}}

TL;DR:

{{summary}}
```

1.13 Topic Classification

1.13.1 ag_news

Dataset from Zhang et al. (2015b). Used in training.

Data Example

Key	Value
text	Wall St. Bears Claw Back Into the Black (Reuters) ...
label	2

Prompts

```
What label best describes this news article?
{{text}}

{{answer_choices[label] }}
```

```
Is this a piece of news regarding {{"world politics, sports, business, or
science and technology"}}?
{{text}}

{{answer_choices[label] }}
```

```
Would you recommend the following article to a {{"politician"}}, an
{{"athlete"}}, a {{"business executive"}}, or a {{"scientist"}}?

{{ text }}

{{answer_choices[label]}}
```

```
{{text}}

Which of the following sections of a newspaper would this article likely
appear in? {{"World News"}}, {{"Sports"}}, {{"Business"}}, or {{"Science
and Technology"}}?
```

```
{{answer_choices[label] }}
```

```
{{text}}

Which section of a newspaper would this article likely appear in?
```

```
{{answer_choices[label] }}
```

```
{{text}}
Is this a piece of news regarding {{"world politics, sports, business, or
science and technology"}}?
```

```
{{answer_choices[label] }}
```

```
{{text}}
What label best describes this news article?
```

```
{{answer_choices[label] }}
```

1.13.2 DBPEDIA_14

Dataset from Lehmann et al. (2015). Used in training.

Data Example

Key	Value
content	Abbott of Farnham E D Abbott Limited was a Britis...
label	0
title	E. D. Abbott Ltd

Prompts

```
{{content}} Given a list of categories: {{"company, educational
institution, artist, athlete, office holder, mean of transportation,
building, natural place, village, animal, plant, album, film or written
work"}}, what category does the paragraph belong to?
```

```
{{ answer_choices[label] }}
```

```
Pick one category for the following text. The options are - {{"company,
educational institution, artist, athlete, office holder, mean of
transportation, building, natural place, village, animal, plant, album,
film or written work"}}. {{title}} - {{content}}
```

```
{{ answer_choices[label] }}
```

```
{{title}} - {{content}} Given a choice of categories {{"company,
educational institution, artist, athlete, office holder, mean of
transportation, building, natural place, village, animal, plant, album,
film or written work"}}, the text refers to which one?
```

```
{{ answer_choices[label] }}
```

```
"{{title}}", given a list of categories: {{"company, educational
institution, artist, athlete, office holder, mean of transportation,
building, natural place, village, animal, plant, album, film or written
work"}}, what category does the title belong to?
```

```
{{ answer_choices[label] }}
```

1.13.3 TREC

Dataset from Li and Roth (2002). Used in training.

Data Example

Key	Value
label-coarse	0
label-fine	0
text	How did serfdom develop in and then leave Russia ?

Prompts

```
Categories: {{', '.join(answer_choices)}}

What category best describes: {{text}}
Answer:
```

```
{{ answer_choices [label_coarse] }}
```

Prompt not from the original task.

```
{% set label_mapping = {21:0, 18:1, 24:2, 11:3, 14:4} %}
{% if label_coarse == 5 %}
Is this question asking for {{', '.join(answer_choices)}}?
{{text}}
```

```
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

Prompt not from the original task.

```
{% set label_mapping = {39:0, 13:1, 8:2, 40:3, 25:4, 43:5, 27:6, 38:7,
35:8, 41:9, 32:10, 45:11, 14:12} %}
{% if label_coarse == 4 %}
{{text}}

Is this question asking for {{', '.join(answer_choices)}}?
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

Prompt not from the original task.

```
{% set label_mapping = {2:0, 22:1, 19:2, 1:3, 46:3, 23:4, 10:5, 17:6,
33:7, 37:8, 15:9, 30:10, 26:11, 16:12, 28:13, 42:14, 31:15, 20:16, 44:17,
36:18, 14:19} %}
{% if label_coarse == 1 %}
Is this question asking for {{', '.join(answer_choices)}}?
{{text}}
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

Prompt not from the original task.

```
{% set label_mapping = {39:0, 13:1, 8:2, 40:3, 25:4, 43:5, 27:6, 38:7,
35:8, 41:9, 32:10, 45:11, 14:12} %}
{% if label_coarse == 4 %}
Is this question asking for {{', '.join(answer_choices)}}?
{{text}}
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

```
Question: {{text}}

Descriptors: {{', '.join(answer_choices)}}

Best Descriptor?
{{answer_choices[label_coarse]}}
```

```
{{text}}

What is this question asking for?
{{answer_choices[label_fine] }}
```

Prompt not from the original task.

```
{% set label_mapping = {21:0, 18:1, 24:2, 11:3, 14:4} %}
{% if label_coarse == 5 %}
{{text}}

Is this question asking for {{', '.join(answer_choices)}}?
```

```
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

```
Which category best describes the following question: {{text}}

Choose from the following list:
{{', '.join(answer_choices)}}
```

```
{{ answer_choices [label_coarse] }}
```

Prompt not from the original task.

```
{% set label_mapping={0:2, 7:1,  12:0, 9:3} %}
{% if label_coarse == 0 %}
Is this question asking for {{', '.join(answer_choices)}}?
{{text}}
```

```
{{ answer_choices[label_mapping[label_fine]] }}
{% endif %}
```

```
{{text}}

Is this asking about {{(', ').join(answer_choices)}}?
```

```
{{ answer_choices [label_coarse] }}
```

Prompt not from the original task.

```
{% set label_mapping={34:0, 3:1} %}
{% if label_coarse == 2 %}
Is this question asking for an {{', '.join(answer_choices)}}?
{{text}}
```

```
{{answer_choices[label_mapping[label_fine]] }}
{% endif %}
```

Prompt not from the original task.

```
{% set label_mapping = {34:0, 3:1} %}
{% if label_coarse == 2 %}
{{text}}

Is this question asking for an {{', '.join(answer_choices)}}?
```

```
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

```
Is the following question asking about {{', '.join(answer_choices)}}?

{{text}}
```

```
{{ answer_choices [label_coarse] }}
```

Prompt not from the original task.

```
{% set label_mapping = {5:0, 4:1, 6:2, 12:3} %}
{% if label_coarse == 3 %}
Is this question asking for {{', '.join(answer_choices)}}?
{{text}}
```

```
{{ answer_choices[label_mapping[label_fine]] }}
{% endif %}
```

```
What is this question asking for?

{{text}}
```

```
{{ answer_choices[label_fine] }}
```

Prompt not from the original task.

```
{% set label_mapping = {5:0, 4:1, 6:2, 12:3} %}
{% if label_coarse == 3 %}
{{text}}

Is this question asking for {{', '.join(answer_choices)}}?
```

```
{{ answer_choices [label_mapping[label_fine]] }}{% endif %}
```

Prompt not from the original task.

```
{% set label_mapping={0:2, 7:1,  12:0, 9:3} %}
{% if label_coarse == 0 %}
{{text}}

Is this question asking for {{', '.join(answer_choices)}}?
```

```
{{ answer_choices [label_mapping[label_fine]] }}
{% endif %}
```

1.14 Word Sense Disambiguation

1.14.1 super_glue wic

Dataset from Pilehvar and os'e Camacho-Collados (2018). Used in evaluation.

Data Example

Key	Value
end1	36
end2	32
idx	0
label	0
sentence1	Do you want to come over to my place later?
sentence2	A political system with no place for the less prom...
start1	31
start2	27
word	place

Prompts

```
{% if label != -1%}
Does the word "{{word}}" have the same meaning in these two sentences?
Yes, No?
{{sentence1}}
{{sentence2}}

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
Does the word "{{word}}" have the same meaning in these two sentences?
{{sentence1}}
{{sentence2}}

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
Homework

Decide whether the word "{{word}}" is used with the same meaning in the
two following sentences. Answer by yes or no.
{{sentence1}}
{{sentence2}}

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
Sentence A: {{sentence1}}
Sentence B: {{sentence2}}

"{{word}}" has a similar meaning in sentences A and B. True or False?

{{answer_choices[label]}}
{% endif %}
```

Prompt from Brown et al. (2020)

```
{% if label != -1%}
{{sentence1}}
{{sentence2}}
Question: Is the word '{{word}}' used in the same sense in the two
sentences above?

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}

Determine whether the word "{{word}}" is used in the same sense in both
sentences. Yes or no?

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
Determine if the word '{{word}}' is used in the same way in the two
sentences below.
{{sentence1}}
{{sentence2}}

{{answer_choices[label]}}
{% endif %}
```

Prompt from Brown et al. (2020)

```
{% if label != -1%}
{{sentence1}}
{{sentence2}}
Question: Is the word '{{word}}' used in the same sense in the two
sentences above? Yes, No?

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
The word "{{word}}" has multiple meanings. Does it have the same meaning
in sentences 1 and 2? Yes or no?

Sentence 1: {{sentence1}}
Sentence 2: {{sentence2}}

{{answer_choices[label]}}
{% endif %}
```

```
{% if label != -1%}
{{sentence1}}
{{sentence2}}
Similar sense of {{word}}?
```

```
{{answer_choices[label]}}
{% endif %}
```