

Table 4: Multinomial beam search sampling produces sampled answers that are less diverse and thus less useful for uncertainty estimation than multinomial sampling.

Sampling method	Semantic Entropy AUROC	Diversity of answers
Multinomial sampling	0.758	0.490
Multinomial beam search sampling	0.735	0.258

## B EXPERIMENTAL DETAILS AND ABLATIONS

We use both the OPT models<sup>2</sup> and the Deberta-large model<sup>3</sup> via the HuggingFace transformers library which can be easily adopted for reproducibility. All of our code is open-source and relies on no proprietary models.

We use the following functions of the HuggingFace API to sample the most likely answers, and the set of answers:

- To obtain the answer which is compared to the reference answer, which determines whether the question is correctly answered, we use beam search using the `generate()` function with `num_beams = 5` and `do_sample = True`.
- To obtain the answer set for uncertainty estimation, by default we use multinomial sampling, that is `generate()` using `do_sample = True` and `num_beams = 1`. If indicated explicitly, we use beam multinomial sampling, that is `generate()` using `num_beams = 5` and `do_sample = True`.

We run all of our experiments on 80GB NVIDIA A100s.

Testing up to 20 samples per answer on the 2.7B, 6.7B and 13B CoQA experiments, we conclude that using more than 10 samples does not significantly improve the performance of the uncertainty measure, we use 10 sampled answers per question in the remaining experiments on TriviaQA. Note, that in Table 2 we compare the 30B model on CoQA and TriviaQA where in both settings we use answer sets of size 10.

We use the following prompts on CoQA and TriviaQA. We find that on CoQA, we obtain accurate model results with zero-shot prompting. While we have to use few-shot prompting to obtain accurate answers on closed-book TriviaQA. We use the following prompts for each of the settings:

### CoQA:

[The provided context paragraph]  
 [additional question-answer pairs]  
 Q: [Provided question]  
 A:

where additional question-answer pairs are preceding turns of the conversation about the paragraph consisting of questions and reference answers.

### TriviaQA:

For TriviaQA, we use a 10-shot prompt of the format:

Q: Which Oscar-nominated film had You Sexy Thing as its theme song? A: The Full Monty Q: Which Joan's career revived in Whatever Happened to Baby Jane? A: Crawford Q: Which much-loved actor won the Best Actor Oscar for The Philadelphia Story? A: James Stewart (...) Q: In which river is the Boulder Dam? A:

To account for generations where the model continues the Q: ... A: ... pattern after providing an answer to the given question, we trim all generations by pattern matching for a selection of stop-words that we observe in the generations: Q:, Question:, QUESTION: and questions:.

<sup>2</sup>[https://huggingface.co/docs/transformers/model\\_doc/opt](https://huggingface.co/docs/transformers/model_doc/opt)

<sup>3</sup>[https://huggingface.co/docs/transformers/model\\_doc/opt](https://huggingface.co/docs/transformers/model_doc/opt)

Table 5: **Automatic evaluation of question answering is highly accurate as compared to human evaluation.** We evaluate how accurate the automatic evaluation metric. The predictions, in this settings are the automatically determined accuracy labels on our question answering task, and the ground truth are human labels for the accuracy of the provided model generation given the reference answer

Data set	Accuracy of automatic evaluation
CoQA	0.89
TriviaQA	0.96

Table 6: TriviaQA: the exact choice of accuracy metric for the free-form QA task has little effect on the assessment of the quality of the uncertainty measure.

Metric	AUROC		Accuracy
	Semantic entropy	Normalised entropy	
$\text{Rouge-L}(y, y') > 0.3$	0.828	0.802	0.506
$\text{Rouge-L}(y, y') > 0.5$	0.835	0.810	0.456
$\text{Rouge-1}(y, y') > 0.5$	0.835	0.810	0.457
Exact matching	0.828	0.808	0.394

### B.1 RELIABILITY OF ACCURACY METRIC AS COMPARED TO HUMAN EVALUATION

In our experiments, we evaluate how well our uncertainty measures predict the model’s accuracy when answering a given question. The choice of accuracy metric is thus a crucial component of our experimental setup. Generally, it has been shown to be difficult to develop automatic metrics for free-form generation that correlate well with human evaluations. We thus verify our choice of accuracy criterion:  $\text{Rouge-L}(y, y') > 0.3$ , for a given reference answer  $y$  and a model generation  $y'$ . We manually evaluate the accuracy of 200 answers of the 30B parameter model on both CoQA and on TriviaQA, and evaluate how closely the human evaluation matches the automatic evaluation. We find that on both data sets, the accuracy of the automatic labels as compared to the human labels as the ground truth is high, see Table 5.

### B.2 TESTING THE BI-DIRECTIONAL ENTAILMENT CLASSIFIER

To the best of our knowledge, this paper is the first application of the bi-directional entailment approach to identifying answers with the same meaning in question answering. Since this is a core component of our approach, we verify how accurately this approach identifies model answers with the same meaning. To this end, we manually label 300 samples for each of TriviaQA and CoQA produced by the 13B parameter model to provide a ground truth as to whether or not they mean the same thing. We find that the model achieves an accuracy of 92.7% and 95.3% respectively.

### B.3 SENSITIVITY OF RESULTS TO ACCURACY METRIC

In principle, the choice of metric to decide whether or not an answer is ‘correct’ might have a large effect on the assessment of our method and baselines. However, we find empirically that our results are relatively insensitive to the choice of accuracy metric.

In Table 6 we show that for TriviaQA the choice of accuracy metric for the question answering has almost no effect on the measured AUROC of the uncertainty estimation, despite making the measured accuracy of the model’s generation significantly different. In particular, the exact matching requirement reduces the accuracy significantly but has little effect on the AUROCs.

For CoQA, which is an open-book QA task with greater answer variability and longer answers the results are broadly similar (see Table 7) except for the exact matching accuracy criterion which is too demanding because of the much larger variety of possible answers for this task.

Table 7: CoQA: the exact choice of the accuracy metric for the free-form open-book QA task has little effect on the assessment of the quality of the uncertainty measure except for the use of exact matching. For CoQA, getting an exact match is significantly harder.

Metric	AUROC		Accuracy
	Semantic entropy	Normalised entropy	
Rouge-L( $y, y'$ ) > 0.3	0.7672	0.7533	0.8239
Rouge-L( $y, y'$ ) > 0.5	0.7379	0.7290	0.7657
Rouge-1( $y, y'$ ) > 0.3	0.7672	0.7533	0.8239
Rouge-1( $y, y'$ ) > 0.5	0.7397	0.7309	0.7677
Exact matching	0.6749	0.6727	0.6459

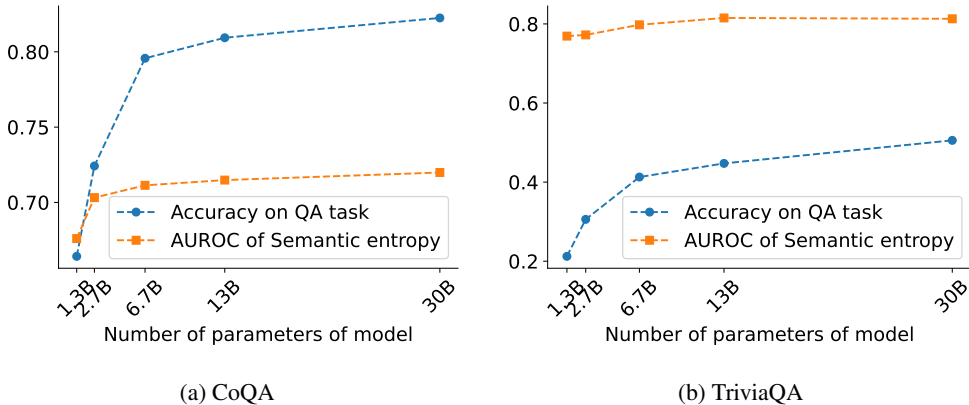


Figure 5: Accuracy improves with model size, as does semantic entropy’s uncertainty performance. At the smallest model size, both accuracy and uncertainty diminish.

#### B.4 ACCURACY ABLATIONS WITH MODEL SIZE

We confirm that increasing the model size improves the accuracy of the generations on both QA datasets (see Fig. 5a and Fig. 5b). Semantic entropy’s uncertainty performance is also shown for context.

#### B.5 EXAMPLE P(TRUE) FORMAT

The format of the prompt, reproduced here for convenient reference from the original source Kadavath et al. (2022), is:

Question: Who was the third president of the United States?  
 Here are some brainstormed ideas: James Monroe  
 Thomas Jefferson  
 John Adams  
 Thomas Jefferson  
 George Washington  
 Possible Answer: James Monroe  
 Is the possible answer:  
 (A) True  
 (B) False  
 The possible answer is:

where the “brainstormed answers” are from the set of sampled answers  $\mathbb{A}$  and  $P(\text{True})$ , i.e. the likelihood of the next token being `True` is taken as the uncertainty measure. The authors note that doing the above needs to be done in a few-shot manner and does not work well as in a zero-shot format. In our experiments, we use a few-shot prompt with 10 examples.