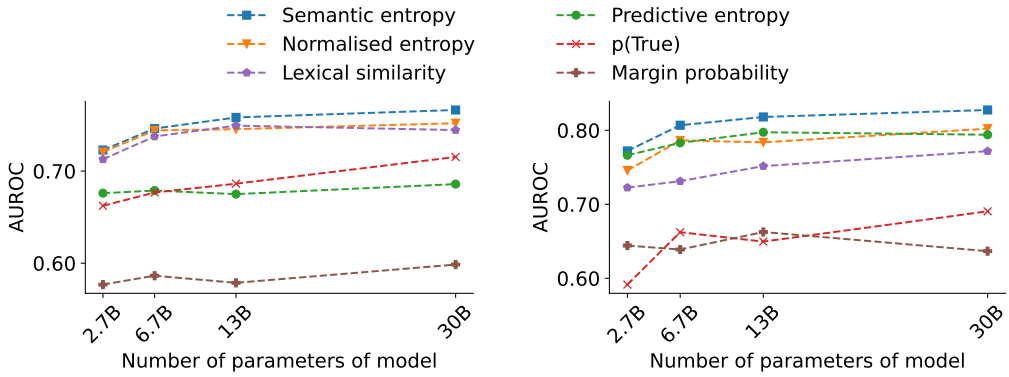

(a) CoQA (b) TriviaQA

Figure 6: The margin probability, i.e. the difference between the likelihood of the most likely answer and the likelihood of the second most likely answer, is not very predictive of models' accuracy on CoQA open-book question answering (a) nor on TriviaQA (b). Identical to Fig. 2 with the addition of *Margin probability* which was previously omitted to avoid stretching the scale.

B.6 Margin-probability baseline

We additionally compare our method to the margin probability method used for neural-symbolic parsing in Lin et al. (2022b):

$$\mathcal{H}_{\text{margin}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})) = p\left(\boldsymbol{y}^{(1)} \mid \boldsymbol{x}, \mathcal{D}\right) - p\left(\boldsymbol{y}^{(2)} \mid \boldsymbol{x}, \mathcal{D}\right),$$

where $\mathbf{y}^{(1)}$ is the top-1 beam search result and $\mathbf{y}^{(2)}$ is the top-2 beam search result.

Initially, running the method as proposed in Lin et al. (2022b) using a 13B parameter model on CoQA, we find that $\mathcal{H}_{\text{margin}}$ is not very predictive of the model's accuracy on answering questions in CoQA achieving an AUROC of 0.54.

We hypothesise that two factors contribute to this poor performance. First, since this measure only looks at the difference of likelihoods, the information about the *magnitude* of the likelihood of a given answer is lost. Second—analogously to the predictive entropy—it would be important to take *semantic uncertainty* into account when computing $\mathcal{H}_{\text{margin}}$. Manually inspecting model answers on CoQA, and the corresponding $\mathcal{H}_{\text{margin}}$, we see that the margin between two semantically equivalent answers and two semantically distinct answers is often similar. That is, this measure does not distinguish between uncertainty between paraphrases of the same meaning (in which case the model might actually be confident about meaning of the answer), and the model's uncertainty about which semantically distinct meaning is correct.

We find that if instead of obtaining $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ by multinomial sampling (as in our other experiments) instead of by beam search, this second problem becomes less pronounced and $\mathcal{H}_{\text{margin}}$ performs better while still being clearly outperformed by the other methods we study. We report our full results in Fig. 6.

Table 8: **Example of challenges for $\mathcal{H}_{\text{margin}}$.** $\mathcal{H}_{\text{margin}}$ does not distinguish between lexical and semantic uncertainty and thus can not distinguish cases where the model is certain about the correct answer (but uncertain about the precise formulation) as in row 1, and cases where the model is uncertain about the correct answer as in row 2. The semantic entropy correctly indicates low uncertainty in the first case and high uncertainty in the second case.

$\mathbf{y}^{(1)}$	$\mathbf{y}^{(2)}$	$\mathcal{H}_{\text{margin}}$	Semantic entropy
Thomas Edison.	Edison.	0.90	0.10
Thomas.	George.	0.36	0.87