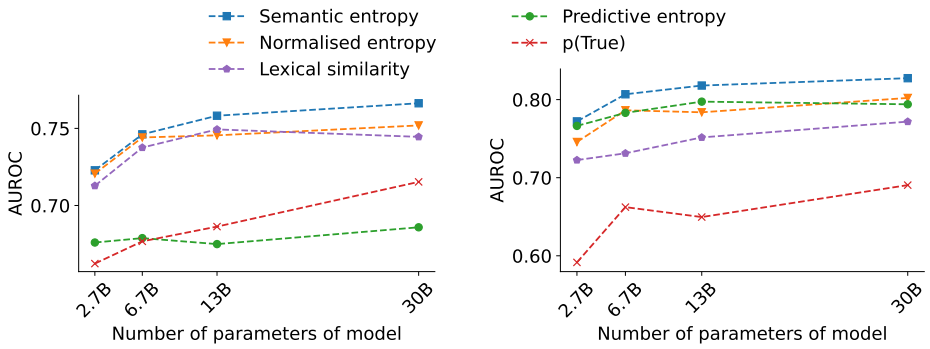


(a) CoQA

(b) TriviaQA

Figure 2: (a) On CoQA open-book question answering semantic entropy demonstrates better uncertainty than ordinary predictive entropy with and without normalisation at larger model sizes. It also performs significantly better than $p(\text{True})$. (b) TriviaQA shows similar results. Identical to Fig. 1b with the addition of $p(\text{True})$, which was previously omitted to avoid stretching the scale.

6 Empirical evaluation

We demonstrate that semantic entropy is an effective way to quantify the uncertainty of NLG on free-form QA tasks. Effective uncertainty measures should offer information about how reliable the model's answers are—that is, very *uncertain* generations should be less likely to be *correct*.

Performance evaluation. Following prior work (e.g. Filos et al. (2019)), we evaluate uncertainty by treating uncertainty estimation as the problem of predicting whether to rely on a model generation for a given context—whether to trust an answer to a question. The area under the receiver operator characteristic curve (AUROC) metric is equivalent to the probability that a randomly chosen correct answer has a higher uncertainty score than a randomly chosen incorrect answer. Higher scores are better, with perfect uncertainty scoring 1 while a random uncertainty measure would score 0.5.

The AUROC is a better measure of uncertainty for *free-form* question answering and NLG than calibration measures like the Brier score, which are often used in classification or for multiple choice QA. This is because the language model outputs a likelihood for a given token-sequence, but not for an entire meaning. In order to estimate the Brier score, we would need to estimate the entire probability mass assigned to any possible way of saying the correct answer. This is intractable for free form text where we do not have access to probabilities about meanings. In contrast, we can estimate the entropy because it is structured as an expected information, which makes Monte Carlo integration suitable.

Model. We use the GPT-like OPT models (Zhang et al., 2022). We vary the size of the model between 2.7B, 6.7B, 13B and 30B parameters, while our headline results are all reported using the largest computationally feasible model, with 30B parameters. In all cases we use only a single unmodified model. There is no ensembling and no stochastic or Bayesian modification. We chose this because in most cases cutting-edge foundation models are not available as ensembles and are too large to efficiently perform approximate Bayesian inference with. We do not fine-tune these models on TriviaQA or CoQA but use them in their pre-trained form.

Datasets. We use CoQA Reddy et al. (2019) as an open-book conversational question answering problem (in which the model answers a question using a supporting paragraph). We use the development split (∼8000 questions). We also use TriviaQA (Joshi et al., 2017) as a closed-book QA problem (in which the model must answer a question without access to a supporting paragraph). We use a subset of 8000 questions of the training split to match the size of CoQA.

We evaluate correctness of our model's generations on the underlying dataset using the a fuzzy matching criterion: $\mathcal{L}(\mathbf{s}, \mathbf{s}') = \mathbf{1}_{RougeL(\mathbf{s},\mathbf{s}')>0.3}$. That is, we consider an answer $\mathbf{s}$ to be correct if its Rouge-L (Lin & Och, 2004) — a measure of the longest common subsequence — with regards to the reference answer is larger than 0.3. In Appendix B.3 we study other objective functions such as exact matching and Rouge-1 and find our method to be robust to these choices.

Baselines. We compare our method against predictive entropy, length-normalised predictive entropy (Malinin & Gales, 2020), $p(\text{True})$ (Kadavath et al., 2022), and lexical similarity (similar to (Fomicheva et al., 2020)). **Predictive entropy** is a widely used measure of uncertainty in other

Table 2: Incorrectly answered questions have more semantically distinct answers than correct ones. On its own, this count is a reasonable uncertainty measure, though semantic entropy is better.

Dataset	Average # of semantically distinct answers		AUROC	
	Correctly answered	Incorrectly answered	Semantic entropy	# distinct answers
CoQA	1.27	1.77	0.77	0.66
TriviaQA	1.89	3.89	0.83	0.79

domains, and has been used as a baseline without length-normalisation in, e.g., Kadavath et al. (2022). Here, the score is just the predictive entropy of the output distribution as described in Eq. (1). **Length-normalised predictive entropy** divides the joint log-probability of each sequence by the length of the sequence, as proposed by Malinin & Gales (2020) in the case of NLG uncertainty and further discussed in Section 3.3. Note that unlike Malinin & Gales (2020), we use only a single model, not an ensemble, and use multinomial sampling as we do for all other methods. $p(\textbf{True})$ proposed by (Kadavath et al., 2022) as a way to estimate the probability that a model's generation is correct by 'asking' the model if its answer is correct. They propose sampling M answers and constructing a new natural language question using these possible answers as context before asking whether the proposed answer is correct and measuring the probability of the completion being `True`. An example of the format is provided in Appendix B. Note that our implementation of this uses OPT models with up to 30B parameters, while Kadavath et al. (2022) use a proprietary 52B parameter model. We are also limited computationally to 10-shot prompting while the original paper uses 20-shot prompting. **Lexical similarity** uses the average similarity of the answers in the answer set $\mathbb{A}$: $\frac{1}{C}\sum_{i=1}^{|\mathbb{A}|}\sum_{j=1}^{|\mathbb{A}|}\text{sim}\left(s_i, s_j\right)$, where $C = |\mathbb{A}| * (|\mathbb{A}| - 1)/2$, and sim is Rouge-L. Additionally, we evaluate a margin-probability baseline (Lin et al., 2022b) in Appendix B.6, and study why it is not very predictive of model accuracy in this setting. All code and data used in our experiments is available at `https://github.com/lorenzkuhn/semantic_uncertainty`.

6.1 Semantic Entropy uncertainty

For both TriviaQA and CoQA, semantic entropy improves over baselines in predicting whether a model's answer to a question is correct. For TriviaQA, using the largest model we show in Fig. 1a we show that semantic entropy has a significantly higher AUROC than entropy in sequence-probability-space with and without length-normalisation, as well as the lexical similarity baseline. At the same time, it performs dramatically better than $p(\text{True})$. Similarly, we find in Fig. 1b that our method outperforms more for larger model sizes and continues to steadily improve as the model size increases, with the performance of the $p(\text{True})$ baseline added in Fig. 2b (not shown in the opening figure for visual clarity). For CoQA, in Fig. 2a we show that semantic entropy predicts model correctness significantly better than the baselines at larger model sizes.

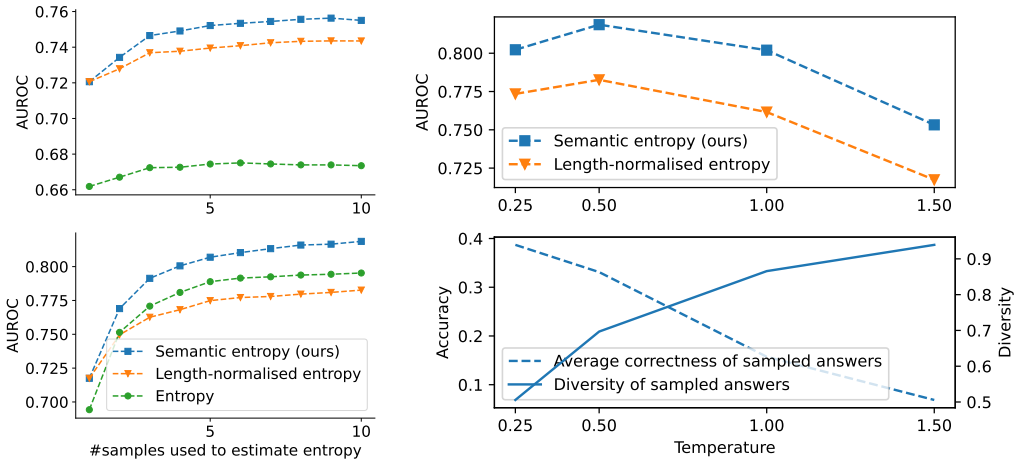
The ground truth answers for TriviaQA are generally single words or very short phrases, while CoQA contains both longer and shorter ground truth answers. This is why performing length-normalisation has a large effect for CoQA but no effect for TriviaQA (compare Fig. 2a and Fig. 2b). TriviaQA is also a more challenging dataset: accuracy of 50.6% against 82.3% for CoQA.

We can better understand the mechanism of action for semantic entropy by examining the difference between incorrect and correct answers generated by the model. In Table 2 we show that the average number of semantically distinct clusters of answers ($|C|$) from the 30B parameter model is significantly greater for incorrectly answered questions than correctly answered ones. This fits our predictions, which is that the model is more likely to generate incorrect answers when it is uncertain about the most likely generation. There are 10 answers generated per question, so there is substantial overlap in meaning. We also show that simply using the number of semantically distinct answers as an uncertainty measure on its own performs reasonably well. Semantic entropy has a higher AUROC than the number of distinct answers, especially for CoQA whose difficulty causes greater spread in predicted probabilities between possible answers.

Finally, we can see that much of the performance gain comes from making better use of more samples. In Fig. 3a we show that for both CoQA (top) and TriviaQA (bottom) the gap between semantic entropy and length-normalised entropy widens as the number of samples increases.

6.2 Hyperparameters for effective sampling

Adjusting the temperature used for multinomial sampling has two competing effects on the generated sequences produced by the model. Increasing the temperature increases the diversity of samples



(a) (top) CoQA, (bottom) TriviaQA (b)

Figure 3: (a) Semantic entropy makes better use of additional samples because it handles duplication better, the performance gap therefore continues to improve. (b) (bottom) Higher temperatures result in more diversity but less accurate generations. (top) The best performing uncertainty comes from an intermediate temperature that balances these two forces. Results on TriviaQA.

(Fig. 3b, bottom figure, solid line). One would expect more diverse generations to cover the space of possible meanings more fully. Here we measure the diversity using the average overlap of the longest sub-sequence among sampled answers $(1 - \binom{M}{2}^{-1} \sum_{\mathbf{s} \neq \mathbf{s}' \in C} \text{Rouge-L}(\mathbf{s}, \mathbf{s}'))$. At the same time, reducing the temperature improves the average correctness of the answer (Fig. 3b, bottom figure, dashed line). Typically, more accurate models are also better at estimating uncertainty.

In fact, we find that these two effects compete and the highest AUROC for semantic entropy and length-normalised entropy is optimised by an intermediate temperature of 0.5 (Fig. 3b, top figure). A lower temperature would improve accuracy, while a higher temperature would improve diversity. A similar figure for CoQA can be found in Appendix B. Note that prior work using predictive entropy as a baseline uses a temperature of 1.0 (Kadavath et al., 2022), which our evaluation suggests would significantly weaken the baseline relative to our implementation.

7 Discussion

Many natural language problems display a crucial invariance: sequences of distinct tokens *mean* the same thing. Addressing this directly, we introduce semantic entropy—the entropy of the distribution over meanings rather than sequences—and show that this is more predictive of model accuracy on QA than strong baselines. Our unsupervised approach using 'out-of-the-box' models improves reproducibility and is easier to deploy. Unsupervised uncertainty may also help address the observation raised in prior work that supervised uncertainty measures struggle with distribution shift.

For semantic entropy, we introduce a novel bidirectional entailment clustering algorithm which uses a smaller natural language inference model. Our method therefore represents a middle ground between fully probabilistic methods and methods that use language models to exploit aspects of natural language that are not transparently present in the model activations. We believe that this sort of joint approach is more promising than relying on either perspective on its own, especially as language models continue to improve. This will become more important in cases where language models are capable of deception, something which our method does not protect against, rather than merely being uncertain between many possible meaningful options. By combining model internals with model prediction one can hope to stay a step ahead of model capabilities.

We focus on question answering because this is a particularly important free-form NLG problem with relatively clear ground truths. In the future, however, we hope our work on semantic equivalence can pave the way towards progress in settings like summarisation where correctness requires more human evaluation although additional progress on paraphrase identification in these settings is likely required first. Semantic likelihoods could also be extended to other tools for probabilistic uncertainty like mutual information, potentially offering new strategies for NLG uncertainty.