Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*, 2014. 6

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2, 7

Table 3: Illustration of semantic, syntactic, and lexical equivalence. Work with foundation models implicitly focuses on *lexical* equivalence, which entails the others, but we usually care about *semantic* equivalence.

		Equivalence		
Sentence A	Sentence B	Lexical	Syntactic	Semantic
Paris is the capital of France.	Paris is the capital of France.	✓	✓	✓
	Berlin is the capital of France.		✓	
	France's capital is Paris.			✓

A Further details on semantic entropy

A.1 Further discussion of semantic equivalence

We illustrate the distinction between different kinds of equivalence in Table 3. Lexically equivalent sequences use exactly the same symbols. They are always also semantically and syntactically equivalent (in a given context). Syntactically equivalent sentences have the same grammatical form. But they can have different meanings (not semantically equivalent) and can use different symbols (not lexically equivalent). Semantically equivalent sentences mean the same thing, but they might have different grammatical form (not syntactically equivalent) or symbols (not lexically equivalent). Two sentences can also be both syntactically and semantically equivalent but not lexically equivalent if they match up to a synonym.

Soft equivalence and transitivity. Formally, semantic equivalence is transitive. That is, if $E(\mathbf{s}, \mathbf{s}')$ and $E(\mathbf{s}', \mathbf{s}'')$ then it follows that $E(\mathbf{s}, \mathbf{s}'')$. However, the implementation of our bidirectional equivalence algorithm permits some classification errors and it is slightly 'soft'—it will sometimes return `equivalent` for pairs that are not *quite* equivalent. As a result, it is not strictly true that our equivalence relation is transitive, and therefore not strictly true that there is a unique set of equivalence classes. For example, the clusters might depend on the order in which the comparisons are made. In practice, however, we find that this does not pose a noticeable problem—usually, inspecting the outputs shows that the equivalence appears clear cut. However, we acknowledge this potential issue as an area for improvement in future clustering algorithms.

Unequal token importance. From the perspective of meaning, some tokens can matter more than others—key words. Naive methods like predictive entropy do distinguish between key words or unimportant tokens. Supervised uncertainty methods that make use of language models in the uncertainty evaluation can potentially take this into account better. In addition, our semantic entropy approach partly adjusts for this, as discussed in Section 4.1.

A.2 Further algorithmic details

In addition to the description of the method provided in the main body, in Algorithm 1 we provide the pseudocode for our bi-directional entailment algorithm.

A.3 Impact of sampling method on quality of uncertainty estimate

In Section 4, we study the impact of the temperature hyper-parameter on the performance of the uncertainty measures. Here, we show a variant of Fig. 3b for the CoQA dataset showing an almost identical pattern. Like TriviaQA, the optimal temperature is 0.5 despite a significantly harder problem with lower accuracy, suggesting that this choice hyperparameter may generalize well. Unlike TriviaQA, normalised entropy outperforms semantic entropy at high temperatures.

Beyond the temperature, there are a number of other design choices to be made when sampling: the sampling method and hyper-parameters such as `top-p` and `top-k`. Our contribution in this paper is to show the importance of these choices for uncertainty estimation which has been overlooked previously, and study the temperature in particular. While we leave the detailed study of these hyperparameters to future work, we do compare our default multinomial sampling method, to multinomial beam search sampling which focuses more on high-likelihood regions of the output space.

Algorithm 1 Bidirectional Entailment Clustering

Require: context x, set of seqs. $\{\mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(M)}\}$, NLI classifier $\mathcal{M}$, set of meanings $C = \{\{\mathbf{s}^{(1)}\}\}$
  for $2 \leq m \leq M$ **do**
    for $c \in C$ **do** ▷ Compare to already-processed meanings.
      $\mathbf{s}^{(c)} \leftarrow c_0$ ▷ Use first sequence for each semantic-class.
      `left` $\leftarrow \mathcal{M}(\text{cat}(x, \mathbf{s}^{(c)}, \text{“<g/>”}, x, \mathbf{s}^{(m)}))$ ▷ Does old sequence entail new one?
      `right` $\leftarrow \mathcal{M}(\text{cat}(x, \mathbf{s}^{(m)}, \text{“<g/>”}, x, \mathbf{s}^{(c)}))$ ▷ Vice versa?
      if `left` is `entailment` **and** `right` is `entailment` **then**
        $c \leftarrow c \bigcup \mathbf{s}^{(m)}$ ▷ Put into existing class.
      end if
    end for
    $C \leftarrow C \bigcup \{\mathbf{s}^{(m)}\}$ ▷ Semantically distinct, gets own class.
  end for
  return C

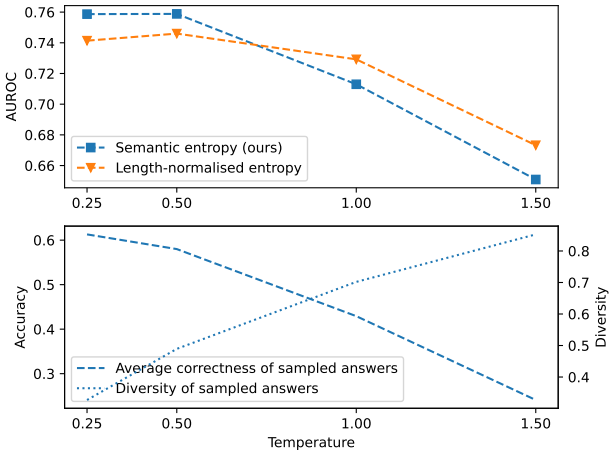

Figure 4: CoQA temperature ablation. (bottom) Similar to TriviaQA, higher temperatures mean higher diversity and lower accuracy. (top) The best performance for both methods comes at a temperature of 0.5. Unlike TriviaQA, normalised entropy outperforms semantic entropy at high temperatures.

In Table 4 we show that multinomial beam search sampling yields uncertainty measures that are less predictive of model accuracy than multinomial sampling. Beam search also generates much less diverse samples. We conjecture that multinomial beam search sampling focuses too much on the most likely sequences. The diversity of this beam search corresponds to the lowest temperature result in Fig. 4. As in the main body of the paper, we measure diversity as the average lexical overlap of the answers in the answer set. Additionally, we investigate, why the semantic entropy underperforms the length-normalised entropy at high temperatures. To that end, we manually inspect and label 100 classifications of our semantic equivalence method at T=1.5, and we find that at these temperatures, many of the generated model answers are nonsensical combinations of words from the context that is provided for the question. While the likelihood of these sequences still seems somewhat predictive of the model's accuracy, semantic clustering becomes very difficult and an unreliable signal for uncertainty estimation. At this temperature, the accuracy of the semantic equivalence methods is only 61%, whereas it is over 92% at lower temperatures (see Appendix B.2)

Note, that at low-temperatures, where one does gets plausible and well-formed model generations, semantic entropy does clearly outperform the baselines. This finding further underlines the importance of choosing appropriate sampling hyper-parameters when using entropy-based uncertainty measures in NLG.