Strength in Numbers: Estimating Confidence of Large Language Models by Prompt Agreement

Gwenyth Portillo Wightman and **Alexandra DeLucia** and **Mark Dredze**
Center for Language and Speech Processing, Johns Hopkins University
{gwightman, aadelucia, mdredze}@jhu.edu

Abstract

Large language models have achieved impressive few-shot performance on a wide variety of tasks. However, in many settings, users require confidence estimates for model predictions. While traditional classifiers produce scores for each label, language models instead produce scores for the generation which may not be well calibrated. We compare generations across diverse prompts and show that these can be used to create confidence scores. By utilizing more prompts we can get more precise confidence estimates and use response diversity as a proxy for confidence. We evaluate this approach across ten multiple-choice question-answering datasets using three models: T0, FLAN-T5, and GPT-3. In addition to analyzing multiple human written prompts, we automatically generate more prompts using a language model in order to produce finer-grained confidence estimates. Our method produces more calibrated confidence estimates compared to the log probability of the answer to a single prompt. These improvements could benefit users who rely on prediction confidence for integration into a larger system or in decision-making processes.

https://github.com/JHU-CLSP/Confidence-Estimation-TrustNLP2023

1 Introduction

The modern framing of language modeling problems now includes the ability to perform numerous tasks previously handled by specialized supervised discriminative systems. For example, binary and multi-class classification tasks can be framed as text generation, where a large language model (LLM) is given the input and the possible labels, and it generates the best label. More broadly, many reading comprehension, reasoning, and question-answering (QA) tasks can be framed in this multiple-choice style. An advantage to framing tasks in this manner is the ability to perform few-shot learning via in-context learning, in which a task can be performed based on only a handful of examples (Arora et al., 2022; Brown et al., 2020; Kojima et al., 2022; Sanh et al., 2021; Zhou et al., 2022a). Instead of collecting a large dataset and investing time in training a new model, a user could utilize an existing LLM for a new task by labeling a few examples and crafting a prompt: the input template which instructs the model to perform the given task.

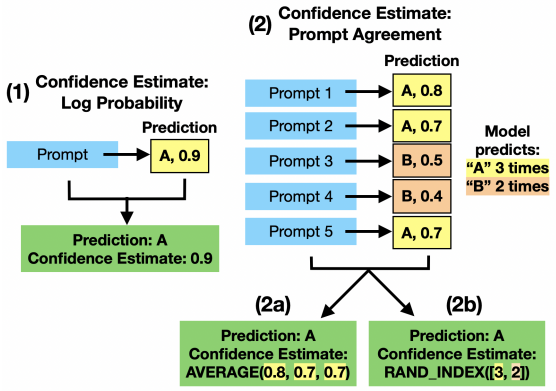

Figure 1: A comparison of our proposed *prompt agreement* confidence scores (2) and the commonly used log probability (1). Log probability is based on a single prompt, while the prompt agreement confidence estimate uses multiple prompts to determine a confidence estimate using (2a) the log probabilities from majority label or (2b) the Rand index of label frequencies.

One challenge to utilizing LLMs in this manner is producing well-calibrated confidence scores for model predictions. A calibrated confidence score aids in the interpretation of model predictions (Guo et al., 2017) and may be crucial if models become integrated into high-risk domains like healthcare and finance (Jiang et al., 2021). A model is considered well calibrated if its prediction probabilities are aligned with the actual probability of its predictions being correct (Jiang et al., 2021). If a model says an answer has 90% confidence, then we should expect it to be correct 90% of the time.

Formally, the probability that the predicted label $\hat{Y}$ is equal to the correct label Y for input X should be equal to the model's predicted confidence (Nguyen and O'Connor, 2015). For supervised discriminative systems, confidence scores emerge from output probabilities or normalizing model scores to be between 0 and 1. For linear models, posterior probability serves as a reasonable confidence score because as the amount of evidence that supports prediction Y increases, confidence also increases (Dong et al., 2018). However, prior work shows that these probabilities are not well calibrated for non-linear models (Johansen and Socher, 2017).

It is less clear how we can obtain confidence scores from LLMs. One approach is to use the (log) probability of the generation. However, these scores correspond to the likelihood of a text sequence given some context, as opposed to the actual probability of the label. For example, the model may assign probability mass to alternate generations that reflect the same answer (e.g. "Answer A" vs. "The answer is A"). Other creative approaches include asking the model to generate statements of confidence (e.g. "90% confidence in the label"), but it is unclear how to calibrate these open-ended statements (Lin et al., 2022). Model self-consistency can be used to identify the most confident model output, but it is unclear how to produce a meaningful score (Wang et al., 2022). Instead, we turn to another trend in LLMs: diverse prompts. Sanh et al. (2021) showed that by writing variations of prompts for a range of tasks, they produced models better able to generalize to new domains. Similarly, Chung et al. (2022) found that training on a diverse set of tasks improved model performance. We consider whether measuring the stability of an answer across a diverse set of prompts can be used to estimate model confidence.

We propose to measure LLM answer confidence by *prompt agreement*, whether the response of a model remains consistent across multiple prompts for a given instance. We prompt an LLM with multiple different prompts that instruct the model to perform the same task for a single input instance and measure the agreement of the model responses across these prompts. We consider two approaches, represented in Figure 1. First, we measure the log probability of the response across multiple prompts that agree on the answer. Second, we measure the diversity in answers across different prompts in the model output, concluding that answers which appear in more responses have relatively higher confidence. We compare these methods to the log probability of the answer produced in response to the official task prompt. We find that across a range of datasets and models, our methods consistently provide more accurate estimates of confidence.

Our contributions are as follows:

- We show that the confidence estimate based on multiple prompts more accurately reflects the chance that a model is correct as compared to log probabilities from a single prompt.
- We demonstrate these results on ten multiple-choice QA datasets and three models: T0++ (Sanh et al., 2021), FLAN-T5 XXL (Chung et al., 2022), and GPT-3 (Brown et al., 2020).
- We utilize automated prompt generation methods to test whether they can be used in place of human-authored prompts to create better confidence estimates.

2 Related Work

We present the relevant background concepts of in-context learning and prompt sensitivity, and then outline approaches to confidence estimation.

2.1 In-Context Learning

Recent work has shown that model performance can be improved by in-context learning (ICL), in which the model is conditioned on a natural language instruction and several demonstrations of the task (few-shots) and then completes additional instances of the task by predicting what comes next (Radford et al., 2019; Brown et al., 2020).

However, the efficacy of ICL varies depending on the prompt. Prompts that appear semantically similar to humans can still yield different results (Gao et al., 2021; Schick and Schütze, 2021), and many efforts have explored best practices for few-shot learning. Techniques have emerged to assist prompt engineers with creating and selecting the best prompts (Sorensen et al., 2022). In addition to the choice of prompt, performance varies based on the choice of training examples and the order of the training examples (Zhao et al., 2021). This sensitivity makes ICL less reliable in practice.

Chen et al. (2022) found that sensitive predictions were less likely to be accurate. This suggests that a model's predictions may be less accurate when they lack *consistency* (Zhou et al., 2022b), defined as the model's ability to make the same prediction across generations for the same input

(Wang et al., 2020). Consistency has been used in semi-supervised learning and ensemble learning to encourage predictions to be consistent across perturbations of the input, such as noise or paraphrasing (Bachman et al., 2014; Sajjadi et al., 2016; Xie et al., 2019; Zhai et al., 2019). Consistency inspires our approach to estimating confidence based on model behavior across a set of prompts.

2.2 Confidence Estimation

Confidence estimation is the counterpart to uncertainty estimation, which quantifies a model's lack of confidence in its predictions. Previous work has shown that modeling uncertainty improves task performance on neural machine translation (Wang et al., 2019), document quality prediction (Shen et al., 2019), sentiment analysis, named entity recognition, and language modeling using convolutional and recurrent neural network models (**?**).

Work on model confidence estimation for NLP has included a range of models–Naive Bayes and logistic regression (Nguyen and O'Connor, 2015), neural networks (Jagannatha and yu, 2020)–and tasks—structured prediction (Jagannatha and yu, 2020), natural language understanding (Desai and Durrett, 2020; Kamath et al., 2020; Kong et al., 2020), and neural machine translation systems (Kumar and Sarawagi, 2019). Kamath et al. (2020) found that QA models are overconfident in out-of-domain tasks when asked to answer as many questions as possible while maintaining high accuracy. More recently, this work has turned to language models, and researchers have struggled to obtain sensible confidence measures. Jiang et al. (2021) found that language models such as T5, BART, and GPT-2 did not produce well-calibrated scores based on generation probabilities for QA tasks.

A variety of methods have been proposed to obtain calibrated confidence measures from LLMs. Jiang et al. (2021) experiment with several calibration methods, including fine-tuning, post hoc probability modification, or adjustment of the predicted outputs or inputs. Kong et al. (2020) use a regularized fine-tuning method to obtain better calibration for both in-distribution and out-of-distribution data. Xiao et al. (2022) focus on the design choices for pre-trained language model-based prediction pipelines, suggesting that the calibration of the model depends on the choice of the fine-tuning loss function. Desai and Durrett (2020) demonstrated a more calibrated model trained with label smoothing. Unfortunately, these methods are not feasible for LLMs such as GPT-3, which have already been trained and cannot be easily modified without substantial compute power or model access.

An alternative approach is to rely on post hoc calibration methods. Established techniques include training a separate, smaller model to identify incorrect predictions (Kumar and Sarawagi, 2019; Kamath et al., 2020) or to adjust predictions (Isotonic Regression (Niculescu-Mizil and Caruana, 2005) and forecaster (Jagannatha and yu, 2020)), but these methods require a separate validation set. Similarly, a validation set can also be used for tuning decoding hyperparameters for better calibration, as in temperature scaling (Desai and Durrett, 2020; Jiang et al., 2021). Dong et al. (2018) present metrics to measure three kinds of uncertainty (model uncertainty, data uncertainty, and input uncertainty) that may lead to miscalibration. Our work contributes to the ongoing work of calibration through post hoc techniques, which are still feasible for larger models, particularly when we lack access to the model weights or don't have the compute to fine-tune the model. Instead of requiring access to validation sets or training external models, we introduce a stand-alone method.

Our approach utilizes a post hoc confidence estimate for a generated model prediction by measuring agreement across multiple prompts. The idea of majority voting and prompts appears in several related studies. Zhou et al. (2022a) rely on the idea that a single task can be described by multiple prompts, and encourage model behavior to be consistent across different prompts (*prompt consistency*). They use consistency across prompts to engineer new prompts as written by an LLM. Wang et al. (2022) use self-consistency to improve chain-of-thought reasoning. They found a correlation between consistency and accuracy, suggesting that consistency provides an estimate of how certain the model is about its generations. Arora et al. (2022) use voting in their Ask Me Anything (AMA) prompting method to determine an input's label by collecting noisy votes from a set of machine-generated prompts that vary in quality. A version of BARTSCORE (BARTSCORE-PROMPT) utilizes a similar prompt-ensembling scheme (with generated prompts) for prompting BART to score summarization quality (Yuan et al., 2021). These studies provide support for our idea that majority voting can

inform confidence scores.

Finally, Lin et al. (2022) take a unique approach to obtaining confidence from LLMs: they ask the model! For example, GPT-3 generates confidence estimates when asked to verbalize its confidence with statements like "90% confidence." While these generations cannot easily be compared and calibrated across tasks, it further suggests that models have some notion of confidence.

The idea of model confidence is related to the style of generation and the certainty with which a model expresses answers. Informal analyses of models, especially those focused on scientific generations like Galactica (Taylor et al., 2022), have found that models frame answers in a confident tone regardless of the actual factuality of the statement. This observation of answer framing may be related to our task of assigning a confidence score to a generation.

3 Estimating Confidence through Multiple Prompts

We propose estimating model confidence through multiple prompts based on *prompt agreement*, i.e., the consistency among a model's generations in response to a set of diversely worded prompts. We prompt the model multiple times using different prompts, each of which asks the model to respond to a given question-answer (QA) input. Intuitively, the more often that different prompts favor the same generation, the greater confidence the model has in that generation. For example, suppose that for a given question queried across ten prompts, the model always replies *eggplant*. For a second question queried with the same prompts, the model answers *potato* (5 times) and *eggplant*, *cucumber*, *squash*, *carrot* and *kale*. We would say the model is more confident in its answer to the first question.

We score confidence via prompt agreement in two ways: (1) log probabilities across multiple prompts and (2) answer agreement across multiple prompts. We compare these to a baseline of the log probability of the response to a single prompt.

3.1 Log Probabilities

Log probability of the generation is a common method for confidence estimation (Jiang et al., 2021; Nguyen and O'Connor, 2015; Dong et al., 2018). For each instance we query the model using the single, official task prompt for the dataset and use the log probability of the generation.[1]

3.2 Log Probabilities Across Prompts

For each instance, we query the model with each available prompt and record the resulting answer and associated log probability. We compute the *majority label* across these prompts and assign it a confidence of the average log probabilities across these prompts. Figure 1 shows this technique in practice (2a), where the model predicts A three times and B twice, making A the majority label. The confidence estimate is the average of the log probabilities from each time A was predicted. In case of a tie, we compute the average log probability of each tied answer and select the answer with the highest average log probability.

3.3 Answer Agreement Across Multiple Prompts

A drawback to averaging the log probabilities of the majority is that it does not reflect overall agreement across the prompts. Consider the example in Figure 1, where the model predicts "A" three times and "B" twice and compare to a case where the model predicts "A" three times, "B" once and "C" once. The model appears to be more uncertain in the second case, yet averaging the majority log probability would yield the same score.

We create a confidence score that reflects answer agreement across multiple prompts. We count the number of times the model predicts each answer and view this agreement list as a form of clustering of the prompts into answer bins. We use Rand index (Rand, 1971), a metric that measures similarity between two clusterings, to quantify the amount of agreement within this list. We compute the Rand index between the observed "clustering" and the "ideal" clustering, where the model predicts the same answer for every prompt (highest confidence). This measure naturally incorporates cases with varying numbers of prompts.

The resulting Rand index is a confidence score for answer agreement across multiple prompts. We note that unlike our other methods, this does not yield a probability. We address this in our evaluation metrics below.

[1] Section 4 details how we obtain these scores for each model.

4 Models

Our confidence estimation methods are compatible with multiple language models. We evaluate our methods on three popular models, chosen because of their strong few-shot task performance, and focus on the largest models in each model "family" because they are the highest performing.

For T0++ and FLAN-T5, we use the Hugging Face implementations locally.[2]

T0++ is an 11B parameter T5-based model that was trained with a multitask mixture and multiple prompts on 55 datasets to improve zero-shot task generalization (Sanh et al., 2021).

FLAN-T5-XXL is an 11B parameter T5 model (Raffel et al., 2019) fine-tuned on 1.8k instruction oriented tasks (Chung et al., 2022). Task fine-tuning (FLAN) produces state-of-the-art results on few-shot performance across several benchmarks.

GPT-3 is a 176B parameter GPT-style model trained with a causal language modeling objective (Brown et al., 2020). We use `text-davinci-002`, an instruction-tuned version of GPT-3 (Ouyang et al., 2022), through the OpenAI API. Due to the restrictions in obtaining all token logits in a single API call, we generate a model response and match it to the closest answer choice for cost efficiency. See Appendix A for details.

For each prompt and for each QA instance, we need to obtain (1) the model's selected answer from the multiple-choice list and (2) the log probability of the selected answer. To obtain the best answer we use *rank scoring*, which evaluates the model log probability for generating each answer from the multiple-choice list and selects the best option (Brown et al., 2020; Sanh et al., 2021). For T0++ and FLAN-T5-XXL we use Sanh et al. (2021)'s publicly available evaluation code,[3] modifying it to return log probabilities of the answers. We run these models on a compute instance with 4 A100 40GB GPUs, with a per-device batch size of 8 for all datasets except Dream (batch size of 1).

Finally, we omit results for automatically generated prompts for GPT-3 due to the high financial cost of using the API for so many prompts. We include these results for the other methods.

[2] `https://huggingface.co/bigscience/T0pp` and `https://huggingface.co/google/flan-t5-xxl`

[3] `https://github.com/bigscience-workshop/t-zero/blob/master/evaluation/run_eval.py`

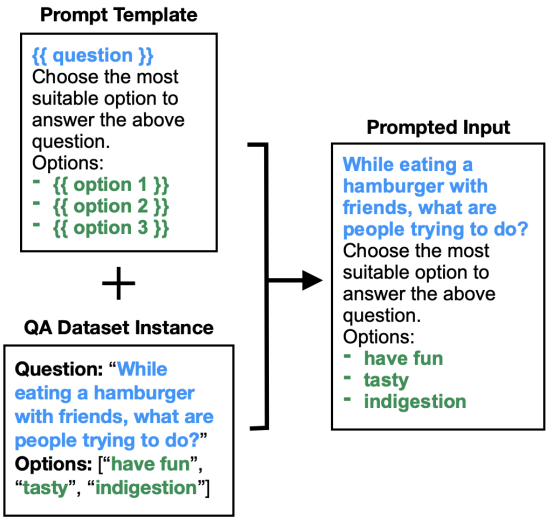

Figure 2: An example of a prompt template applied to a QA instance.

5 Data

We evaluate our method across ten multiple-choice question-answering datasets. For each dataset, we have the official task prompt and a source of diverse prompts for the same task. Within a dataset, each instance contains contextual information, a series of multiple-choice answers, and annotations indicating the correct answer.

We use the following multiple-choice QA datasets from the T0 training mixture (Sanh et al., 2021): CoS-E v1.11 (Rajani et al., 2019), Cosmos QA (Huang et al., 2019), DREAM (Sun et al., 2019), QASC (Khot et al., 2020), Quail (Rogers et al., 2020), Quarel (Tafjord et al., 2019a), Quartz (Tafjord et al., 2019b), SciQ (Welbl et al., 2017), Social IQA (Sap et al., 2019), and WIQA (Tandon et al., 2019). We exclude WikiHop (Tu et al., 2019) due to the extra computational resources needed for this dataset. We use only the validation splits.

6 Prompts

We pair these datasets with three sources of prompts: the official task prompt and two sources for diverse set prompts for each task. First, we use the official task prompt as defined in the original paper for each dataset.

Second, we use the diverse human-authored prompts provided by Sanh et al. (2021). Each prompt is a template that contains text strings and placeholders to insert the question and answer choices (see Figure 2). We only use the T0 prompts that correspond to the original task intended by the

dataset's authors. We refer to these as the Multiple Human prompts. We apply these prompts to the QA data using the PromptSource library (Bach et al., 2022) and evaluation code for T0.[3]

Third, we create a larger set of prompts through automated prompt generation. While having multiple prompts leads to better confidence scores, not every task has multiple human-authored prompts available. Furthermore, if multiple prompts are helpful, perhaps a larger set would provide more fine-grained confidence estimates. Automatically generating prompts addresses both of these cases.

Many methods have been proposed for automatically generating LLM prompts. Most prompt generation methods assume either a single prompt for a task (Shin et al., 2020; Zhong et al., 2021; Gao et al., 2021) or a unique prompt for each input (Wu et al., 2022; Zhang et al., 2022). Instead, we seek to generate multiple prompts for each task. We draw inspiration from the iterative prompt generation process of Zhou et al. (2022b), which generates paraphrases of a prompt by asking a LLM to paraphrase instructions with different *prompt generation prompt*s (PGP). For example, by providing an LLM the PGP "Generate a variation of the following instruction while keeping the semantic meaning," we can obtain prompt variations. We use a total of 15 PGPs (listed in Table 8 in Appendix E.2), 14 of which we authored and the final PGP being from Zhou et al. (2022b). Figure 3 summarizes the prompt generation process.

We use this method with GPT-3 `text-davinci-002` to generate a set of **Automatically Generated Prompts** (AGPs) based on 31 instruction statements extracted from the T0 prompts (listed in Table 9 in Appendix E). We generate multiple prompts for each GPT-3 query with a temperature of 0.7 to allow for randomness and repeat each query 3 times. We obtained 465 paraphrase queries (31 T0 instructions $\times$ 15 PGPs), which repeated 3 times gives 1395 paraphrases. After removing duplicates, the number of unique paraphrases per dataset varies, ranging from 16 for WIQA to 158 for Quartz. We insert the paraphrased instructions into the existing dataset templates (which indicate where the question and answer choice should go) to generate new prompt templates. For each dataset, we limit the total number of AGPs to 50 by random selection.

Table 3 in Appendix C shows the number of prompts for each dataset: a single official prompt, a set of Multiple Human prompts, and a larger set of AGPs.

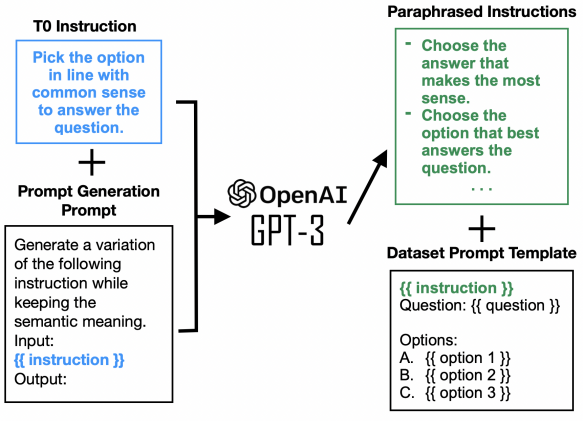

Figure 3: We create prompts by using GPT-3 to generate paraphrased instructions and inserting the paraphrased instructions into a dataset prompt template.

7 Evaluation

Does measuring confidence across multiple prompts yield better calibrated confidence scores? A common approach to measuring calibration is **expected calibration error (ECE)** (Guo et al., 2017), which buckets the prediction probabilities and measures the empirical accuracy of each bucket with its average estimated probability (confidence).[4] The discrepancy between these terms is the calibration gap; lower gaps indicate better calibration. ECE ranges from 0 (perfect calibration) to 1 (lowest calibration). We utilize this method to compare log probabilities obtained from a single prompt to those from multiple prompts. For each dataset, we use 10 evenly-spaced bins and set the min and max of the bins according to the minimum and maximum average log probability in the dataset.

We measure agreement across prompts using **Rand index**, which does not give normalized scores that can be interpreted as probabilities. We could convert these scores into probability confidence scores in two ways. 1) Measure the empirical accuracy of different ranges of Rand index on a held-out validation set, then assign confidence scores based on those accuracies. The drawback to this approach is it requires a separate held-out set for calibration, which may be an unrealistic assumption, especially in few-shot settings. 2) Normalize the empirical Rand index scores to form a

[4]While Nixon et al. (2019) found shortcomings of ECE to measure calibration for deep learning models, it still serves as best practice in this area.

probability distribution. We experimented with this approach but found that how we bucketed and normalized the scores heavily influenced ECE results, which produced an unfair evaluation setting.

Instead, we view Rand index scores as a *relative confidence score* between instances, where a higher score means "more confident." We propose an evaluation metric that considers relative confidence of answers between instances. We rank instances in a dataset according to their confidence scores (log probability or Rand index), with the highest scoring instance (e.g., largest log probability or Rand index) at the top of the list. We evaluate each confidence estimation method on how well it ranks correct predictions higher than incorrect ones.

Most evaluation metrics for ranking are geared towards an information retrieval setting where the number of items in the list can vary, different items can be included by each model, only a few items are "relevant", or we have close to a total ordering over the ranked items. Our ranked lists differ significantly from these settings. Therefore, we choose a simple, intuitive ranking evaluation: **swapped pairs**, based on the ranking loss function from Díez Peláez et al. (2006); Joachims (2002). A list is scored based on the number of item pairs that need to be swapped to create a correct ordering. This penalizes methods that have higher confidence in predictions that were incorrect over correct predictions. Swapped pairs is not normalized and grows with the number of items in a ranked list (from 0, i.e., perfect rank ordering, to $\frac{n*(n-1)}{2}$, i.e., worst rank ordering, where n is the number of items to be ranked). We report macro-averaged results by dividing the total swapped pairs by dataset size, after filtering out any invalid predictions.[5]

8 Results

Multiple prompts provide a more calibrated confidence estimate than a single prompt. Table 1 shows the results for ECE and Swapped pairs across confidence methods and models. Estimating confidence using multiple prompts consistently provides a better calibrated score as compared to confidence scores based on a single prompt. For ECE, using the log probability for multiple human-authored prompts always improves over a single prompt. Additionally, we observe that the ECE and swapped pairs metrics are in agreement with each other; across each method and model they yield the same ordering of the results, supporting our assertion that swapped pairs is a sufficient metric for measuring relative confidence scores. This indicates that swapped pairs can be used to evaluate calibration. Additionally, we observe that different models vary considerably in their scores. Specifically, we find that T0++ and GPT-3 are much better calibrated than FLAN-T5-XXL, although using our method dramatically decreases the gap. This may be partly explained by the differences in model accuracy on these QA tasks, as discussed below.

Measuring confidence using prompt agreement with human-authored prompts also improves over using a single prompt as measured by swapped pairs. There is not a clear winner between the log probability and agreement methods, as each obtains the most calibrated scores for some models. However, both ways of using multiple human-authored responses improve over a single prompt.

Automatically generated prompts show mixed results. Sometimes automatically generated prompts improve over a single prompt (ECE on FLAN-T5-XXL), and sometimes they do not. We suspect that this may be related to the quality of the prompts. Poorly written prompts that obtain worse accuracy on the task give worse confidence scores. To test this hypothesis, for each dataset we select the top 10 prompts with the highest accuracy on the validation set. We compare the confidence scores from using these 10 prompts with the scores from using all AGPs. However, this filtering still does not yield consistent improvements on ECE or swapped pairs. There may be other factors that prevent automatically generated prompts from producing better confidence scores. For example, they may have insufficient diversity or may be worse in some other manner. In contrast, we know that the human-authored prompts were carefully written by people who have experience prompting language models. Despite the poor performance of AGPs, they still show improved performance over a single prompt, indicating that AGPs could serve as a substitute for human-authored prompts if human-authored prompts are not available.

The multiple human-written prompts method appears to be the most calibrated overall. There is not a clear trend as to which method should be used in practice. For example, Table 1 shows that the best method for T0++ is Human + Mul-

[5]We experimented with other normalized methods but the ordering of the methods were unchanged in the results.

Confidence Method	ECE (↓) T0++	FLAN-T5-XXL	GPT-3	Swapped Pairs (↓) T0++	FLAN-T5-XXL	GPT-3
Human Prompts						
- Single / log-prob	5.66	7.35	4.18	137.14	203.93	133.68
- Multiple / log-prob	**1.61**	**2.39**	**2.23**	**89.53**	135.52	130.23
- Multiple / agreement	-	-	-	125.75	128.87	**105.36**
Automatically Generated Prompts						
- Top 10 / log-prob	6.17	4.89	-	154.05	166.81	-
- Top 10 / agreement	-	-	-	168.56	123.08	-
- All / log-prob	6.20	5.23	-	153.57	169.97	-
- All / agreement	-	-	-	164.28	**118.52**	-

Table 1: Expected Calibration Error (ECE) and Swapped Pairs results by model (T0++, FLAN-T5-XXL, GPT-3), prompt type (human written or automatically generated; single or multiple), and confidence estimation method (log probability or agreement).

Confidence Method	Accuracy T0++	FLAN-T5-XXL	GPT-3
Human Prompts			
- Single / max log-prob/agreement	0.69	0.61	0.56
- Multiple / max log-prob	0.76	0.74	0.65
- Multiple / agreement	**0.80**	**0.80**	**0.69**
Automatically Generated Prompts			
- Top 10 / max log-prob	0.72	0.74	-
- Top 10 / agreement	0.72	0.75	-
- All / max log-prob	0.71	0.72	-
- All / agreement	0.72	0.74	-

Table 2: Accuracy by model (T0++, FLAN-T5-XXL, GPT-3) and prompt type (human written or automatically generated; single or multiple), where the prediction is either the label with the maximum log probability or the majority label. Note that because the Single prompt setting contains only one prompt, Single / max log-prob and Single / agreement result in the same accuracy.

tiple / log-prob, while AGP + All / agreement is best for FLAN-T5-XXL. However, we can see that across all models, using multiple prompts (typically human-written prompts, opposed to AGPs) performs the best, suggesting that it would be the most promising confidence method in practice.

Higher accuracy is linked to a larger improvement in calibration. We now consider how the accuracy for each type of prompt is correlated with improvements in calibration from using multiple prompts. From the accuracy results in Table 2, we observe that T0++ achieves the highest accuracy and is the best calibrated among the models, while FLAN-T5-XXL achieves the same level of accuracy with lower calibration. Using multiple prompts rather than a single prompt consistently results in higher accuracies across all models, which may be why T0++ is better calibrated. However, we find that GPT-3 has a worse accuracy than FLAN-T5-XXL, yet GPT-3 is better calibrated than FLAN-T5-XXL according to ECE.

9 Conclusion

Our experiments with T0++, FLAN-T5-XXL, and GPT-3 suggest that prompt agreement provides a more calibrated confidence estimate than the typical approach of log probability from a single prompt. We find mixed results in scaling up the number of prompts using automatically generated prompts. Experimenting with additional prompt generation methods may enable the automatically generated prompt approach to produce even better calibrated confidence scores. We leave this to future work.

Limitations

The main limitation of this work is the lack of human evaluation. Since confidence scores are typically used for model explainability, a practical evaluation of scores from our method would be a human-in-the-loop scenario where a user is tasked with understanding a system and making decisions based on the output. The primary questions for this human study would be to determine if our scores are more useful to users than other methods, such as log probabilities, and would being presented with confidence scores lead to different decisions.

Second, we focused on multiple-choice questions, with a specific set of possible options. Since MC QA and classification are so similar, our analysis of many MC QA datasets is sufficient to show that our method works for text classification. However, there are other use cases for these models that do not have pre-determined answer choices, such as open-ended questions or summarization.

Third, while we supported our decision to only use the largest models in each model family due to their superior performance, we acknowledge that replicating our study across different model sizes (e.g., FLAN-T5-Small, -Base, -Large, -XL, -XXL), is useful for ensuring our method is robust to the number of parameters.

Further, we acknowledge a drawback of our method is the difficulties in comparing across other calibration techniques since the Rand index scores are not normalized. While there are ways to normalize the scores (see Section 7), we decided against these methods in our evaluation because they were either against our zero-shot setting or heavily influenced ECE results based on how scores were normalized.

We leave these questions to future work.

References

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. ArXiv:2210.02441 [cs].

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with pseudo-ensembles.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2022. On the relation between sensitivity and accuracy in in-context learning.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. *arXiv:2003.07892 [cs]*. ArXiv: 2003.07892.

Jorge Díez Peláez, Juan José del Coz Velasco, Antonio Bahamonde Rionda, et al. 2006. A support vector method for ranking minimizing the number of swapped pairs.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

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

Abhyuday Jagannatha and hong yu. 2020. Calibrating Structured Output Predictors for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, Online. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know *When* Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

Alexander Johansen and Richard Socher. 2017. Learning when to skim and when to read. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 257–264, Vancouver, Canada. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective Question Answering under Domain Shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated Language Model Fine-Tuning for In- and Out-of-Distribution Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1326–1340, Online. Association for Computational Linguistics.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 625–632, New York, NY, USA. Association for Computing Machinery.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2019. Measuring calibration in deep learning.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8722–8731.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv:2110.08207 [cs]*. ArXiv: 2110.08207.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Timo Schick and Hinrich Schütze. 2021. 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. Association for Computational Linguistics.

Aili Shen, Daniel Beck, Bahar Salehi, Jianzhong Qi, and Timothy Baldwin. 2019. Modelling uncertainty in collaborative document quality assessment. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 191–201, Hong Kong, China. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset and models for answering questions about qualitative relationships. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7063–7071.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. Quartz: An open-domain dataset of qualitative relationship questions. *arXiv preprint arXiv:1909.03553*.

Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. 2019. Wiqa: A dataset for" what if..." reasoning over procedural text. *arXiv preprint arXiv:1909.04739*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. *arXiv preprint arXiv:1905.07374*.

Lijing Wang, Dipanjan Ghosh, Maria Teresa Gonzalez Diaz, Ahmed Farahat, Mahbubul Alam, Chetan Gupta, Jiangzhuo Chen, and Madhav Marathe. 2020. Wisdom of the ensemble: Improving consistency of deep learning models.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802, Hong Kong, China. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, V. G. Vinod Vydiswaran, and Hao Ma. 2022. Idpg: An instance-dependent prompt generation method. In *North American Chapter of the Association for Computational Linguistics*.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. 2019. S4l: Self-supervised semi-supervised learning.

Yue Zhang, Hongliang Fei, Dingcheng Li, and Ping Li. 2022. PromptGen: Automatically generate prompts using generative models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 30–37, Seattle, United States. Association for Computational Linguistics.

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

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022a. Prompt consistency for zero-shot task generalization.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022b. Large language models are human-level prompt engineers.

A Generating Predictions from GPT-3

The GPT-3 API does not allow direct access to all the token probabilities, and the method of gathering logits through multiple API calls for each answer choice is cost-prohibitive. In order to perform rank scoring with GPT-3, we generate the best answer from the model by asking for deterministic generations (temperature of 1) and using `<|endoftext|>` as the stop token. We take the log probability as the sum of the token log probabilities up to and including the first `<|endoftext|>` token.

While GPT-3 generally does well at following directions, it often does not generate an answer which exactly matches one of the multiple-choice options. We map each GPT-3 generation to one of the valid options by finding the answer that has the greatest 1,2-gram overlap with the generation (after lowercasing and removing punctuation and whitespace). We label a GPT-3 response as *invalid* if it has no overlap with a valid option. When evaluating confidence estimates, we filter out instances that resulted in at least one invalid prediction for a prompt. See Table 10 for statistics about the number of valid GPT-3 predictions.

B Dataset Information

We present information about the datasets in Table 10: links to access to datasets on Hugging Face, the size of validation split, the number of instances that GPT-3 generated valid predictions for on the official prompt, and the number of instances that GPT-3 generated valid predictions for across all Multiple Human prompts.

C Number of Prompts Per Dataset

Table 3 shows the number of Multiple Human (MH) prompts and automatically generated prompts (AGP) per dataset. In addition to these prompts, each dataset has a single prompt (Official Prompt) which comes from the paper in which the dataset authors introduced the dataset.

D Confidence and Accuracy per Dataset

D.1 Confidence

Table 4 shows the ECE and swapped pairs results for each dataset when using human-written prompts. Table 5 shows the ECE and swapped pairs results for each dataset when using automatically-generated prompts.

Dataset	MH	AGP
CoS-E v1.11	6	48
Cosmos QA	10	50
DREAM	2	19
QASC	5	50
Quail	10	50
Quarel	5	39
Quartz	8	50
SciQ	4	50
Social IQA	4	25
WIQA	2	16

Table 3: The number of Multiple Human (MH) Prompts and Automatically Generated Prompts (AGPs) per dataset.

D.2 Accuracy

Table 6 shows the accuracy results for each dataset when using human-written prompts. Table 7 shows the accuracy results for each dataset when using automatically generated prompts.

E Automatically Generated Prompts

E.1 Instructions Used for Prompt Generation

In Table 9, we list the instructions that were used to generate additional prompts. These instructions come from the prompts used to train T0 (Sanh et al., 2021).

E.2 Prompt Generation Prompts

In Table 8, we list the prompt generation prompts (PGP) that were used generate new prompts. Within each PGP, we substitute an instruction from Table 9 in place of "{{ instruction }}" before gathering a response from GPT-3.

E.3 Paraphrased Instructions

In Table 11, we provide the number of paraphrased instructions per dataset. We include statistics about the total number of unique paraphrased instructions and the final number of paraphrased prompts (after randomly selecting up to 50 prompts per dataset). In Tables 12 to 21 we provide the paraphrased instructions for each dataset.

Confidence Method (Human Prompts)	ECE (↓) T0++	FLAN-T5-XXL	GPT-3	Swapped Pairs (↓) T0++	FLAN-T5-XXL	GPT-3
Log prob (single)						
cos_e	1.62	6.10	3.87	101.50	121.64	69.02
cosmos_qa	10.83	5.17	3.59	248.92	353.44	203.20
dream	12.05	9.73	4.87	152.08	207.76	101.71
qasc	4.06	5.02	4.99	1.14	8.24	20.08
quail	1.40	8.75	4.92	131.88	202.99	170.50
quarel	4.60	7.69	5.90	28.49	32.95	14.70
quartz	6.95	6.70	5.56	22.23	33.75	7.85
sciq	1.09	4.36	0.94	37.39	94.10	30.56
social_i_qa	2.37	12.07	3.51	219.66	209.21	42.50
wiqa	11.59	7.94	3.69	428.07	775.16	676.66
Average	5.66	7.35	4.18	137.14	203.93	133.68
Log prob (multiple)						
cos_e	1.02	1.45	1.65	51.28	104.11	43.73
cosmos_qa	1.57	3.27	2.00	141.05	265.93	238.44
dream	1.01	1.41	2.23	46.27	61.33	90.22
qasc	2.06	2.85	2.64	3.06	2.85	12.99
quail	1.02	1.11	1.54	74.99	97.23	199.40
quarel	0.91	1.47	2.43	27.19	28.08	16.03
quartz	1.15	1.51	2.41	8.62	9.21	5.25
sciq	2.06	2.98	2.17	23.81	16.30	11.93
social_i_qa	4.29	6.20	3.15	167.03	163.74	35.55
wiqa	1.02	1.63	2.12	352.06	606.44	648.76
Average	1.61	2.39	2.23	89.53	135.52	130.23
Agreement (multiple)						
cos_e	-	-	-	65.32	49.25	24.56
cosmos_qa	-	-	-	116.60	152.34	158.99
dream	-	-	-	119.79	108.39	105.60
qasc	-	-	-	0.75	0.67	6.30
quail	-	-	-	94.37	84.73	91.74
quarel	-	-	-	31.69	22.39	14.92
quartz	-	-	-	11.12	6.94	3.71
sciq	-	-	-	5.61	5.29	6.19
social_i_qa	-	-	-	138.84	151.37	41.13
wiqa	-	-	-	673.39	707.31	600.43
Average	-	-	-	125.75	128.87	105.36

Table 4: Expected Calibration Error (ECE) and Swapped Pairs results for Human Prompts (single or multiple) by model (T0++, FLAN-T5-XXL, GPT-3), confidence estimation method (log probability or agreement), and dataset.

Confidence Method (AGP)	ECE (↓) T0++	FLAN-T5-XXL	GPT-3	Swapped Pairs (↓) T0++	FLAN-T5-XXL	GPT-3
Top 10 (log prob)						
cos_e	12.12	5.31	-	132.08	108.60	-
cosmos_qa	11.17	8.90	-	262.76	353.77	-
dream	1.15	1.38	-	47.25	56.94	-
qasc	1.03	1.31	-	0.71	2.16	-
quail	10.63	10.34	-	157.22	228.76	-
quarel	0.95	2.37	-	27.01	30.22	-
quartz	1.17	1.22	-	8.96	8.69	-
sciq	0.96	1.08	-	9.34	6.15	-
social_i_qa	12.08	9.01	-	217.91	216.98	-
wiqa	10.48	7.98	-	677.22	655.82	-
Average	6.17	4.89	-	154.05	166.81	-
Top 10 (agreement)						
cos_e	-	-	-	124.88	62.92	-
cosmos_qa	-	-	-	254.07	226.66	-
dream	-	-	-	110.02	78.65	-
qasc	-	-	-	2.04	1.64	-
quail	-	-	-	225.35	144.74	-
quarel	-	-	-	29.81	21.06	-
quartz	-	-	-	15.27	9.50	-
sciq	-	-	-	30.51	10.33	-
social_i_qa	-	-	-	211.44	208.51	-
wiqa	-	-	-	682.20	466.81	-
Average	-	-	-	168.56	123.08	-
All (log prob)						
cos_e	12.15	5.68	-	133.30	107.24	-
cosmos_qa	11.27	9.85	-	266.24	367.09	-
dream	1.16	1.65	-	48.56	69.55	-
qasc	1.03	1.54	-	0.76	2.90	-
quail	10.64	11.11	-	160.07	232.35	-
quarel	0.94	2.22	-	27.25	30.66	-
quartz	1.26	2.40	-	10.01	13.33	-
sciq	0.97	1.11	-	11.45	6.16	-
social_i_qa	12.10	8.92	-	216.16	214.50	-
wiqa	10.43	7.85	-	661.90	655.91	-
Average	6.20	5.23	-	153.57	169.97	-
All (agreement)						
cos_e	-	-	-	125.71	51.60	-
cosmos_qa	-	-	-	234.45	230.02	-
dream	-	-	-	92.89	65.07	-
qasc	-	-	-	0.59	0.80	-
quail	-	-	-	249.17	137.03	-
quarel	-	-	-	29.46	23.78	-
quartz	-	-	-	11.49	9.41	-
sciq	-	-	-	28.44	4.61	-
social_i_qa	-	-	-	208.03	209.52	-
wiqa	-	-	-	662.53	453.35	-
Average	-	-	-	164.28	118.52	-

Table 5: Expected Calibration Error (ECE) and Swapped Pairs results for Automatically Generated Prompts (AGP) (either 10 or all) by model (T0++, FLAN-T5-XXL, GPT-3), confidence estimation method (log probability or agreement), and dataset.

Confidence Method (Human Prompts)	Accuracy T0++	FLAN-T5-XXL	GPT-3
Single (max log-prob/agreement)			
cos_e	0.59	0.38	0.56
cosmos_qa	0.72	0.60	0.48
dream	0.80	0.70	0.88
qasc	0.99	0.95	0.76
quail	0.71	0.67	0.58
quarel	0.59	0.60	0.70
quartz	0.85	0.68	0.62
sciq	0.77	0.62	0.30
social_i_qa	0.34	0.33	0.35
wiqa	0.49	0.54	0.39
Average	0.69	0.61	0.56
Multiple (max log-prob)			
cos_e	0.73	0.68	0.72
cosmos_qa	0.68	0.59	0.51
dream	0.85	0.81	0.82
qasc	0.98	0.96	0.80
quail	0.77	0.77	0.62
quarel	0.59	0.63	0.54
quartz	0.87	0.86	0.75
sciq	0.87	0.88	0.86
social_i_qa	0.63	0.59	0.48
wiqa	0.65	0.59	0.38
Average	0.76	0.74	0.65
Multiple (agreement)			
cos_e	0.75	0.75	0.77
cosmos_qa	0.82	0.75	0.61
dream	0.85	0.84	0.83
qasc	0.99	0.99	0.86
quail	0.79	0.80	0.66
quarel	0.60	0.63	0.56
quartz	0.89	0.92	0.78
sciq	0.94	0.96	0.91
social_i_qa	0.71	0.70	0.52
wiqa	0.65	0.63	0.37
Average	0.80	0.80	0.69

Table 6: Accuracy by model (T0++, FLAN-T5-XXL, GPT-3) and dataset for Human Prompts (single or multiple), where the prediction is either the label with the maximum log probability or the majority label (agreement). Note that because the Single prompt setting contains only one prompt, Single (max log-prob) and Single (agreement) result in the same accuracy.

Confidence Method (AGP)	Accuracy T0++	FLAN-T5-XXL	GPT-3
Top 10 (max log-prob)			
cos_e	0.62	0.72	-
cosmos_qa	0.72	0.65	-
dream	0.85	0.85	-
qasc	0.99	0.98	-
quail	0.66	0.70	-
quarel	0.61	0.63	-
quartz	0.88	0.88	-
sciq	0.92	0.95	-
social_i_qa	0.34	0.34	-
wiqa	0.63	0.71	-
Average	0.72	0.74	-
Top 10 (agreement)			
cos_e	0.62	0.76	-
cosmos_qa	0.73	0.67	-
dream	0.85	0.85	-
qasc	0.99	0.99	-
quail	0.66	0.71	-
quarel	0.62	0.65	-
quartz	0.88	0.89	-
sciq	0.92	0.95	-
social_i_qa	0.34	0.34	-
wiqa	0.64	0.73	-
Average	0.72	0.75	-
All (max log-prob)			
cos_e	0.61	0.71	-
cosmos_qa	0.71	0.60	-
dream	0.85	0.83	-
qasc	0.99	0.98	-
quail	0.65	0.66	-
quarel	0.60	0.61	-
quartz	0.86	0.83	-
sciq	0.91	0.94	-
social_i_qa	0.34	0.34	-
wiqa	0.62	0.70	-
Average	0.71	0.72	-
All (agreement)			
cos_e	0.62	0.74	-
cosmos_qa	0.72	0.62	-
dream	0.85	0.85	-
qasc	0.99	0.99	-
quail	0.66	0.67	-
quarel	0.61	0.64	-
quartz	0.87	0.89	-
sciq	0.92	0.94	-
social_i_qa	0.34	0.34	-
wiqa	0.64	0.73	-
Average	0.72	0.74	-

Table 7: Accuracy by model (T0++, FLAN-T5-XXL, GPT-3) and dataset for Automatically Generated Prompts (either top 10 or all), where the prediction is either the label with the maximum log probability or the majority label (agreement).

Table 8: Prompt Generation Prompts that are fed to GPT-3 in order to generate prompts. The PGP in row 1 is taken from Zhou et al. (2022b).

ID	Prompt Generation Prompt
1	Generate a variation of the following instruction while keeping the semantic meaning. Input: {{ instructions }} Output:
2	What's another way of saying "{{ instructions }}" while keeping the same semantic meaning? Output:
3	Rephrase the following instructions while keeping the same semantic meaning. Input: {{ instructions }} Output:
4	Can you tell me another way of saying the following instructions while keeping the semantic meaning? Input: {{ instructions }} Output:
5	Paraphrase the following instructions while keeping the same semantic meaning. Input: {{ instructions }} Output:
6	Tell me another way of stating "{{ instructions }}" while keeping the same semantic meaning. Output:
7	How can I rephrase the instructions "{{ instructions }}" while keeping the same semantic meaning? Output:
8	Give me a sentence that expresses the following instructions in different words. Input: {{ instructions }} Output:
9	Generate a variation of the following instruction. Input: {{ instructions }} Output:
10	What's another way of saying "{{ instructions }}"? Output:
11	Rephrase the following instructions. Input:{{ instructions }} Output:
12	Can you tell me another way of saying the following instructions? Input: {{ instructions }} Output:
13	Paraphrase the following instructions. Input: {{ instructions }} Output:
14	Tell me another way of stating "{{ instructions }}". Output:

Continued on next page

Table 8 – continued from previous page

ID	Prompt Generation Prompt
15	How can I rephrase the instructions "{{ instructions }}"? Output:

Table 9: Instructions from T0 prompts (Sanh et al., 2021) that were used to generate new prompts.

Dataset	Instruction
CoS-E v1.11	Pick the option in line with common sense to answer the question.
CoS-E v1.11	Choose the most suitable option to answer the above question.
CoS-E v1.11	The best answer is
Cosmos QA	According to the above context, choose the best option to answer the following question.
Cosmos QA	According to the above context, answer the following question.
Cosmos QA	Pick the best answer from the following options
Cosmos QA	Read the following context and choose the best option to answer the question.
Cosmos QA	Read the following context and answer the question.
DREAM	Read the following conversation and answer the question.
QASC	Given the two facts above, answer the question with the following options:
QASC	You are presented with the question and the following answer choices. Now knowing the facts, choose the best answer.
QASC	You are presented with the quiz. But you don't know the answer, so you turn to your teacher to ask for hints. He says the following facts. So, what's the best answer to the question?
Quail	According to the above context, choose the correct option to answer the following question.
Quail	The correct answer is
Quail	Pick the correct answer from the following options
Quail	Read the following context and choose the correct option to answer the question.
Quarel	Choose between "X" and "Y".
Quarel	Do not use A and B to answer the question but instead, choose between "X" and "Y".
Quarel	What is the most sensical answer between "X" and "Y"?
Quarel	Choose the answer between "X" and "Y".
Quarel	I am testing my students' logic. What is the answer they should choose between "X" and "Y"?
Quartz	Answer the question based on the following text.
Quartz	Answer the question below
Quartz	Given the facts below, answer the question
Quartz	Having read the above passage, choose the right answer to the following question
Quartz	Read the passage below and choose the right answer to the following question
Quartz	Use information from the paragraph to answer the question.
SciQ	Answer the following question given this paragraph
SciQ	Read this paragraph and choose the correct option from the provided answers:
Social IQA	Which one of these answers best answers the question according to the context?
WIQA	How does the supposed perturbation influence the second effect mentioned? Answer by more, less or no effect.

Dataset	Hugging Face URL	Validation Size	Valid GPT-3 Predictions for OP	Valid GPT-3 Predictions for MH
CoS-E v1.11	https://huggingface.co/datasets/cos_e/	1221	947	838
Cosmos QA	https://huggingface.co/datasets/cosmos_qa/	2985	2974	2624
DREAM	https://huggingface.co/datasets/dream	2040	2040	1943
QASC	https://huggingface.co/datasets/qasc	926	796	461
Quail	https://huggingface.co/datasets/quail	2164	2141	1917
Quarel	https://huggingface.co/datasets/quarel	278	277	182
Quartz	https://huggingface.co/datasets/quartz	384	211	162
SciQ	https://huggingface.co/datasets/sciq	1000	991	521
Social IQA	https://huggingface.co/datasets/social_i_qa	1954	1751	872
WIQA	https://huggingface.co/datasets/wiqa	6894	6894	6172

Table 10: Dataset information: Hugging Face URL, size of validation split, number of instances that GPT-3 generated valid predictions for on the official prompt (OP), and number of instances that GPT-3 generated valid predictions for across all Multiple Human (MH) prompts.

Dataset	Unique Generated Paraphrases	Final Number of Paraphrased Prompts
CoS-E v1.11	48	48
Cosmos QA	121	50
DREAM	19	19
QASC	98	50
Quail	89	50
Quarel	75	39
Quartz	158	50
SciQ	61	50
Social IQA	25	25
WIQA	16	16

Table 11: The total number of unique paraphrased instructions and the final number of paraphrased prompts (up to 50 per dataset).

Table 12: Automatically generated instructions for CoS-E v1.11.

ID	Instruction
1	Choose the option that makes the most sense to answer the question.
2	Choose the most logical answer to the question.
3	Choose the most practical option to answer the question.
4	Choose the answer that makes the most sense.
5	Choose the option that best answers the question.
6	What is the best answer to the question above?
7	Select the best option to answer the question.
8	Select the option that best answers the question.
9	What is the best answer to the question?
10	Select the best answer for the question above.
11	Choose the best option to answer the question.
12	What is the best option to answer the question?
13	Please select the option that best answers the question.
14	Choose the best answer to the question above.
15	The most correct answer is
16	One possible answer is...
17	The most accurate answer is
18	The most accurate answer is, The most precise answer is
19	What is the best answer?
20	The most accurate answer is the one that is closest to the correct answer
21	Choose the option that most makes sense to answer the question.
22	Choose the most sensible option to answer the question.
23	What is the best option to answer the question above?
24	Pick the best option to answer the question.
25	Select the most appropriate response to the question above.
26	What is the most suitable option to answer the above question?
27	Please select the option which you believe best answers the question.
28	Pick the best answer for the question above.
29	What is the best response to the question?
30	The answer that is most advantageous/ beneficial/ favorable is the best answer
31	The most correct answer is the one that is closest to the answer key
32	The answer that is most accurate or precise is the best answer.
33	The most ideal answer is.
34	Choose the option that seems most reasonable to answer the question.
35	Choose the answer that makes the most sense given the question.
36	The most sensible answer to the question is the one you should choose.
37	Pick the option that you think makes the most sense to answer the question
38	The most logical answer to the question is the best option.
39	Please select the option which you believe is the most sensible answer to the
40	From the given options, select the one that best answers the question.
41	Select the option that best responds to the question.
42	From the options below, select the one that best responds to the question
43	There is more than one correct answer to the question. Please choose the
44	Pick the best option to respond to the question.
45	Choose the most appropriate option to answer the question.

Continued on next page

Table 12 – continued from previous page

ID	Instruction
46	The most optimal answer is
47	Choose the answer that you think is most correct.
48	The most correct answer is the one that is most accurate and precise.

Table 13: Automatically generated instructions for Cosmos QA.

ID	Instruction
1	Read the following context and answer the question below.
2	Read the following context and answer the question below. What does
3	Choose the option that best answers the question based on the context above.
4	Read the text below and answer the question that follows.
5	Choose the most correct answer from the following options.
6	What does the author say about the best option?
7	What is the author's purpose in writing this text?
8	Which of the following is the best option to answer the question?
9	After reading the context, answer the question.
10	In light of the information provided, please answer the following question.
11	What is the main idea of the text? The main idea
12	Please choose the option that best answers the question.
13	Choose the option that best answers the question based on the information given.
14	Read the following text and choose the best option to answer the question.
15	Read the context and choose the best option to answer the question.
16	To complete this task, read the text and then choose the best answer
17	What is the most important advice from the text? The most
18	What is the best option to answer the following question, based on the
19	Please select the option which you think is correct, based on the context
20	Based on the information given, select the most appropriate response.
21	Please select the option that you believe best answers the question based on the
22	Assuming you want a similar phrase with different words: Please
23	Choose the most correct answer from the given choices.
24	What is the answer to the question, based on the context above?
25	Please read the following information and select the best option to answer the question
26	According to the context above, choose the best option to answer the following
27	In light of the information given, please answer the following question.
28	Choose the most suitable answer from the given choices.
29	What is the author's opinion on the matter?
30	Please read the following information and answer the question that follows.
31	Choose the best answer from the following options.
32	Choose the most accurate answer from the given choices.
33	Based on the information given, answer the following question.
34	Read the text below and answer the question.
35	In light of the above information, select the most appropriate response to the
36	Please answer the following question given the context above.
37	Choose the most correct answer from the following choices.
38	What are the instructions asking you to do? Read the following
39	Read the text below and select the best answer to the question.
40	What does the author say about the relationship between the two countries?
41	What is the best answer from the following options?

Continued on next page

Table 13 – continued from previous page

ID	Instruction
42	What's the best answer from the following options?
43	What is the main idea of the text? The text is
44	What is the best option?
45	Read the following context and select the best option to answer the question.
46	Choose the best option to answer the question based on the following context.
47	Please select the option that best answers the question based on the context above
48	Please read the following text and select the best answer to the question below
49	Please answer the following question based on the information given above.
50	Choose the best option to answer the question based on the context provided.

Table 14: Automatically generated instructions for DREAM.

ID	Instruction
1	Please read the following conversation and answer the question.
2	What does the following conversation reveal about the speaker?
3	Read the following conversation and answer the question below. Who is
4	Read the conversation below and answer the question.
5	What is the conversation about? What is the question about?
6	What is the conversation about?
7	Read the following conversation and then answer the question.
8	Please read the conversation below and answer the question that follows.
9	Read the conversation below and answer the question. Who is the
10	Read the following conversation and answer the question.
11	Read the following conversation and answer the question. Who is speaking
12	What is the conversation about? What is the main topic of
13	What does the conversation below reveal about the speaker? Read the
14	Read the following conversation and answer the question below. Two friends
15	Please read the conversation and answer the question.
16	Read the following conversation and answer the question. Who is the
17	Read the conversation and answer the question.
18	What is the next line in the conversation?
19	Read the conversation below and answer the question. At what time

Table 15: Automatically generated instructions for QASC.

ID	Instruction
1	You are given the quiz, but you are unsure of the answers.
2	Based on the information provided, please select one of the following options:
3	You are presented with a quiz, and you don't know the answer
4	You are given a question and the following answer choices. Choose the best
5	Choose the best answer from the given choices, based on the information given
6	Choose the best answer from the given choices that best aligns with the
7	The teacher said that the answer to the question is one of the following
8	Taking into account the two facts mentioned above, please select one of the
9	Given the two facts above, please answer the question with one of the
10	1. Given the two facts, answer the question with the following options
11	You are taking a quiz and you don't know the answer to
12	What is the best way to answer the question given the two facts?
13	Choose one of the following options: - A - B
14	You are given the quiz, but you are unsure of the answer.
15	Choose one of the following options based on the two facts given above.
16	Choose one of the following options that best answers the question:
17	Choose the answer that best fits the question, based on the information given
18	Assuming that the average person sleeps eight hours a day, how long will
19	The teacher provides you with the following information: You are presented
20	Choose one of the following answers: A) The moon orbits
21	Based on the information given, select one of the following options:
22	Assuming the two facts are true, which of the following is most likely
23	Based on the two facts provided, please select from the following options to
24	You are given the question and the following answer choices. With the information
25	Assuming the two facts above, answer the question with the following options:
26	You are taking a quiz and don't know the answer to one of
27	What is the most likely explanation for the data? -The
28	Given the question and the following answer choices, select the most accurate answer
29	What is the probability that the person is a Democrat? What
30	The teacher provides you with the following information to help you answer the question
31	You are given a quiz, but you are unsure of the answers.
32	You are given a quiz, but you don't know the answer.
33	You are presented with a quiz, but you are unsure of the answer
34	Read the question and the answer choices carefully, then select the most correct
35	Assuming the aforementioned facts are accurate, please select from the following options:
36	What is the probability that the box contains a white ball?
37	What is the result of subtracting 4 from 9? -
38	You are taking a quiz and are unsure of the answer to one of
39	Choose the best answer from the given choices.

Continued on next page

Table 15 – continued from previous page

ID	Instruction
40	Choose one of the following options that best answers the question based on the
41	You are presented with the quiz. But you don't know the answer
42	Choose one of the following options: a) The moon orbits
43	Given the question and the following answer choices, select the most accurate response
44	What's the best answer to the question, given the following facts?
45	Choose the best answer from the given choices that best fits the question.
46	Choose the answer that best fits the question based on the given information.
47	Choose the answer that best fits the question, based on the given information
48	You are given the question and the following answer choices. Choose the best
49	What is the conclusion based on the two facts?
50	What is the best answer given the question and the following answer choices?

Table 16: Automatically generated instructions for Quail.

ID	Instruction
1	Choose the option that best answers the question given the context.
2	The answer is right.
3	Choose the option that best answers the question based on the information given.
4	What does the author say about the relationship between the two countries?
5	Choose the option that best answers the question below, based on the
6	Choose the correct option to answer the following question, based on the context
7	Please read the following text and select the appropriate answer to the question.
8	The right answer is
9	Choose the option that best answers the question based on the information given in
10	Choose the correct answer from the following options.
11	Based on the information given, select the option that best answers the question
12	Read the following context and choose the best answer to the question.
13	In light of the context above, please select the most appropriate option to
14	Choose the right option from the given choices.
15	After reading the text, select the option that best answers the question.
16	Choose the option that answers the question based on the context above.
17	The answer you are looking for is.
18	After reading the following context, select the option that best answers the question
19	Choose the correct option to answer the question based on the context.
20	Choose the option that best answers the question based on the context.
21	What is the best answer from the following options?
22	Read the text above and then select the best answer to the following question
23	Read the context below and choose the best answer to the question.
24	Select the option that best answers the question.
25	Read the following context and then select the best answer to the question.
26	Read the following context and choose the correct option to answer the question.
27	Choose the correct option to answer the following question based on the context above
28	Based on the context above, select the appropriate option to answer the question
29	Based on the information provided, select the most appropriate answer to the question
30	Given the context above, please select the most appropriate answer to the
31	In reference to the text above, please select the appropriate response to the
32	Choose the option below that best answers the question based on the context above
33	In reference to the context above, select the most appropriate response to the
34	The answer that is correct is
35	What is the best answer to the following question?
36	Read the following context and choose the best option to answer the question.
37	Choose the option that best answers the question.
38	Choose the right response from the given choices.
39	Read the provided context and select the option that best answers the question.
40	The answer you are looking for is correct.
41	Select the option which best answers the question based on the information provided.
42	Choose the most accurate response from the given choices.
43	Which of the following options best completes the sentence? I'm
44	In the context above, please select the most appropriate response to the following
45	Select the most appropriate answer from the following choices.

Continued on next page

Table 16 – continued from previous page

ID	Instruction
46	Choose the answer that best fits the context.
47	The answer that is correct is the one that you should select.
48	Read the following context and select the option that best answers the question.
49	What is the best way to respond to the following question?
50	What does the author mean by "a variation of the following instruction?

Table 17: Automatically generated instructions for Quarel. All Quarel prompts written for T0 (Sanh et al., 2021) incorporate the multiple choice options into the instruction (e.g., “Choose between X and Y”), so when generating prompts for Quarel, we exclude generated prompts that do not include two placeholders, X and Y.

ID	Instruction
1	You can have either "X" or "Y".
2	What is your preference between "X" and "Y"?
3	You can choose either "X" or "Y".
4	You can either choose "X" or "Y".
5	Pick either "X" or "Y".
6	You can either have "X" or "Y".
7	Choose either "X" or "Y".
8	Please choose either "X" or "Y".
9	Pick "X" or "Y".
10	You have the option of choosing either "X" or "Y".
11	What would you like to do, "X" or "Y"?
14	Choose between "X" and "Y" to answer the question,
15	Choose "X" or "Y", but not "A" and
16	Choose between "X" and "Y" instead of using A and
19	"X" or "Y"?
20	Please choose either "X" or "Y" to answer the question
22	Use either "X" or "Y" to answer the question,
23	Choose either "X" or "Y" to answer the question,
24	What is the most logical answer between "X" and "Y"?
25	What is the most reasonable answer between "X" and "Y"?
26	What is the most sensible answer between "X" and "Y"?
27	Which of "X" and "Y" is the most reasonable answer
28	Select the answer between "X" and "Y".
29	Select the response either "X" or "Y".
30	Choose the answer between "X" and "Y".
31	Select the answer from the options "X" and "Y".
32	Please pick one of the following options: "X" or "Y"
33	Select the correct response from "X" or "Y".
34	Choose either "X" or "Y" as your answer.
35	Select either "X" or "Y".
36	Choose between "X" and "Y".
37	What is your choice between "X" and "Y"?
38	Select either "X" or "Y" as your answer.
39	What is the correct answer between "X" and "Y"?
41	What is the answer they should choose between "X" and "Y"
42	What is the correct answer between "X" and "Y" from
43	What should the answer be between "X" and "Y" when
44	What is the correct answer, "X" or "Y"?
47	What is the difference between "X" and "Y"?

Table 18: Automatically generated instructions for Quartz.

ID	Instruction
1	Respond to the question using the given text as reference.
2	Read the passage below and choose the best answer to the following question.
3	Using the information given below, answer the question.
4	What does the author think about people who are good at math?
5	Based on the information given, please answer the question.
6	What is the answer to the question, based on the information provided
7	After reading the passage, choose the best answer to the question.
8	Read the passage above and then select the correct answer to the question below
9	What is the answer to the question, given the following facts?
10	After reading the text, select the most appropriate answer to the question below
11	What is your favorite color?
12	What information from the paragraph can you use to answer the question?
13	Respond to the question below.
14	Based on the following text, answer the question.
15	The facts are as follows: -The average person needs about
16	Facts: 1. Lisa is taller than Sarah.
17	The question can be answered using information from the paragraph.
18	Read the passage below and choose the right answer to the following question.
19	Assuming the information given is true, answer the question.
20	What does the author say about the relationship between the sun and Jupiter?
21	Read the passage below and choose the answer to the question that best completes
22	What is your answer to the question below?
23	Respond to the question using the given information.
24	What is the main idea of the following text?
25	Given the facts that it is currently snowing outside and the temperature is
26	What is the capital of France? The capital of France is
27	Read the text and select the correct response to the question.
28	After reading the passage, please select the correct answer to the following question
29	What is the author's view on the relationship between the two countries?
30	Read the passage below and then select the best answer to the question that
31	After reading the passage, select the answer that best responds to the
32	Based on the text, answer the following question.
33	What is the probability of drawing two cards from a standard deck of cards
34	Read the passage and then select the answer that best fits the question.
35	What does the text say about the author's feelings? The
36	Facts: John is taller than Bill. Bill is
37	Refer to the paragraph for guidance in answering the question.
38	Choose the right answer to the following question based on the passage you just
39	In order to answer the question, use the information found in the paragraph
40	Refer to the paragraph for the answer to the question.

Continued on next page

Table 18 – continued from previous page

ID	Instruction
41	Below are the facts. Please answer the question based on them.
42	What is the probability of being dealt a flush in poker?
43	Please provide an answer to the question based on the text you have been
44	Skim the passage for the answer to the following question.
45	Choose the right answer to the following question, after reading the passage above
46	What is the author's purpose in writing the text?
47	Answer the question based on the information provided.
48	What is the value of X? X is the value of
49	Choose the right answer to the following question, having read the passage above
50	What can you infer from the text?

Table 19: Automatically generated instructions for SciQ.

ID	Instruction
1	What is the main idea of the paragraph? What is the
2	What is the question that needs to be answered based on the given paragraph
3	What is the question that must be answered based on the given paragraph?
4	What does the author say about the relationship between the two countries?
5	Read the paragraph and choose the correct option from the answers provided.
6	Read the following paragraph and select the most appropriate answer from the given options
7	What is the question that must be answered given the paragraph?
8	What is the question that you need to answer based on the given paragraph
9	Choose the correct option from the provided answers that best completes the following
10	Read the paragraph and select the best answer from the given options.
11	What is the main idea of this paragraph?
12	Please read the following paragraph and select the most accurate response from the
13	Choose the correct option from the provided answers that best completes the paragraph
14	What is the main idea of the paragraph?
15	What does the author say about the book? The author says
16	Read the paragraph and select the most appropriate answer from the given options
17	What is the author's purpose in writing this paragraph?
18	Please answer the question below based on the given paragraph.
19	Which of the following best completes the sentence?
20	What is the author's opinion of the book? The author
21	Please read the following paragraph and select the most appropriate response from the
22	Please read the following paragraph and choose the best answer from the given options
23	What is the main idea of the paragraph? After reading the
24	What does the author say about the benefits of a plant-based diet
25	Please read the following paragraph and then select the most appropriate answer from the
26	Read the following paragraph and select the most appropriate response from the given choices
27	Choose the option that best completes the paragraph: There are four
28	Please read the paragraph and choose the most appropriate answer from the given options
29	What is the main idea of the paragraph? The paragraph is
30	What does the author say about the benefits of studying abroad?
31	What does the author say about the role of government in a market economy
32	Choose the correct answer from the options provided below the paragraph.
33	What is the question that must be answered given the following paragraph?
34	Choose the correct option from the provided answers that best completes the following paragraph
35	Read the paragraph and select the correct option from the provided answers.
36	Read this paragraph and choose the best option from the given answers.
37	In what ways does the author use pathos in the essay?
38	Read this paragraph and select the most appropriate option from the given choices.

Continued on next page

Table 19 – continued from previous page

ID	Instruction
39	Read the paragraph and select the most accurate answer from the given choices.
40	Read the following paragraph and choose the option that best answers the question.
41	What is the question that you must answer based on the given paragraph?
42	Based on the paragraph, answer the following question.
43	Give an answer to the following question based on the given paragraph.
44	Read the following paragraph and choose the best answer from the provided options:
45	What is the main idea of the passage? The main idea
46	What is the main point the author is making in the paragraph?
47	Select the correct answer from the provided options after reading the following paragraph.
48	Read the paragraph below and choose the best answer from the provided options.
49	Read the paragraph and choose the best answer from the provided options.
50	What does the paragraph say about the author's feelings towards his work?

Table 20: Automatically generated instructions for Social IQA.

ID	Instruction
1	What is the best answer to the question according to the context?
2	What is the most accurate answer to the question given the context?
3	Which of these answers best answers the question according to the context?
4	What is the most accurate response to the question given the context?
5	Which of these answers is most relevant to the question?
6	Which of these answers is best according to the context?
7	Which answer is the most accurate for the question given the context?
8	Which answer is the most relevant to the question?
9	Which answer provides the best response to the question?
10	Which one of these answers is most relevant to the question?
11	Which answer best fits the context of the question?
12	What is the most accurate response to the question?
13	Which answer is most relevant to the question?
14	Which one of these answers is the most accurate in relation to the question
15	Which answer best fits the question's context?
16	Which answer best responds to the question in the given context?
17	Which option provides the most accurate response to the question?
18	What is the most accurate answer to the question?
19	Which one of these best answers the question according to the context?
20	Which of these answers is the most relevant to the question at hand?
21	Which of these answers best fits the question's context?
22	Which answer best suits the question?
23	What is the most appropriate answer to the question?
24	Which answer best responds to the question given the context?
25	Which of these answers most accurately responds to the question given the surrounding context

Table 21: Automatically generated instructions for WIQA.

ID	Instruction
1	How does the supposed perturbation influence the second effect mentioned? Answer
2	What is the extent to which the supposed perturbation influences the second
3	What is the supposed effect of the perturbation on the second mentioned
4	What is the supposed perturbation's effect on the second mentioned effect
5	What effect does the supposed perturbation have on the second mentioned effect
6	What is the supposed perturbation?
7	What is the extent to which the supposed perturbation affects the second
8	How does the supposed perturbation influence the second effect mentioned? More
9	What is the supposed perturbation's influence on the second effect?
10	What is the supposed impact of the perturbation on the second mentioned
11	To what extent does the supposed perturbation affect the second mentioned outcome
12	What is the expected effect of the perturbation on the second mentioned
13	To what extent does the supposed perturbation influence the second effect mentioned
14	Does the supposed perturbation have more, less, or no effect
15	What is the supposed effect of the perturbation on the second effect
16	What is the nature of the supposed perturbation's influence on the