UNCERTAINTY QUANTIFICATION IN RETRIEVAL AUGMENTED QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

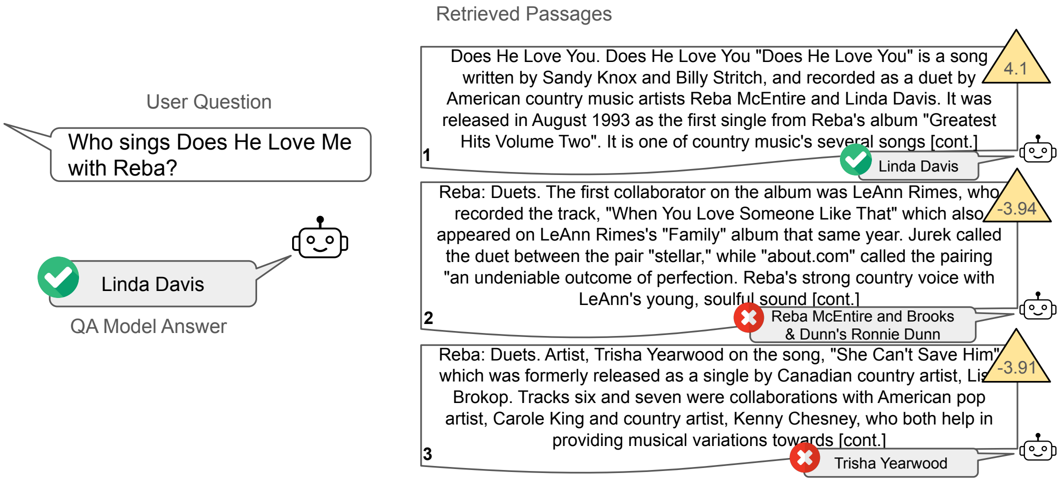
Retrieval augmented Question Answering (QA) enables QA models to overcome knowledge gaps when answering questions at test time by taking as input the question together with retrieved evidence, that is usually a set of passages. Previous studies show that this approach has numerous benefits such as improving QA performance and reducing hallucinations, without, however, qualifying whether the retrieved passages are indeed useful at answering correctly. In this work, we evaluate existing uncertainty quantification methods and propose an approach that predicts answer correctness based on utility judgements on individual input passages. We train a small neural model that predicts passage utility for a target QA model. We find that simple information theoretic metrics can predict answer correctness up to a certain extent, more expensive sampling based approaches perform better, while our lightweight approach can efficiently approximate or improve upon sampling-based approaches.

1 INTRODUCTION

Retrieval augmented Question Answering (QA) enables QA models to overcome knowledge gaps when answering user questions at test time by giving them access to input evidence, i.e., a set of passages, retrieved for the user questions (Lewis et al., 2020; Guu et al., 2020; Izacard et al., 2024). Recent work exploits the language understanding and generation abilities of Large Language Models (LLMs; (Brown et al., 2020; Ouyang et al., 2024)) and makes use of external retrievers to find the input evidence (Chen et al., 2017; Izacard & Grave, 2021a). That is, the retrieved evidence is given to the LLM-based QA model as input context in tandem with the question; the QA model will read this evidence and formulate an answer. For instance, in Figure 1, for the user question *Who sings Does He Love Me with Reba?*, the QA model is provided with a set of evidence passages together with the question; and correctly formulates the answer *Linda Davis*.

Such retrieval augmented QA architectures have proven beneficial enabling access to external knowledge (Izacard et al., 2024), increasing the performance on tail knowledge (Mallen et al., 2023), reducing hallucinations in model answers, and even improving model calibration (Jiang et al., 2021). However, there are various ways in which a retrieval augmented QA approach can go wrong at production time. The set of passages obtained using retrieval methods is far from perfect (Sciavolino et al., 2021; Yoran et al., 2024; Kasai et al., 2024) containing irrelevant or misleading evidence, the model might be under-trained to read certain passages and reason over these and the question (Izacard et al., 2024; Liu et al., 2024b), or the question can simply be ambiguous or unanswerable (Kasai et al., 2024). In these cases where the QA system lacks the knowledge to formulate an answer (i.e., it is uncertain about what the answer is), we want it to refrain from answering rather than providing an erroneous answer. Thus, predicting answer uncertainty is key.

Approaches to answer uncertainty prediction can be grouped in two main categories, sampling- and LLM-based methods. Sampling-based methods to QA uncertainty detection rely on the output discrepancies among multiple predictors on the same input (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017); i.e., this variance in outputs indicates that the model is uncertain. Concretely, these methods sample via temperature scaling (Guo et al., 2017) and then measure diversity on the set of sampled answers (Kuhn et al., 2023; Chen & Mueller, 2024). These approaches are expensive to run for in-production QA systems and the quality of the semantic similarity will degrade on long



Does He Love You. Does He Love You "Does He Love You" is a song written by Sandy Knox and Billy Stritch, and recorded as a duet by American country music artists Reba McEntire and Linda Davis. It was released in August 1993 as the first single from Reba's album "Greatest Hits Volume Two". It is one of country music's several songs [cont.]

Reba: Duets. The first collaborator on the album was LeAnn Rimes, who recorded the track, "When You Love Someone Like That" which also appeared on LeAnn Rimes's "Family" album that same year. Jurek called the duet between the pair "stellar," while "about.com" called the pairing "an undeniable outcome of perfection. Reba's strong country voice with LeAnn's young, soulful sound [cont.]

Reba: Duets. Artist, Trisha Yearwood on the song, "She Can't Save Him" which was formerly released as a single by Canadian country artist, Liz Brokop. Tracks six and seven were collaborations with American pop artist, Carole King and country artist, Kenny Chesney, who both help in providing musical variations towards [cont.]

Figure 1: Example of user question from the Natural Questions dataset with the set of three top retrieved passages with Contriever (Izacard et al., 2022) (the other two passages below the rank are less relevant and not shown in the figure); the gold answer is *Linda Davis*. The target QA model GEMMA2-9B correctly answers the question when provided with the top five passages. Below each passage, it is shown the answer generated by the QA model when only prompted with that passage and the question. The QA model correctly answers when prompted with the first passage and produces an incorrect answer when prompted with each of the other ones. The yellow triangles on the top right of the passages are the predicted utility scores by our utility ranker. Higher values indicate more useful passages and our model correctly identifies that the top passage is better.

answers (Zhang et al., 2024).[1] LLM-based methods explore to what extent language models are able to correctly express uncertainty about their own predictions (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Zhou et al., 2023). These look into whether the model's confidence in its outputs coincides with their correctness (i.e., calibration), methods to fix calibration, and ways to elicit from the model a verbal expression of that confidence (i.e., linguistic calibration). Findings about model calibration are diverse and model dependent, fixing relies on approximations for the case of black-box models and fine-tuning what could be infeasible in practice given current LLMs' sizes. None of these answer uncertainty detection approaches has been applied in the context of retrieval augmented QA, most of them are applied on closed-book QA tasks where the answer is predicted based on the question and the models' encoded knowledge.

In this work, we propose a **secondary model that makes predictions at individual retrieved passage level that are useful to estimate answer uncertainty of retrieval augmented QA models**. We hypothesize that the type of retrieved passages and questions, the relation between them and their implicit interaction with the QA model's own knowledge are indicators of answer correctness. If the passages are informative and priming the QA model towards appropriate knowledge, we expect the QA model to produce a correct answer. In contrast, if the passages are not informative or misleading and the posed question is out of the QA model's knowledge, we expect it to generate an erroneous answer (i.e., either factually incorrect or completely made up content). We operationalise this as retrieved passage *utility*. Given a question, a passage is *useful*, if a QA model can correctly answer the question based on it. We train a small neural model to predict passage utility which we refer to as *utility ranker*. We train the utility ranker on utility judgements generated by the target QA model. We borrow ideas from direct uncertainty quantification approaches (Van Amersfoort et al., 2020; Lahlou et al., 2023) but we do not decompose uncertainty or outline shifts in the input distribution.

We show that individual passage utilities are good predictors of retrieval augmented QA accuracy. This means that it is possible to train an answer uncertainty predictor independently

---

[1] By expensive we mean both latency as well as cost as a long prompts might need to be processed and QA systems may rely on paid proprietary language models.

from the choice of number of retrieved passages used to prompt the target QA model. Because retrieved passages are scored individually, our approach is independent of the *number* of retrieved passages chosen for the target QA model. We evaluate our approach on short-form question answering tasks. Figure 1 shows an example of input question, set of retrieved passages, and correct answer from the Natural Questions dataset (Kwiatkowski et al., 2019). Results on six QA datasets show that our approach performs on par with existing sampling-based uncertainty quantification approaches while being more efficient at test time. It requires a small model pass over the set of input passages and the question (see inference cost comparison in Appendix C.1). Surprisingly, in more complex reasoning questions (SQuAD) and adversarial QA settings (e.g., rare entities or unanswerable questions) our approach surpasses existing uncertainty quantification methods. Moreover, we show that the utility scores predicted by the Utility Ranker can be used to re-rank retrieved passages obtained via the external retrieval system to improve QA accuracy (Liu et al., 2024b).

2 RELATED WORK

Uncertainty Quantification for Question Answering Several methods have been proposed to predict answer uncertainty in QA; however, none of them has analysed uncertainty in retrieval augmented QA models. Many existing approaches rely on capturing output variation as the expression of model uncertainty (Kuhn et al., 2023; Farquhar et al., 2024; Chen & Mueller, 2024). On a sample of model outputs, Kuhn et al. (2023) propose to first cluster answers with similar meaning via natural language inference before computing entropy. Chen & Mueller (2024) propose an approach for *black-box* models, they also compute similarities in the set of answers but associate them with a model self-judgement of confidence. These approaches are expensive to run at inference time for a production QA system, they require several inference steps plus the similarity computations. In addition, as the length of the answers increases, measuring similarity becomes more complex (Zhang et al., 2024). Hou et al. (2024) propose a decomposition of predictive uncertainty and focus on quantifying aleatoric uncertainty (i.e., uncertainty in the data) caused by ambiguous questions. This approach is orthogonal to ours.

Judging the Utility of Retrieved Passages Previous work has analysed the set of retrieved passages (Yu et al., 2023; Asai et al., 2024; Wang et al., 2024; Xu et al., 2024; Yoran et al., 2024) following the observation that passages can be irrelevant or misleading making the QA model prone to producing incorrect answers. Asai et al. (2024) make use of an external critic model to judge whether a question requires retrieval (or not), whether the retrieved passages are relevant to formulate the answer, and whether the final response elaborated by the QA model is useful. While they analyse retrieved passage *relevance*, this decision is taken by an external extreme-scale critic (e.g., GPT-4) and used to fine-tune the QA model. In contrast, we do not fine-tune the target QA model but rather we elicit utility judgements from it to train a secondary model to predict passage utility. Other work creates auxiliary tasks around retrieved passages enforcing the QA model to reason on them; e.g., by taking notes about each passage (Yu et al., 2023) or generating passage summaries (Xu et al., 2024). These methods also use extreme-scale LLMs to generate training data to fine-tune the retrieval augmented QA model. Park et al. (2024) select specific in-context examples to improve the LLM's reasoning on the input passages, their focus is on detecting input passages with conflicting content (e.g., different dates for a given event). These approaches aim at improving QA performance while our primary goal is modelling QA uncertainty.

Improving Retriever via Reader Performance Previous work with pre-trained language models has focused on jointly training the retriever and reader modules end-to-end (Lee et al., 2019; Lewis et al., 2020; Izacard & Grave, 2021b). That is, the performance of the question answering model is propagated also to the retriever. This joint training scheme can be very expensive for current (extreme-scale) LLMs. Our approach can be seen as an intermediate module between the QA model (reader) and the external retriever. It would be interesting to explore our utility ranker to provide feedback (e.g., to label data) to fine-tuning the retriever. In recent work, Salemi & Zamani (2024) evaluate the performance of information retrieval systems via retrieval augmented QA performance. Interestingly, they show that external judgements (e.g., query-document relevance labels) of passage utility correlate poorly with retrieval augmented QA performance.

Learning to Predict Confidence Some approaches train a specific model to predict a confidence score (Dong et al., 2018; Kamath et al., 2020; Mielke et al., 2022). For semantic parsing, Dong et al. (2018) train a confidence predictor based on a set of uncertainty features from the input and the model. Mielke et al. (2022) also train a calibrator that, given the user question and model generated answer, predicts a confidence score. In our approach, we simple aggregate predicted individual passage utilities but it would also be possible to train a confidence module that takes utilities with other features into account (e.g., output sequence probability), and predicts a confidence score.

3 MODELLING ANSWER UNCERTAINTY

Formally, we define retrieval augmented QA as follows. Given question x and set of retrieved passages $R = \{p_1, p_2, \cdots, p_{|R|}\}$ obtained with retriever $\mathcal{R}$, a LLM-based QA model $\mathcal{M}$ is prompted to generate answer $y_\mathcal{M}$ to question x token-by-token as $y_\mathcal{M} = \arg\max_{y_\mathcal{M}} \prod_{t=1}^{|y_\mathcal{M}|} p_\mathcal{M}(y_t|y_{1..t-1}, x, R)$. We want to estimate the uncertainty or error of $\mathcal{M}$ on generating $y_\mathcal{M}$ given x and R; i.e., we want an estimator $\{x, R\} \mapsto \mathbf{u}_\mathcal{M}(\{x, R\})$ of $\mathcal{M}$'s answer uncertainty. In our approach, the answer uncertainty predictor $\mathbf{u}_\mathcal{M}$ is based on individual passage utilities. Our hypothesis is that individual passage utilities of retrieved passages in R are indicators of the QA model uncertainty when generating $y_\mathcal{M}$ when prompted with R. For instance, in Figure 1, given that the first passage in the set has a high utility score, this indicates that the QA model is likely to be confident when providing the answer *Linda Davis*. Thus, we want a passage utility estimator $\{x, p\} \mapsto \upsilon_\mathcal{M}(\{x, p\})$ of every $p \in R$. In what follows, we focus on defining passage utility and how to estimate and predict it. Next, we discuss a simple answer uncertainty estimator $\mathbf{u}_\mathcal{M}$ based on $\upsilon_\mathcal{M}$.

3.1 PASSAGE UTILITY RANKING

Passage Utility Intuitively, a passage p retrieved for question x is useful for a QA model $\mathcal{M}$, if $\mathcal{M}$ can correctly answer x when prompted with p. In addition, $\mathcal{M}$'s reliance on passage p to formulate the answer may vary. That is, the QA model may formulate a correct answer even though p does not provide the answer itself; instead, p positively primes $\mathcal{M}$ to use its memorised knowledge. The utility of the first passage, in Figure 1, is high as the QA model generates a correct answer when prompted with it and the fact that *Linda Davis* sings together with *Reba McEntire* can be derived from it. The second and third passages, although related to the topic of the question, are not useful. The QA model is uncertain about how to answer the question and the passages do not help, the model incorrectly answers when prompted with each of them. Thus, the second and third passages' utilities are low.

Concretely, we estimate the utility of passage p for QA model $\mathcal{M}$ to answer question x by combining two measures. These are *accuracy*, denoted as $a(y_\mathcal{M})$, whether the generated answer $y_\mathcal{M}$ is correct, and *entailment*, denoted as $e(y_\mathcal{M})$, how much does passage p supports the generated answer $y_\mathcal{M}$. Accuracy is computed by a critic model $\mathcal{A}$ and entailment by a Natural Language Inference (NLI) classifier model $\mathcal{E}$. We define the combined passage utility as $\upsilon_\mathcal{M} = (a(y_\mathcal{M}) + e(y_\mathcal{M}))/2$ and takes values in the closed interval $[0, 1]$ given that the ranges are the set $\{0, 1\}$ for a and the closed interval $[0, 1]$ for e.

Utility Ranker We train a small neural model to predict passage utility scores, $\{x, p\} \mapsto \upsilon_\mathcal{M}(\{x, p\})$. We use observed answer accuracy and entailment by QA model $\mathcal{M}$ on a training set $D = \{(x, p)\}$ to train the utility predictor. That is, we run the QA model $\mathcal{M}$ on examples from D and compute passage utilities to form a training set for our utility predictor $D_\mathcal{M} = \{(x, p, \upsilon_\mathcal{M})\}$.

For recall purposes, retrieval augmented QA generally retrieves more than one input passage for each question x, i.e., $|R| > 1$. To generate training data for the passage utility predictor, we retrieve $|R|$ passages per question in order for to cover passages with different usefulness. From the set of passages R for question x, we derive training instances $\{(x, p_i, \upsilon_{\mathcal{M}i}) \,|\, p_i \in R\}$. We exploit this to train the passage utility predictor with a contrastive learning scheme.[2] That is, if p_i and p_j are passages in R and p_i is more useful than passage p_j to answer question x, the predicted utility score

---

[2] We also run initial experiments training the predictor with a target value and the Mean Squared Error (MSE) objective and performance was also competitive.

$\upsilon_{\mathcal{M}i}$ should be higher by a margin m than the predicted score $\upsilon_{\mathcal{M}j}$ for p_j (i.e., p_i should be ranked higher than p_j). We train the utility predictor with the following pair-wise ranking objective:

$$\mathcal{L}_{rank} = \sum_{((x,p_i),(x,p_j)) \in R \times R, i \neq j} \max(0, -z(\upsilon_{\mathcal{M}i} - \upsilon_{\mathcal{M}j}) + m)), \tag{1}$$

where z controls the gold order between p_i and p_j (e.g., if $z = 1$, p_i has higher utility, conversely $z = -1$ indicates the opposite ordering) and m is a hyper-parameter. The passage utility predictor is trained with a Siamese neural network. Its architecture is constituted by a BERT (Devlin et al., 2019) based encoder followed by a pooling and two MLP layers stacked on top of BERT outputs (Fang et al., 2024). The output layer computes the utility score as $\upsilon_{\mathcal{M}i} = W_o h^L + b_o$ where h^L is the vector representation for (x, p_i) from the last hidden layer (the L-th layer) of the network. At inference time, we compute a single utility score for each passage. We provide implementation and training details in Section 4.

To enforce the signal on accuracy prediction and to regularise the range of utility values learned by the ranking scheme, we combine the ranking objective in Equation 1 with the following Binary Cross Entropy (BCE) objective (Sculley, 2010):

$$\mathcal{L}_{BCE} = \sum_{(x,p) \in \{(x,p_i),(x,p_j)\}} a_{\mathcal{M}} \times (\log(p(x,p)) + (1 - a_{\mathcal{M}}) \times \log(1 - p(x,p)), \tag{2}$$

where $p(x,p) = \text{sigmoid}(\upsilon_{\mathcal{M}})$ and $a_{\mathcal{M}}$ is the target accuracy label taking values in the set $\{0, 1\}$. We train the utility predictor with the following combined objective:

$$\mathcal{L} = \mathcal{L}_{rank} + \lambda \mathcal{L}_{BCE}, \tag{3}$$

where λ is a hyper-parameter. Both the ranking and BCE objectives are compatible with gold annotations that could be obtained via human intervention in an interactive and active learning learning setting. That is, it would be feasible to elicit from human judges (e.g., moderators of the QA system) answer accuracy labels (e.g., *correct/incorrect*) and level of passage support for the generated answer (e.g., *best* or *worse*) (Simpson et al., 2020; Fang et al., 2024). Note that the Utility Ranker could also be trained with different variants of this objective that also exhibit competitive performance. We report in Appendix D.1 a study on the ablation of the different components of the training objective.

The passage utility predictor is related to the direct error prediction approach in (Lahlou et al., 2023). Lahlou et al. (2023) train a secondary model to estimate target model loss; instead, we train the passage utility predictor with sequence level metrics, i.e., accuracy and entailment, which indirectly measure error. This choice is best suited for our task for various reasons. First, in the context of text generation and its possibly diverse (e.g., paraphrases) but correct set of possible generated answers (Kuhn et al., 2023), predicting loss against a unique single paraphrase would result in a too narrow estimation. Our choice is also adequate for proprietary LLMs where it is not possible to create training data with model losses. Finally, our approach is suited for collecting data from user feedback for active model adaptation (Simpson et al., 2020; Fang et al., 2024). In the image domain, van Amersfoort et al. (2020) map inputs to feature representations and take the distance between new inputs and their closest cluster centroids as a measure of uncertainty. In retrieval augmented QA with LLMs, text passages, and questions, it is less clear what the boundary between seen and unseen texts or topics is. Because our Utility Ranker is trained on a target dataset it could be exploited to detect out-of-domain instances for a target application. It would be interesting to pursue future work on using our Utility Ranker as a content controller for the target LLM-based QA model.

Some approaches to answer uncertainty prediction that train a secondary model are in (Kamath et al., 2020; Zhang et al., 2021). However, none of them is applied to retrieval augmented QA; but instead to Reading Comprehension (RC), i.e., the task of generating an answer based on a single positive (i.e., supposed to contain the answer) context document. There are two major differences with our work. One is that in their scenario, all input documents are useful while in ours the utility of retrieved passages is varied. The second one is that we show that individual passage utilities are good predictors of retrieval augmented QA with a set of retrieved passages.

3.2 ANSWER UNCERTAINTY ESTIMATION FOR RETRIEVAL AUGMENTED QA

For retrieval augmented QA, we want an estimator $\{x, R\} \mapsto \mathbf{u}_{\mathcal{M}}(\{x, R\})$ of the answer uncertainty of a target QA model $\mathcal{M}$ when generating answer $y_{\mathcal{M}}$ from a prompt with set of passages R and question x. We propose the direct estimation of $\mathbf{u}_{\mathcal{M}}$ from individual passage utilities predicted for passages in R. The intuition is that, the highest the utility in one (or more) passages in R the less uncertain $\mathcal{M}$ will be when generating answer $y_{\mathcal{M}}$. Concretely, we take the maximum utility score that is given to passages in R as an estimate of answer uncertainty $\mathbf{u}_{\mathcal{M}}$, i.e.,

$$\mathbf{u}_{\mathcal{M}}(\{x, R\}) = \max(\upsilon_{\mathcal{M}}(\{x, p\}) \,|\, p \in R). \tag{4}$$

Note that other more complex estimators $\{x, R\} \mapsto \mathbf{u}_{\mathcal{M}}(\{x, R\})$ could be learned by training, for instance, a regression model on individual passage utilities in addition to other features of the target model $\mathcal{M}$ such as probability of the generated answer $y_{\mathcal{M}}$ (Dong et al., 2018).

4 EXPERIMENTAL SETUP

Accuracy Evaluation A precise metric for measuring accuracy is key when evaluating the quality of uncertainty estimation. Token overlap metrics are far from being precise and can over- or under-estimate accuracy, e.g., Acc yields a higher score for the pair of gold and generated answers (*a politician*, *not a politician*) than for the pair (*a politician*, *a congressperson*). Thus, our main metric to evaluate QA model performance and as the accuracy evaluator $\mathcal{A}$ to create data to train the passage utility predictor, is based on a LLM judgement of accuracy proposed by Sun et al. (2024) (**AccLM**). A critic LLM is prompted with the gold and generated answer and asked to judge whether they are the equivalent. This metric is more robust to surface string differences between generated and gold answers, e.g., the *5* and *five* strings will be considered as equal. In a sample of 840 generated answers human and LLM-based judgment of correctness agreed 98% of the time (Sun et al., 2024). We use the prompt as proposed in (Sun et al., 2024), we include it in Appendix B for completeness. We use Qwen2-72B-Instruct (Yang et al., 2024) to obtain accuracy judgments. For compatibility with previous work and as a lower bound, in Appendix D.2, we report QA model performance with token overlap accuracy (**Acc**) defined as whether the gold answer is contained in the generated answer (Mallen et al., 2023; Asai et al., 2024).

Utility Ranker Implementation Details To create the training set $D_{\mathcal{M}}$ to train the Utility Ranker, we consider the first top five retrieved passages for each question, i.e., $|R| = 5$. Note that this is a hyper-parameter and other values would also be possible, e.g., with larger sizes of $|R|$ further training data would be available. We use the target QA model $\mathcal{M}$ to generate answers $y_{\mathcal{M}}$ for each of the five passages p in R (i.e., $\mathcal{M}$ is prompted with passage p and question x). We then ge utility scores using the LLM-based accuracy judge $\mathcal{A}$ as described above and an ALBERT-xlarge Lan et al. (2020) model optimized on MNLI (Williams et al., 2018) and VitaminC (Schuster et al., 2021) as our entailment judge $\mathcal{E}$.

Comparison Approaches and Baselines Several uncertainty quantification approaches have been proposed. We choose the stronger methods from previous work (Fadeeva et al., 2023) to compare our approach with.

Information Based. We compare against the stronger information based uncertainty quantification approaches reported in previous work Fadeeva et al. (2023). These are based on predictive probabilities; recall that the predictive distribution under QA model $\mathcal{M}$ prompted with question x and set of passages R is $P(y_{\mathcal{M}}|x, R, \mathcal{M}) = \prod_{t=1}^{|y_{\mathcal{M}}|} p_{\mathcal{M}}(y_t|y_{1..t-1}, x, R)$ for a target QA model $\mathcal{M}$.

Maximum Sequence Probability (MSP) based uncertainty estimation is based on the probability of the most likely answer and computed as $\text{MSP}(y_{\mathcal{M}} \,|\, x, R, \mathcal{M}) = 1 - [\arg\max_{y_{\mathcal{M}}}(P(y_{\mathcal{M}}|x, R, \mathcal{M}))]$. In the Appendix D, we also report Perplexity (PPL) computed as $PPL(y_{\mathcal{M}}, x, R; \mathcal{M}) = \exp\{-\frac{1}{|y_{\mathcal{M}}|}\sum_{t=1}^{|y_{\mathcal{M}}|} P_{\mathcal{M}}(y_t|y_{1..t-1}, x, R)\}$, i.e., based on the average negative log-likelihood of the generated tokens. The other uncertainty estimation approach is the negative mean Point-wise Mutual Information (PMI) Takayama & Arase (2019); i.e., it compares the probability of generating answer $y_{\mathcal{M}}$ given the prompt with question x and passages R w.r.t the probability given by $\mathcal{M}$ to $y_{\mathcal{M}}$ without context. Intuitively, the higher the PMI the more certain on generating $y_{\mathcal{M}}$. PMI is computed as $PMI(y_{\mathcal{M}}, x, R; \mathcal{M}) \frac{1}{|y_{\mathcal{M}}|}\sum_{t=1}^{|y_{\mathcal{M}}|} \log \frac{p_{\mathcal{M}}(y_t|y_{1..t-1})}{p_{\mathcal{M}}(y_t|y_{1..t-1}, x, R)}$.
The other two methods are based on entropy. We compare with Regular Entropy (RE), i.e., the

entropy on the predictive distribution computed at sequence level $\mathbb{E}[-\log P(y_{\mathcal{M}}|x, R, \mathcal{M})]$ with $\mathbb{E}$ computed on sequences $y_{\mathcal{M}}$ sampled from $P(y_{\mathcal{M}} \mid x, R, \mathcal{M})$. In practice, this is approximated via Monte-Carlo integration, i.e., sampling N random answers from $P(y_{\mathcal{M}} \mid x, R, \mathcal{M})$. Thus, Regular Entropy is computed as $-\frac{1}{N}\sum_{n=1}^{N} \log \tilde{P}(y_{\mathcal{M}}^{(n)} \mid x, R, \mathcal{M})$, where $\tilde{P}(y_{\mathcal{M}}^{(n)} \mid x, R, \mathcal{M})$ is the length normalised version of $P(y_{\mathcal{M}}^{(n)}|x, R, \mathcal{M})$.

Answer Variation. Kuhn et al. (2023) propose a variant of regular entropy, named Semantic Entropy (SE), that accounts for uncertainty in the surface form of the generated answers rather than on meaning. Concretely, Semantic Entropy clusters the set of N samples into M, $M \leq N$, clusters with the same meaning via bidirectional entailment. Then computes the average answers' probability within each cluster, $SE(x, \mathcal{M}) = -\sum_{m=1}^{M} \hat{P}_m(x, \mathcal{M}) \log \hat{P}_m(x, \mathcal{M})$ where $\hat{P}_m(x, \mathcal{M}) = \frac{\sum_{y_{\mathcal{M}} \in C_m} P(y_{\mathcal{M}} \mid x, R, \mathcal{M})}{\sum_{m=1}^{M} \sum_{y_{\mathcal{M}} \in C_m} P(y_{\mathcal{M}} \mid x, R, \mathcal{M})}$. Cluster Assignment (CA) is the variant of SE without answers' probabilities where $\hat{P}_m(x, \mathcal{M})$ is approximated from the number of answers in the cluster (CA values are very close to SE values; thus we report them in Appendix D).

Reflexive. We compare with p(true) proposed by Kadavath et al. (2022). This approach uses the same target QA model (LLM) evaluate whether the answers it produces are correct. It is prompted with the question and a set of candidate answers, i.e., the most likely answer plus a sample of size N answers, and instructed to respond whether the most likely answer is true or false (i.e., correct/incorrect). The score produced by this approach is the probability of the model $\mathcal{M}$ generating the token True. p(true) needs several in-context examples to work well, so we fit as many examples as can be in the context.

Baselines. The sets of passages in R are originally ranked by the IR system, so each passage in R has a retriever score which can be seen as baseline passage utility. We thus take the Retriever Score as a baseline. Despite the QA models are instructed to produce a short answer, these often generate longer answers. The length of the answer could be a feature indicating that the model is uncertain about the answer. Thus, we estimate answer uncertainty from the Answer Length (Ans.Len) as the number of words in the answer.

Following previous work (Farquhar et al., 2024), we take $N = 10$ samples and use multinomial sampling to generate samples. That is, we set the sampling temperature to 1, with nucleus sampling ($P = 0.9$) (Holtzman et al., 2020) and top$-K$ sampling ($K = 50$) (Fan et al., 2018), and use a different random seed to draw each sample. Most likely answers are generated with greedy sampling at temperature equal to 0. We use the implementation provided by Farquhar et al. (2024) to compute RE, SE, CA, and p(true). We report inference cost of each approach in Appendix C.1.

QA Models Our target retrieval augmented QA models $\mathcal{M}$ are based on the following instruction fine-tuned LLMs. To assess the performance of the Utility Ranker for QA models that potentially exhibit different answer uncertainty, we consider different families of similar size. These are Llama-3.1-8B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.3 Jiang et al. (2023), and Gemma2-9B-it Riviere et al. (2024). For all QA models, we use a simple prompt including the retrieved passages and the question in the context, the prompt is shown in Table 6 of the Appendix. We use vLLM for inference (Kwon et al., 2023). Following previous work on retrieval augmented generation, we use Contriever Izacard et al. (2022) as our external retrieval model (Asai et al., 2024) and the target QA models are based on a set of retrieved passages of size $|R| = 5$ Yu et al. (2023); Asai et al. (2024); Xu et al. (2024).

Datasets We evaluate our answering uncertainty prediction approach on short-form answer generation tasks. Concretely, we evaluate on the Natural Questions Kwiatkowski et al. (2019), TriviaQA Joshi et al. (2017), WebQuestions Berant et al. (2013), and SQuAD (Rajpurkar et al., 2016) datasets. We follow the training/validation/test splits in prior work Lee et al. (2019); Min et al. (2019); Karpukhin et al. (2020). To test the generalisation robustness of our approach we carry out additional experiments on PopQA Mallen et al. (2023), a dataset with questions about rare entities, and RefuNQ Liu et al. (2024a), a dataset with unanswerable questions about non-existing entities. Statistics about our datasets are given in the Appendix in Table 5.

Evaluation of the Quality of Uncertainty Estimation To assess the quality of answer uncertainty prediction, we follow Farquhar et al. (2024) and report the Area Under the Receiver Operator

Table 1: Answer uncertainty estimation for QA models Gemma2-9B, Llama-3.1-8B, and Mistral-7B-v0.3 on NaturalQuestions, TriviaQA, WebQuestions, and SQuAD (evaluation with in-distribution test data for the Utility Ranker). We report AUROC and AURAC.

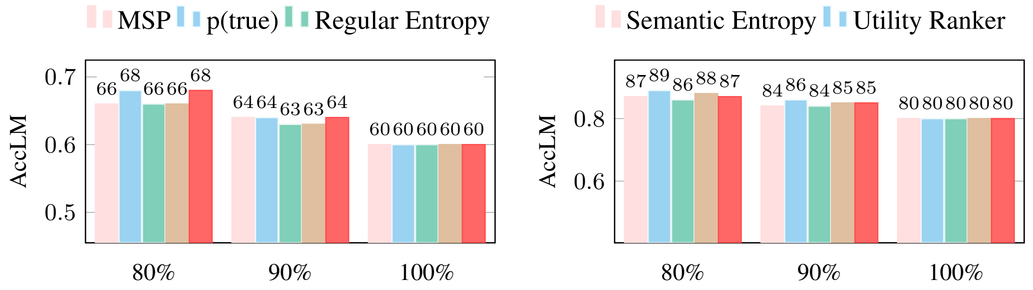
	NaturalQuestions		TriviaQA		WebQuestions		SQuAD	
	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
Gemma2-9B								
MSP	0.69	0.67	0.68	0.83	0.63	0.66	0.65	0.63
PMI	0.51	0.58	0.53	0.78	0.45	0.58	0.50	0.55
p(true)	0.72	0.70	0.78	0.86	**0.74**	**0.74**	0.67	0.66
Regular Entropy	0.69	0.68	0.65	0.82	0.63	0.67	0.65	0.62
Cluster Assignment	0.67	0.66	0.69	0.83	0.60	0.65	0.66	0.63
Semantic Entropy	0.68	0.67	0.68	0.83	0.60	0.65	0.66	0.63
Ans.Len	0.65	0.65	0.59	0.80	0.62	0.66	0.61	0.60
Retriever Score	0.60	0.65	0.68	0.84	0.53	0.62	0.61	0.62
Utility Ranker	**0.76**	**0.72**	**0.81**	**0.88**	0.72	0.71	**0.81**	**0.74**
Llama-3.1-8B								
MSP	0.71	0.69	0.83	**0.88**	0.71	0.74	0.77	0.69
PMI	0.56	0.60	0.57	0.78	0.51	0.65	0.61	0.59
p(true)	**0.79**	**0.74**	**0.84**	0.87	**0.76**	0.76	0.65	0.61
Regular Entropy	0.72	0.69	0.83	**0.88**	0.72	0.74	0.78	0.69
Semantic Entropy	0.69	0.67	0.81	0.86	0.68	0.73	0.75	0.68
Ans.Len	0.59	0.61	0.60	0.78	0.61	0.68	0.57	0.55
Retriever Score	0.58	0.62	0.64	0.81	0.50	0.63	0.65	0.61
Utility Ranker	0.73	0.70	0.78	0.86	**0.76**	**0.78**	**0.84**	**0.73**
Mistral-7B-v0.3								
MSP	0.68	0.63	0.73	0.87	0.65	0.68	0.71	0.65
PMI	0.53	0.59	0.55	0.79	0.50	0.62	0.58	0.60
p(true)	0.72	0.67	**0.84**	**0.88**	0.72	0.70	0.69	0.63
Regular Entropy	0.60	0.60	0.71	0.85	0.61	0.68	0.66	0.62
Semantic Entropy	0.67	0.63	0.78	**0.88**	0.69	0.68	0.71	0.66
Ans.Len	0.68	0.63	0.67	0.84	0.64	0.69	0.66	0.63
Retriever Score	0.59	0.60	0.67	0.82	0.53	0.65	0.64	0.62
Utility Ranker	**0.76**	**0.68**	0.79	0.86	**0.76**	**0.71**	**0.80**	**0.68**

Curve on detecting answer uncertainty, i.e., incorrect answers, (**AUROC**) and the area under the rejection accuracy curve (**AURAC**). AURAC summarises the accuracy of QA models when answer uncertainty is used to refuse to answer questions. It summarises accuracy at different percentages of rejection. Instruction fine-tuned models are known to refuse to answer questions, i.e., they produce answers such as *This information is not available in the text*. In some cases, the refusal response will be adequate (e.g., no input passage contains the information to answer) but in many cases QA models may refuse when they should have provided an answer Adlakha et al. (2024); Liu et al. (2024a). Thus, to simplify the assessment of answer correctness, we did not explicitly instruct the QA models to abstain and treat occurring refusal answers as cases of uncertainty where the QA model is expressing the uncertainty in the answer (Farquhar et al., 2024). We report the percentage of refusal answers for each QA model and dataset.

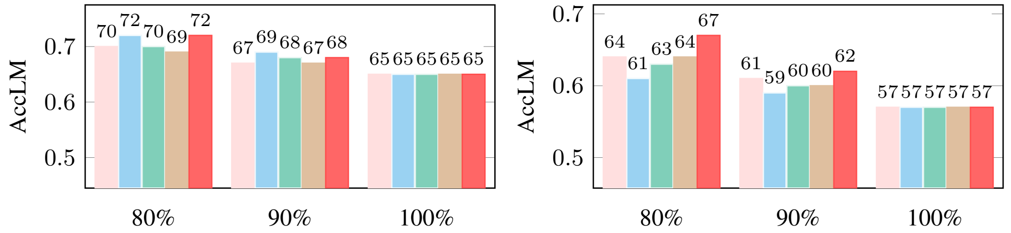
5 RESULTS

5.1 UNCERTAINTY QUANTIFICATION

Answer uncertainty estimation results for the three QA models (Gemma2-9B, Llama-3.1-8B, and Mistral-7B-v0.3) are shown in Table 1 (results on the development set are included in Appendix D). In terms of predicting answer uncertainty (i.e., model incorrect answers), column AUROC in Table 1, simple metrics based on models' probabilities such as MSP perform better for some models. It exhibits high performance for Llama-3.1-8B while lower performance for Gemma2-9B and Mistral-7B-v0.3. Sampling-based approaches (meaning diversity and reflexive), can better identify model uncertainty but at the cost of running inference several times to have a good size sample for the estimation. Our Utility Ranker has similar or better performance with a single inference step on each input passage. We speculate that clustering approaches can suffer in phrase



(a) NaturalQuestions (b) TriviaQA



(c) WebQuestions (d) SQuAD

Figure 2: Average QA model performance on test sets with $|R| = 5$. We show model based accuracy (AccLM) at different percentages of rejecting to answer (i.e., when choosing to respond on 80%, 90%, and all the cases) given uncertainty estimations by the different approaches.

or sentence level correct answers where these contain different levels of details Zhang et al. (2024); thus, not being clustered together wrongly suggesting variation.

On improving question-answering accuracy, AURAC column in Table 1, with the exception of TriviaQA, all uncertainty prediction approaches outperform the information theoretic approaches (i.e., MSP, PMI). The Utility Ranker performs on par or better than the more expensive sampling based approaches. To have a clearer picture of baseline retrieval augmented QA accuracy w.r.t. accuracy when the uncertainty estimation is used to decide whether to abstain nor not, we show in Figure 2 the accuracy of the model at different thresholds for answer rejection. That is, we report when the QA model chooses to answer only the 80% or 90% of the most confident cases as well as when always answers. Retrieval augmented QA accuracy per model and dataset on the full test and development sets is included in Appendix D. Across all datasets, the Utility Ranker performs on par with of better than more expensive uncertainty estimation approaches. The easiest QA task is TriviaQA where QA models show very good performance and information theoretic methods work on par with more complex ones. On the most difficult task, SQuAD, the utility ranker outperforms all other methods both at 20% and 10% of rejected answers.

5.2 ROBUSTNESS AND GENERALISATION OF UNCERTAINTY ESTIMATION

We assess the robustness and generalisation of our Utility Ranker on test cases that are different from those examples seen during training (i.e., out-of-distribution). These examples encompass real cases that a QA model will face at test time such as different type of questions, e.g., longer and more complex. We also study the generalisation and robustness in extreme adversarial cases such as questions about tail knowledge for both retrievers and instruction tuned LLMs (PopQA), and unanswerable questions (RefuNQ).

Distribution Shift Table 2 shows the performance of the Utility Ranker when evaluated in out-of-distribution data. The first column indicates the training data and the first row indicates the evaluation data. Results in the diagonal correspond to the Utility Ranker trained and evaluated in the same data distribution; the off-diagonal cells correspond to the Utility Ranker evaluated zero-shot

Table 2: Performance of GEMMA2-9B's Utility Ranker on distribution shift. That is, trained on one dataset and evaluated zero-hot on another one. We report all combinations of train and test data. The first column indicates train data while the first row test data.

	NaturalQuestions		TriviaQA		WebQuestions		SQuAD	
	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
NaturalQuestions	**0.76**	**0.72**	0.72	0.86	0.65	0.67	0.72	0.68
TriviaQA	0.64	0.67	**0.81**	**0.88**	0.63	0.68	0.71	0.68
WebQuestions	0.60	0.64	0.72	0.86	**0.72**	**0.71**	0.58	0.59
SQuAD	0.65	0.67	0.77	0.87	0.61	0.65	**0.81**	**0.74**

Table 3: Answer uncertainty estimation for GEMMA2-9B on adversarial QA tasks (PopQA and RefuNQ). Its Utility Ranker is trained on Natural Questions.

	PopQA		RefuNQ	
	AUROC	AURAC	AUROC	AURAC
MSP	0.66	0.58	0.66	0.39
PMI	0.51	0.50	0.54	0.35
p(true)	0.71	**0.62**	0.73	0.45
Regular Entropy	0.66	0.58	0.66	0.39
Semantic Entropy	0.69	0.59	0.68	0.41
Ans.Len	0.62	0.55	0.65	0.38
Retriever Score	0.63	0.58	0.76	0.47
Utility Ranker (NQ)	**0.72**	**0.62**	**0.82**	**0.51**

in a different dataset. As expected, the Utility Ranker variants evaluated on a different dataset show a decrease in performance. However, for some training data the decrease is small yet providing a competitive prediction. That is, NaturalQuestions and SQuAD provide the best training data, what agrees with previous experiments in reading comprehension settings (Chen et al. (2021) choose NaturalQuestions to train the base model, Kamath et al. (2020); Zhang et al. (2021) SQuAD). The Utility Ranker variants trained on WebQuestions (smallest training set) and TriviaQA (the easiest task) have the worst generalisation performance. Note that we focus on zero-shot to assess bare transfer performance, in real scenarios it would make sense as proposed in previous work, to train the model with few examples of the out-of-distribution data (Kamath et al., 2020; Zhang et al., 2021).

Adversarial Questions Table 3 reports results for GEMMA2-9B's Utility Ranker trained on NaturalQuestions and evaluated zero-shot on to predict answer uncertainty for retrieval augmented QA with $|R| = 5$ on PopQA and RefuNQ. These datasets are made of adversarial cases so we report AUROC (predicting incorrect answers) and AURAC (summary of different rejection thresholds). The Utility Ranker (NQ) is effective at detecting answer uncertainty across datasets and improving the QA accuracy by refusing to answer questions. In particular, we observe superior performance of the Utility Ranker on PopQA and RefuNQ. We attribute this to the fact that, either due to knowledge about tail entities or unanswerable questions about nonexistent concepts, the quality of the retrieved passages suffers. Thus, our approach will assign lower utility to these predicting answer uncertainty. Interestingly, information based methods, MSP and PPL, perform worse in this adversarial QA tasks than in the in-distribution test cases (Section 5.1). This suggest that in these cases models produce incorrect answer with potentially high confidence. On RefuNQ, GEMMA2-9B with its Utility Ranker achieves comparable performance to GPT-4-`0613`'s in refusal (65%) and accuracy (53.7) as reported in Liu et al. (2024a).

5.3 IMPROVING QA PERFORMANCE

We also assess the quality of the passage utility scores to identify informative passages via end task QA performance. We compare the original ranking by the external retrieval system with the ranking established by the utility scores by taking the top 3 passages out of 10 passages ordered by the external retriever and re-ranked by the Utility Ranker. We then run the QA models with a budget of $|R| = 3$ input passages. We also run the QA model with the all the 10 passages, i.e., with $|R| = 10$. Results for GEMMA2-9B QA model are shown in Table 4.

Table 4: Retrieval augmented QA performance with three passages $|R| = 3$ is the version with the top three retrieved passages from Contriever and $|R^{UR}| = 3$ is the version with top three re-ranked passages out of ten originally retrieved. We report model based (AccLM) accuracy.

	NaturalQuestions	TriviaQA	WebQuestions	SQuAD
$\|R\| =$ top 3 ranked by external retriever	0.58	0.77	0.63	0.53
$\|R\| =$ top 3 re-ranked by Utility Ranker	0.62	0.79	0.65	0.60
$\|R\| =$ all 10 passages	**0.64**	**0.80**	**0.66**	**0.62**

The QA model with the top 3 passages re-ranked by the Utility Ranker improves 4 points on NaturalQuestions and 7 points on SQuAD over the QA model variant that takes the top 3 ranked by the external retrieval system. This suggest that passages considered relevant for user questions by the external retriever do not coincide with what is useful for the target QA model. Moreover, the QA model variant with the top 3 passages re-ranked by the Utility Ranker performs very close, i.e., difference of 1 or 2 points across all datasets, to the QA model variant with the 10 passages given as context. This shows that the utility scores are effective at identifying informative passages and can achieve comparable performance with less than a half of the prompt size.

6 CONCLUSIONS

In this work we present an approach to answer uncertainty prediction for retrieval augmented QA models. Importantly, this approach relies on single passage utilities. This approach is based on a small neural model that is trained on a target QA model judgements of retrieved passage usefulness. We show that this approach is competitive or better than existing strong error prediction approaches while being light-weight. Our experiments also show that our approach is particularly better in cases of extreme QA model answer uncertainty like rare entities and unanswerable questions. Future work would explore the approach in the context of log-form generation tasks, e.g., query focused-generation. It would also be interesting to explore to what extent the Utility Ranker model could be used in active learning scenarios.

6.1 ETHICS STATEMENT

Our work does not involve human subjects. We use QA datasets that are publicly available and widely used by the research community.

6.2 REPRODUCIBILITY STATEMENT

We build up on existing base code Farquhar et al. (2024); Fang et al. (2024) and we will make available all code and data together with the docker images for reproducibility.

REFERENCES

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667. URL `https://aclanthology.org/2024.tacl-1.38`.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=hSyW5go0v8`.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural*

Language Processing, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1160`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL `https://aclanthology.org/P17-1171`.

Jifan Chen, Eunsol Choi, and Greg Durrett. Can NLI models verify QA systems' predictions? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3841–3854, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.324. URL `https://aclanthology.org/2021.findings-emnlp.324`.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5186–5200, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.acl-long.283`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Li Dong, Chris Quirk, and Mirella Lapata. Confidence modeling for neural semantic parsing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 743–753, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1069. URL `https://aclanthology.org/P18-1069`.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-polygraph: Uncertainty estimation for language models. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.41. URL `https://aclanthology.org/2023.emnlp-demo.41`.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL `https://aclanthology.org/P18-1082`.

Haishuo Fang, Jeet Gor, and Edwin Simpson. Efficiently acquiring human feedback with Bayesian deep learning. In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe (eds.), *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 70–80, St Julians, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.uncertainlp-1.7`.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024. URL `https://doi.org/10.1038/s41586-024-07421-0`.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/gal16.html`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/guu20a.html`.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=byxXa99PtF`.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL `https://aclanthology.org/2021.eacl-main.74`.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=NTEz-6wysdb`.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=jKN1pXi7b0`.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1), mar 2024. ISSN 1532-4435.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 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, 09 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00407. URL `https://doi.org/10.1162/tacl_a_00407`.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL `https://aclanthology.org/P17-1147`.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL `https://arxiv.org/abs/2207.05221`.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.503. URL `https://aclanthology.org/2020.acl-main.503`.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL `https://aclanthology.org/2020.emnlp-main.550`.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime qa: what's the answer right now? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=VD-AYtP0dve`.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL `https://aclanthology.org/Q19-1026`.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL `https://doi.org/10.1145/3600006.3613165`.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=eGLdVRvvfQ`. Expert Certification.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1eA7AEtvS`.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL `https://aclanthology.org/P19-1612`.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=8s8K2UZGTZ`.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. Examining llms' uncertainty expression towards questions outside parametric knowledge, 2024a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL `https://aclanthology.org/2024.tacl-1.9`.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL `https://aclanthology.org/2023.acl-long.546`.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL `https://aclanthology.org/2022.tacl-1.50`.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. A discrete hard EM approach for weakly supervised question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2851–2864, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1284. URL `https://aclanthology.org/D19-1284`.

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. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

SeongIl Park, Seungwoo Choi, Nahyun Kim, and Jay-Yoon Lee. Enhancing robustness of retrieval-augmented language models with in-context learning. In Wenhao Yu, Weijia Shi, Michihiro Yasunaga, Meng Jiang, Chenguang Zhu, Hannaneh Hajishirzi, Luke Zettlemoyer, and Zhihan Zhang (eds.), *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, pp. 93–102, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.knowledgenlp-1.7. URL `https://aclanthology.org/2024.knowledgenlp-1.7`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, L. Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, S. Mc Carthy, Sarah Perrin, S'ebastien Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Brian Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Cl'ement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118, 2024. URL `https://api.semanticscholar.org/CorpusID:270843326`.

Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2395–2400, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657957. URL `https://doi.org/10.1145/3626772.3657957`.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao

Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL `https://aclanthology.org/2021.naacl-main.52`.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6138–6148, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.496. URL `https://aclanthology.org/2021.emnlp-main.496`.

D. Sculley. Combined regression and ranking. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pp. 979–988, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300551. doi: 10.1145/1835804.1835928. URL `https://doi.org/10.1145/1835804.1835928`.

Edwin Simpson, Yang Gao, and Iryna Gurevych. Interactive Text Ranking with Bayesian Optimization: A Case Study on Community QA and Summarization. *Transactions of the Association for Computational Linguistics*, 8:759–775, 12 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00344. URL `https://doi.org/10.1162/tacl_a_00344`.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 311–325, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.18. URL `https://aclanthology.org/2024.naacl-long.18`.

Junya Takayama and Yuki Arase. Relevant and informative response generation using pointwise mutual information. In Yun-Nung Chen, Tania Bedrax-Weiss, Dilek Hakkani-Tur, Anuj Kumar, Mike Lewis, Thang-Minh Luong, Pei-Hao Su, and Tsung-Hsien Wen (eds.), *Proceedings of the First Workshop on NLP for Conversational AI*, pp. 133–138, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4115. URL `https://aclanthology.org/W19-4115`.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL `https://aclanthology.org/2023.emnlp-main.330`.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9690–9700. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/van-amersfoort20a.html`.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. 2020. URL `https://arxiv.org/abs/2003.02037`.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*, 2024. URL `https://arxiv.org/abs/2402.17497`.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL `https://aclanthology.org/N18-1101`.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=mlJLVigNHp`.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=ZS4m74kZpH`.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models, 2023. URL `https://arxiv.org/abs/2311.09210`.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1701–1722, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.eacl-long.102`.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. Knowing more about questions can help: Improving calibration in question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1958–1970, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.172. URL `https://aclanthology.org/2021.findings-acl.172`.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5506–5524, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.335. URL `https://aclanthology.org/2023.emnlp-main.335`.

A DATASETS

Table 5 shows statistics about the QA datasets we use in our experiments.

B MODEL PROMPTS

The prompt we use for our QA models is shown in Table 6. Table 7 illustrate the prompts used for our LLM based accuracy and p(true) baseline.

Dataset	Train	Dev	Test
Natural Questions	79,168	8,757	3,610
TriviaQA	78,785	8,837	11,313
WebQuestions	2,474	361	2,032
SQuAD	78,713	8,886	10,570
PopQA	11267	-	3000
RefuNQ	-	-	4439

Table 5: Dataset statistics, number of instances per train/dev/test sets. Note that we sample a smaller test set for PopQA in our experiments.

Knowledge:

[1] [passage]
[2] [passage]
...
[N] [passage]

Answer the following question with a very short phrase.

Question: [question]

Table 6: Minimal prompt selected as user turn for the QA models.

C COMPARISON AND BASELINE UNCERTAINTY ESTIMATION METHODS

C.1 TEST TIME COST OF UNCERTAINTY ESTIMATION METHODS

Table 8 shows the cost of executing uncertainty estimation for a user question x in terms of model inference calls required. Simple information theoretic methods require a single call (PPL, MSP) or two (PMI) calls to the QA model with the full prompt (N retrieved passages and user question x); similarly the Ans.Len baseline. However, approaches that estimate uncertainty based on diversity (Regular Entropy, Cluster Assignment, Semantic Entropy, and p(true)) require generating N answers, i.e., N inference calls with the full prompt. In addition, Cluster Assignment and Semantic Entropy require the computation of the answers clusters, so additional calls to an entailment model are required to compare the set of sampled answers. p(true) requires one additional LLM call to elicit a True/False answer but with a very long prompt including in-context examples and candidate answers. In contrast, our approach requires $|R|$ utility predictions with a BERT-size model.

D ADDITIONAL RESULTS

D.1 DIFFERENT COMPONENTS OF THE TRAINING OBJECTIVE

Table 9 shows results on the ablation of the Utility Ranker training objective (Section 3.1, Equation 3). When trained only with the ranking loss ($\mathcal{L}_{rank}$), in average it achieves better performance when the training signal combines accuracy (a) with entailment (e), i.e., the training ranking is given by $(e+a)/2$. When trained in combination with the full objective ($\mathcal{L}_{rank}$ + $\mathcal{L}_{BCE}$) the ranker shows an increase of 10 AUROC points. Highlighting the benefit of training the Utility Ranker to predicting QA accuracy for input passages. Interestingly, when we drop the ranking loss (i.e., last line of Table 9) there is a drop in performance. On one hand, the ranking loss enables the comparison of pairs of passages and thus the number of training instances is higher. On the other hand, the entailment -based ranking signal might help the final model to learn features useful for more accurate passage utility prediction.

D.2 UNCERTAINTY ESTIMATION RESULTS

Table 10 and 11 shows retrieval augmented QA performance on the development set for the target QA models. Table 12, 13, and 14 show performance of uncertainty quantification approaches.

You need to check whether the prediction of a question-answering system to a question is correct. You should make the judgment based on a list of ground truth answers provided to you. Your response should be ”correct” if the prediction is correct or ”incorrect” if the prediction is wrong.

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: [”William Shakespeare”, ”Roma Gill”]
Prediction: W Shakespeare
Correctness: correct

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: [”William Shakespeare”, ”Roma Gill”]
Prediction: Roma Gill and W Shakespeare
Correctness: correct

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: [”William Shakespeare”, ”Roma Gill”]”
Prediction: Roma Shakespeare
Correctness: incorrect

Question: What country is Maharashtra Metro Rail Corporation Limited located in?
Ground truth: [”India”]
Prediction: Maharashtra
Correctness: incorrect

Question: What’s the job of Song Kang-ho in Parasite (2019)?
Ground truth: [”actor”]
Prediction: He plays the role of Kim Ki-taek, the patriarch of the Kim family.
Correctness: correct

Question: Which era did Michael Oakeshott belong to?
Ground truth: [”20th-century philosophy”]
Prediction: 20th century.”
Correctness: correct

Question: Edward Tise (known for Full Metal Jacket (1987)) is in what department?
Ground truth: [”sound department”]
Prediction: 2nd Infantry Division, United States Army
Correctness: incorrect

Question: What wine region is Finger Lakes AVA a part of?
Ground truth: [”New York wine”]
Prediction: Finger Lakes AVA
Correctness: incorrect

Question: [question]
Ground truth: [answers]
Prediction: [output]
Correctness:

Table 7: Prompt for accuracy evaluation.

	Nb./Type of Inference Call at Test Time
PPL	$1\,$LLM-G
MSP	$1\,$LLM-G
PMI	$2\,$LLM-G
p(true)	$(N+1)\,$LLM-G $+\,1\,$LLM-E
Regular Entropy	$(N+1)\,$LLM-G
Cluster Assignment	$(N+1)\,$LLM-G $+\,N(N-1)/2\,$LLM-E
Semantic Entropy	$(N+1)\,$LLM-G $+\,N(N-1)/2\,$LLM-E
Ans.Len	$1\,$LLM-G
Retriever Score	$0\,$LLM-G
Utility Ranker	$\|R\|$ Bert-F

Table 8: Number and type of inference call required to estimate answer uncertainty for a given user question x. LLM-G means inference with the retrieval augmented QA model, i.e., a forward pass with the prompt including the set of R retrieved passages and the question to generate an answer. LLM-E is inference with an evaluation model, e.g., a forward pass to ask a LLM for correctness in p(true) or a forward pass with an entailment model in the Semantic Entropy method. Bert-F is an inference call to predict passage utility for a passage p in R and user question x.

Table 9: Uncertainty Estimation by the Utility Ranker trained with variants of the training objective. We report AUROC and AURAC for the Utility Ranker for the three target QA models (GEMMA2-9B, LLAMA3.1-8B, and MISTRAL-7B-V0.3) on Natural Questions development data.

	GEMMA2-9B		LLAMA3.1-8B		MISTRAL-7B-V0.3	
	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
$\mathcal{L}_{rank}, (e+a)/2 + \mathcal{L}_{BCE}$	0.77	0.76	0.77	0.76	0.79	0.76
$\mathcal{L}_{rank}, (e+a)/2$	0.67	0.70	0.66	0.70	0.69	0.70
$\mathcal{L}_{rank}, (a)$	0.62	0.67	0.64	0.68	0.67	0.69
$\mathcal{L}_{rank}, (e)$	0.67	0.70	0.64	0.68	0.64	0.67
$\mathcal{L}_{BCE}$	0.76	0.74	0.75	0.74	0.77	0.74

Table 10: Target QA models performance on test sets with $|R| = 5$. Model based accuracy AccLM (column header ALM) is accuracy computed by Qwen2-72B-Instruct.

	NaturalQuestions		TriviaQA		WebQuestions		SQuAD		PopQA		RefuNQ	
	Acc	ALM	Acc	ALM	Acc	ALM	Acc	ALM	Acc	ALM	Acc	ALM
GEMMA2.9B	0.46	0.61	0.73	0.78	0.40	0.64	0.41	0.58	0.49	0.51	0.26	0.35
LLAMA-3.1-8B	0.47	0.60	0.71	0.77	0.44	0.66	0.41	0.56	0.48	0.49	0.27	0.37
MISTRAL-7B-V0.3	0.47	0.58	0.71	0.75	0.47	0.66	0.40	0.57	0.52	0.49	0.27	0.35

Table 11: Target QA models performance on the development sets with $|R| = 5$. (Acc) is rule based accuracy as used in previous work, (AccLM) is accuracy computed by Qwen2-72B-Instruct.

	Natural Questions		TriviaQA		WebQuestions		SQuAD	
	Acc	AccLM	Acc	AccLM	Acc	AccLM	Acc	AccLM
GEMMA2.9B	0.45	0.62	0.73	0.79	0.45	0.67	0.37	0.58
LLAMA-3.1-8B	0.46	0.60	0.71	0.77	0.52	0.68	0.38	0.58
MISTRAL-7B-V0.3	0.46	0.60	0.71	0.76	0.53	0.69	0.36	0.57

Table 12: Uncertainty estimation for GEMMA2-9B development set.

	Natural Questions		TriviaQA		WebQuestions		SQuAD	
	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
PPL	0.67	0.69	0.61	0.80	0.63	0.70	0.65	0.66
MSP	0.69	0.70	0.66	0.81	0.64	0.70	0.66	0.66
PMI	0.49	0.59	0.42	0.71	0.49	0.63	0.46	0.55
p(true)	0.73	0.73	0.76	0.85	0.73	0.75	0.70	0.69
Regular Entropy	0.70	0.69	0.66	0.81	0.65	0.70	0.68	0.68
Cluster Assignment	0.70	0.70	0.67	0.81	0.65	0.70	0.65	0.66
Semantic Entropy	0.71	0.71	0.65	0.80	0.65	0.71	0.65	0.66
Ans.Len	0.63	0.66	0.62	0.79	0.61	0.69	0.60	0.64
Retriever Score	0.59	0.65	0.62	0.80	0.50	0.62	0.67	0.68
Utility Ranker	**0.75**	**0.74**	**0.79**	**0.86**	**0.74**	**0.77**	**0.82**	**0.77**

Table 13: Uncertainty estimation for LLAMA3.1-8B development set.

	Natural Questions		TriviaQA		WebQuestions		SQuAD	
	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
PPL	0.75	0.75	0.80	0.85	0.68	0.73	0.71	0.70
MSP	0.79	0.77	0.83	0.86	0.69	0.73	0.72	0.70
PMI	0.61	0.68	0.56	0.75	0.55	0.67	0.55	0.60
p(true)	0.79	0.77	**0.89**	**0.88**	0.72	0.75	0.69	0.69
Regular Entropy	**0.81**	**0.78**	0.82	0.86	0.69	0.74	0.75	0.72
Cluster Assignment	0.77	0.75	0.82	0.85	0.72	0.75	0.75	0.72
Semantic Entropy	0.76	0.75	0.84	0.86	0.71	0.75	0.76	0.73
Ans.Len	0.63	0.67	0.66	0.79	0.61	0.69	0.56	0.60
Retriever Score	0.57	0.65	0.62	0.78	0.49	0.64	0.67	0.67
Utility Ranker	0.79	0.77	0.81	0.85	**0.77**	**0.79**	**0.83**	**0.76**

Table 14: Uncertainty estimation for MISTRAL-7B-V0.3 development set.

	Natural Questions		TriviaQA		WebQuestions		SQuAD	
	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC	AUROC	AURAC
PPL	0.65	0.69	0.65	0.80	0.62	0.70	0.66	0.65
MSP	0.70	0.71	0.74	0.82	0.67	0.73	0.72	0.68
PMI	0.49	0.60	0.57	0.76	0.56	0.68	0.54	0.58
p(true)	0.73	0.71	**0.80**	**0.85**	0.69	0.75	0.70	0.67
Regular Entropy	0.65	0.69	0.66	0.80	0.63	0.71	0.70	0.68
Cluster Assignment	0.71	0.72	0.76	0.82	0.71	0.75	0.75	0.69
Semantic Entropy	0.72	0.72	0.77	0.83	0.71	0.74	0.75	0.70
Ans.Len	0.65	0.68	0.69	0.80	0.64	0.72	0.66	0.64
Retriever Score	0.59	0.65	0.61	0.77	0.58	0.69	0.64	0.63
Utility Ranker	**0.76**	**0.74**	0.77	0.84	**0.73**	**0.77**	**0.80**	**0.72**