

Overlap-Based OOD Detection in Closed-Book Question Answering

Dries Wedda and Matias Valdenegro-Toro

Department of Artificial Intelligence, University of Groningen

The Netherlands

dries.wedda@student.uva.nl, m.a.valdenegro.toro@rug.nl

Abstract

Closed-book question answering requires models to answer using knowledge stored in their parameters, making uncertainty estimation important when that knowledge is limited. We study how total, aleatoric, and epistemic uncertainty respond to progressively weaker train-test overlap, from question paraphrase overlap to answer-only overlap to no overlap, and how well these estimates detect overlap-based out-of-distribution questions. Using T5-Large-SSM on WebQuestions and Open Natural Questions, we estimate token-level uncertainty with Deep Ensembles and Monte Carlo Dropout and sequence-level uncertainty with Deep Ensembles. Weaker overlap is associated with higher total and epistemic uncertainty, but aleatoric uncertainty also tends to increase. Epistemic uncertainty performs best for OOD detection on Open Natural Questions, but does not consistently outperform total or aleatoric uncertainty on WebQuestions. These findings show that train-test overlap is reflected in uncertainty estimates, but the response to weaker overlap is not consistently isolated to the epistemic component.

1 Introduction

Language models are increasingly used as interfaces for answering factual questions. However, fluent and confident generation does not guarantee correctness. Models may produce plausible but false or unsupported answers, commonly referred to as hallucinations or confabulations. One cause of such behaviour is a knowledge boundary, where a question falls beyond the information that the model can reliably recover from its parameters (Huang et al., 2025).

Closed-book question answering (CBQA) provides a controlled setting for studying this problem. A CBQA model answers factual questions using only knowledge encoded in its parameters, without retrieving external sources at inference time

(Roberts et al., 2020). Ideally, such a model should answer when it possesses the required knowledge and indicate uncertainty when it does not.

Not all uncertainty indicates a knowledge gap. Aleatoric uncertainty arises from irreducible variability in the possible outputs, such as an ambiguous question or multiple valid answers. Epistemic uncertainty instead reflects limitations in the model’s knowledge and may be reduced through additional data. This distinction can be practically relevant. High epistemic uncertainty may require information retrieval, whereas aleatoric uncertainty may require clarification or providing one of multiple correct answers. Distinguishing these sources could therefore make uncertainty more informative than a single predictive uncertainty score. Figure 1 illustrates this distinction. Despite having similar total uncertainty, the two predictions differ substantially in their estimated aleatoric and epistemic components.

We study uncertainty across the train-test overlap partitions of Lewis et al. (2021). These distinguish test questions with a paraphrase or near-duplicate in training, questions for which only the answer occurs in training, and questions with neither question nor answer overlap. If epistemic uncertainty reflects limitations in learned knowledge, it should increase as overlap weakens, whereas aleatoric uncertainty should not respond. This motivates the following research questions:

RQ1 Does weaker train-test overlap correspond to higher epistemic uncertainty while leaving aleatoric uncertainty approximately unchanged?

RQ2 How well do the uncertainty estimates detect out-of-distribution questions, where distribution shift is defined as decreasing overlap with the training data?

We evaluate these questions on WebQuestions and Open Natural Questions using Deep Ensembles and Monte Carlo Dropout.

Q: What countries included in Oceania?

Reference answers: Northern Mariana Islands; Norfolk Island; Guam; American Samoa; (+6 more)

Prediction: Northern Mariana Islands



Q: What part of DC is Howard University in?

Reference answers: Northwest Washington; Northwest; in Northwest Washington

Prediction: Northeastern



Figure 1: Two Deep Ensemble predictions with similar total uncertainty but different estimated aleatoric and epistemic components. The first question permits multiple valid answers and has a larger aleatoric component. The second asks for a single factual location, but the model incorrectly predicts *Northeastern* instead of *Northwest Washington*, and has a larger epistemic component.

The contributions of this paper are: (1) evaluating out-of-distribution detection based on the train-test overlap for closed-book question answering. (2) Evaluating Deep Ensembles at both the token and sequence level and MC Dropout at the token level for this task, we find that weaker overlap leads to increased total and epistemic uncertainty, going beyond Lewis et al. (2021), which shows degraded QA performance in the same setting, and we relate uncertainty to this degradation. (3) We additionally evaluate the use of soft labels and training set reduction to evaluate uncertainty disentanglement, and selective prediction using each uncertainty component.

2 Related Work

2.1 Closed-Book Question Answering and Train-Test Overlap

Roberts et al. (2020) showed that pretrained T5 models can be fine-tuned to answer factual questions using knowledge stored in their parameters. Performance improved substantially with model scale, demonstrating the capacity of pretrained large language models to learn factual knowledge. These results established closed-book question answering as a setting for evaluating parametric knowledge.

However, question-answering performance does not distinguish generalization from reliance on overlap with the training data. Lewis et al. (2021) analysed WebQuestions, Open Natural Questions, and TriviaQA and found that 58-71% of test answers also occurred in the corresponding training sets, while 28-34% of test questions had a paraphrase or near-duplicate in training. Closed-book models performed substantially worse on questions without train-test overlap. We build upon

these overlap annotations to study how uncertainty changes as overlap with the training data becomes weaker.

2.2 Uncertainty Quantification

Uncertainty may arise either from inherent variability that *cannot* be reduced with more data, such as ambiguity or noise, or from a lack of knowledge that *can* be reduced with more data. These sources are commonly associated with aleatoric and epistemic uncertainty, respectively.

A common method to estimate these components is to apply an information-theoretic decomposition to predictive uncertainty. Under this decomposition, epistemic uncertainty is quantified through disagreement between predictions obtained under different parameter configurations. Deep Ensembles obtain these different predictive distributions from independently trained models (Lakshminarayanan et al., 2017), whereas Monte Carlo Dropout obtains them from stochastic forward passes by retaining dropout at inference time and has been interpreted as approximate Bayesian inference (Gal and Ghahramani, 2016).

The entropy of the averaged predictive distribution represents total uncertainty, the mean entropy of the individual distributions represents aleatoric uncertainty, and their difference represents epistemic uncertainty. This leads to an additive relation between aleatoric, epistemic, and total uncertainty. Malinin and Gales (2021) extended this framework to autoregressive structured prediction, distinguishing uncertainty at the level of individual tokens from uncertainty over complete sequences.

Sequence-level uncertainty in natural language generation is complicated by the possibility that different output strings express the same meaning. Entropy over distinct strings may therefore reflect

variation in wording rather than uncertainty about the underlying answer. Kuhn et al. (2023) address this by clustering generations into semantic equivalence classes using bidirectional entailment and computing entropy over the resulting meanings.

3 Evaluating Overlap OOD in Closed-Book QA

3.1 Datasets

In closed-book question answering (CBQA), a model must answer a question from its parametric knowledge alone, without access to any retrieved passage or supporting context at inference time. Each data point (x, Y) consists of a question x and a set of reference answers Y .

Our experiments use WebQuestions (WQ; Be-rant et al., 2013) and Open Natural Questions (NQ-Open; Lee et al., 2019). We selected these datasets because they provide short-form answer supervision and because train-test overlap annotations are available for both.

WebQuestions. WQ contains questions collected using the Google Suggest API. Annotators answered each question using only the Freebase page associated with the entity mentioned in the question. The original release contains 3,778 training questions and 2,032 test questions. We use the original test set together with the publicly available filtered training and development sets distributed with the Dense Passage Retriever resources (Karpukhin et al., 2020), containing 2,474 training, 278 development, and 2,032 test examples.

Open Natural Questions. NQ-Open is derived from Natural Questions (Kwiatkowski et al., 2019), which was constructed from anonymized queries submitted to the Google search engine. Annotators were shown associated Wikipedia articles and identified long-answer regions and, where applicable, one or more short-answer spans. NQ-Open retains question-answer pairs whose short answers contain at most five tokens (Lee et al., 2019). We use splits containing 79,168 training, 8,757 development, and 3,610 test examples.

The sizes of the dataset splits are summarized in Table 1.

Train-test overlap partitions. To examine how uncertainty varies with the degree of overlap between the training and test data, we use the annotations of Lewis et al. (2021). For our analyses, we

	WQ	NQ-Open
<i>Dataset splits</i>		
Train	2,474	79,168
Development	278	8,757
Test	2,032	3,610
<i>Train-test overlap partitions</i>		
Question overlap	271	318
Answer overlap only	299	315
No answer overlap	853	1,307
Total	1,423	1,940

Table 1: Dataset statistics for WebQuestions (WQ) and Open Natural Questions (NQ-Open). The lower block reports the test-set overlap partitions of Lewis et al. (2021), defined by question and answer overlap with the training data. In our OOD analyses, question overlap, answer overlap only, and no answer overlap are treated as ID, Near-OOD, and Far-OOD, respectively.

use the following in-distribution (ID) and out-of-distribution (OOD) labels for the three annotated categories:

- **Question overlap (ID):** the test question is a paraphrase or near-duplicate of a training question.
- **Answer overlap only (Near-OOD):** no matching training question is identified, but at least one reference answer occurs in the training data.
- **No answer overlap (Far-OOD):** no matching training question is identified and none of the reference answers occurs in the training data.

These labels represent progressively weaker overlap with the training data. We use the partitions to compare uncertainty across overlap levels and to evaluate OOD detection. Because the annotations cover only a subset of the test set, these analyses are restricted to the annotated examples. Counts are provided in Table 1.

Preprocessing and answer normalization. We apply light rule-based cleaning to questions and answers. We remove Wikipedia-style citation markers, standardize selected Unicode characters, and collapse repeated whitespace.

For comparisons between answer strings, we normalize by lowercasing text, removing punctuation and articles, and collapsing whitespace. We use this for exact-match evaluation and for the normalized exact-match clusters used in sequence-level uncertainty estimation.

3.2 Model Architecture

We fine-tune T5-Large-SSM, an encoder-decoder transformer with 770M parameters, 24 encoder blocks and 24 decoder blocks (Raffel et al., 2020). We selected this model as a compromise between strong CBQA performance and computational feasibility. The model applies dropout in the self-attention, cross-attention, and feed-forward sublayers. The model is pretrained with the standard T5 objective, followed by continued pretraining using the salient span masking (SSM) objective on Wikipedia (Roberts et al., 2020). SSM, inspired by REALM (Guu et al., 2020), has been shown to improve closed-book QA performance over standard T5 (Roberts et al., 2020).

3.3 Model Training

All models are initialized from the pretrained T5-Large-SSM checkpoint and fine-tuned separately on WQ and NQ-Open. Questions are formatted by prepending the task prefix ‘question: ’ and tokenized using the T5 tokenizer. Input and target sequences are truncated or padded to maximum lengths of 32 tokens.

Training objective. Let x denote a question and $Y = \{y^{(1)}, \dots, y^{(K)}\}$ its set of reference answers. Training uses teacher forcing, such that the decoder conditions at each position on a reference prefix. Each question is paired with a single reference. When multiple references are available, we deterministically select the first, $y^{(1)} = (y_1^{(1)}, \dots, y_L^{(1)})$. Its sequence-level negative log-likelihood is

$$\mathcal{L}(x, y^{(1)}) = - \sum_{i=1}^L \log p(y_i^{(1)} | x, y_{<i}^{(1)}),$$

where $p(\cdot | x, y_{<i})$ denotes the model’s predictive distribution over the vocabulary at decoder position i . Padding positions are excluded, and losses are averaged over all non-padding target positions in the minibatch.

Optimization and training procedure. We train each model for a fixed number of epochs using AdamW. We do not apply per-run early stopping. In preliminary experiments, development loss reached its minimum before generated-answer performance stabilized, whereas development exact match continued to improve. We therefore selected a common training duration from preliminary development curves and held it fixed across ensemble

members and random seeds. Detailed hyperparameters are reported in Appendix A.

3.4 Uncertainty Quantification

We estimate uncertainty using Deep Ensembles (DE) and Monte Carlo Dropout (MCD). For DE, we fine-tune 5 models independently using different random seeds. For MCD, we keep the standard dropout layers of T5 during inference. Details are in Appendix A.

We write M for the number of parameter realizations used to estimate uncertainty. We use $M = 5$ ensemble members for DE and $M = 30$ stochastic forward passes for MCD. We additionally evaluate a deterministic single-model setting ($M = 1$), using fixed parameters and disabling dropout during inference, as a baseline for both QA performance and uncertainty estimation.

We estimate uncertainty at two levels of granularity. At the token level, uncertainty is computed from the predictive distributions over the vocabulary at each decoding position. At the sequence level, uncertainty is computed from empirical distributions over clusters of sampled complete answers.

3.4.1 Token-level Uncertainty

Let $\mathcal{V}$ denote the model vocabulary. We write

$$p_i^{(m)}(v) = p_{\theta_m}(v | x, \hat{y}_{<i})$$

for the probability assigned to token $v \in \mathcal{V}$ at decoding position i by parameter realization θ_m with $m \in \{1, \dots, M\}$, conditioned on the input question x and the shared generated prefix $\hat{y}_{<i}$. We average these predictive distributions as

$$\bar{p}_i(v) = \frac{1}{M} \sum_{m=1}^M p_i^{(m)}(v),$$

and decode the next token greedily,

$$\hat{y}_i = \arg \max_{v \in \mathcal{V}} \bar{p}_i(v).$$

The predictive distributions at the next position are then recomputed using the updated prefix $\hat{y}_{<i+1}$, continuing until the end-of-sequence token is generated. The token-level uncertainty estimates therefore correspond to the same answer $\hat{y}$.

We write the entropy of a pmf p over a discrete random variable $\mathcal{X}$ as

$$\mathbb{H}[p] := - \sum_{x \in \mathcal{X}} p(x) \log p(x).$$

The total uncertainty at position i is estimated by the predictive entropy of the averaged distribution:

$$T_i = \mathbb{H}[\bar{p}_i].$$

The aleatoric component is the mean entropy over the individual predictive distributions:

$$A_i = \frac{1}{M} \sum_{m=1}^M \mathbb{H}[p_i^{(m)}].$$

Their difference is the epistemic component,

$$E_i = T_i - A_i,$$

which is the mutual information between the next-token prediction and the sampled model realization.

To obtain a single, length-invariant uncertainty score for each predicted answer $\hat{y} = (\hat{y}_1, \dots, \hat{y}_L)$, we average the token-level quantities over decoding positions whose total uncertainty exceeds a threshold ε_{tok} . Applying the same filter to T , A , and E preserves additivity while excluding near-deterministic positions from the aggregation. We define

$$I_{\varepsilon_{\text{tok}}}(\hat{y}) = \{i \in \{1, \dots, L\} \mid T_i \geq \varepsilon_{\text{tok}}\},$$

and, writing $I = I_{\varepsilon_{\text{tok}}}$,

$$U_{\varepsilon_{\text{tok}}}(\hat{y}) = \begin{cases} 1/|I(\hat{y})| \sum_{i \in I(\hat{y})} U_i, & I(\hat{y}) \neq \emptyset, \\ 0, & I(\hat{y}) = \emptyset, \end{cases}$$

with $U \in \{T, A, E\}$. Note how this definition preserves the additive identity $T_{\varepsilon_{\text{tok}}}(\hat{y}) = A_{\varepsilon_{\text{tok}}}(\hat{y}) + E_{\varepsilon_{\text{tok}}}(\hat{y})$. In our analyses, we use $\varepsilon_{\text{tok}} = 0.1$.

3.4.2 Sequence-level Uncertainty

Sequence-level uncertainty is estimated from empirical distributions over complete generated answers. For each parameter realization θ_m , we draw $N = 50$ answers from the model by multinomial sampling at temperature $\tau = 1$, yielding samples $(\hat{y}^{(m,1)}, \dots, \hat{y}^{(m,N)})$.

Each sampled answer is normalized using the procedure described in Section 3.1. Pooling the normalized samples across all M realizations gives the set of distinct answers $\mathcal{C} = \{C_1, \dots, C_K\}$, which we treat as answer clusters. We use normalized exact match to deterministically group answers that are treated as semantically equivalent. The standard alternative is semantic clustering by bidirectional entailment (Kuhn et al., 2023), which

is designed for free-form generation, where a single answer admits many surface realizations. Both datasets used here contain short-form answers, so the space of paraphrases is limited. On a sample of our data, entailment-based clustering produced identical clusters at substantially greater computational cost, and we therefore retain the exact-match criterion.

For each parameter realization θ_m , let

$$q^{(m)}(C_k) = \frac{1}{N} \sum_{n=1}^N \mathbb{1}[\hat{y}^{(m,n)} = C_k], \quad C_k \in \mathcal{C}$$

denote its empirical probability distribution over $\mathcal{C}$. We obtain the predictive cluster distribution by averaging across realizations:

$$\bar{q}(C_k) = \frac{1}{M} \sum_{m=1}^M q^{(m)}(C_k).$$

The predicted cluster is

$$\hat{C} = \arg \max_{C_k \in \mathcal{C}} \bar{q}(C_k),$$

which is used as the sequence-level prediction. The total sequence-level uncertainty is

$$T = \mathbb{H}[\bar{q}],$$

and the aleatoric component is

$$A = \frac{1}{M} \sum_{m=1}^M \mathbb{H}[q^{(m)}].$$

The epistemic component is their difference:

$$E = T - A.$$

Sequence-level uncertainty is evaluated only using DE. Extending the estimator to MCD would require drawing multiple complete answers under each fixed dropout realization. In our implementation, standard dropout masks are resampled across forward passes, so keeping the same masks throughout sequence generation would require a custom fixed-mask procedure.

3.5 Evaluation Metrics

QA performance. We evaluate answer quality using Exact Match (EM). After normalizing the predicted answer $\hat{y}(x)$ to a question x and each reference answer $y \in Y$, the question-level score is

$$\text{EM}(x, Y) = \mathbb{1}[\hat{y}(x) \in Y].$$

The exact match over a dataset is then the mean question-level score.

Method	WebQuestions (EM %, $\uparrow$)				NQ-Open (EM %, $\uparrow$)			
	Total	ID	Near-OOD	Far-OOD	Total	ID	Near-OOD	Far-OOD
<i>Token-level</i>								
Baseline	26.4	64.9	20.1	4.5	24.8	58.5	12.1	2.1
MC Dropout	26.2	62.7	21.7	3.9	23.9	56.9	12.1	1.9
Deep Ensemble	27.8	66.4	25.8	4.3	27.2	67.6	13.7	1.6
<i>Sequence-level</i>								
Baseline	26.7	65.7	19.7	4.7	25.0	59.4	12.7	2.0
Deep Ensemble	27.9	66.8	24.8	4.7	27.5	66.4	14.3	2.3

Table 2: Question-answering performance measured by exact match (%, $\uparrow$) on the test set, as well as overlap-based ID, near-OOD, and far-OOD partitions of WebQuestions and Open Natural Questions.

OOD detection. We evaluate OOD detection as a binary classification problem, where OOD samples are treated as the positive class and ID samples as the negative class. Given an uncertainty scoring function u , we rank all samples by $u(x)$ and report the area under the ROC curve (AUROC). Higher AUROC indicates better separation of ID and OOD examples. We evaluate ID vs. Near-OOD and ID vs. Far-OOD separately.

4 Experimental Results

4.1 QA Performance

Table 2 shows a sharp decline in exact-match accuracy as train-test overlap weakens, consistent with Lewis et al. (2021). Models perform substantially better on questions with training set paraphrases than on questions with only answer overlap, and worst on questions with no answer overlap.

The EM scores on the full test set are in the same broad range as the T5-Large results reported by Roberts et al. (2020), although direct comparison is limited by differences in data splits and training procedure. MCD performs slightly below the corresponding deterministic baseline on both datasets, whereas the DE improves performance. Token- and sequence-level predictions differ by at most 0.3 percentage points, indicating that the choice of prediction level has little effect on answer accuracy.

4.2 Uncertainty Across Partitions

Figure 2 shows the median uncertainty estimates across the train-test overlap partitions. Median total uncertainty and median epistemic uncertainty generally increase as train-test overlap weakens, with the clearest separation between partitions for the sequence-level DE. However, median aleatoric un-

certainty also tends to increase across the partitions. The median profiles therefore show the hypothesized increase in epistemic uncertainty with weaker overlap, but not a response that is specific to the epistemic component.

4.3 OOD Detection Performance

Table 3 shows that all reported AUROC scores exceed 0.5, with Far-OOD questions generally easier to detect than Near-OOD questions. The sequence-level DE achieves the strongest overall results, reaching AUROC values of 0.814 on WQ and 0.833 on NQ-Open for Far-OOD detection. On NQ-Open, epistemic uncertainty is the best predictor for both contrasts, whereas on WQ, total and aleatoric uncertainty are often comparable or better. The stochastic estimators also generally improve OOD detection over their corresponding deterministic baselines.

5 Discussion

5.1 Uncertainty across Train-Test Overlap Partitions

Weaker train-test overlap corresponded to higher median epistemic and total uncertainty, but aleatoric uncertainty also increased. The increase in epistemic uncertainty is consistent with its intended interpretation as reflecting gaps in acquired knowledge. That is, questions with less overlap with the training data tend to receive higher epistemic uncertainty. However, the simultaneous increase in aleatoric uncertainty shows that the response to weaker overlap is not specific to the epistemic component. The aleatoric increase is clearest and monotonic for the sequence-level DE on both datasets, but non-monotonic for token-level DE and

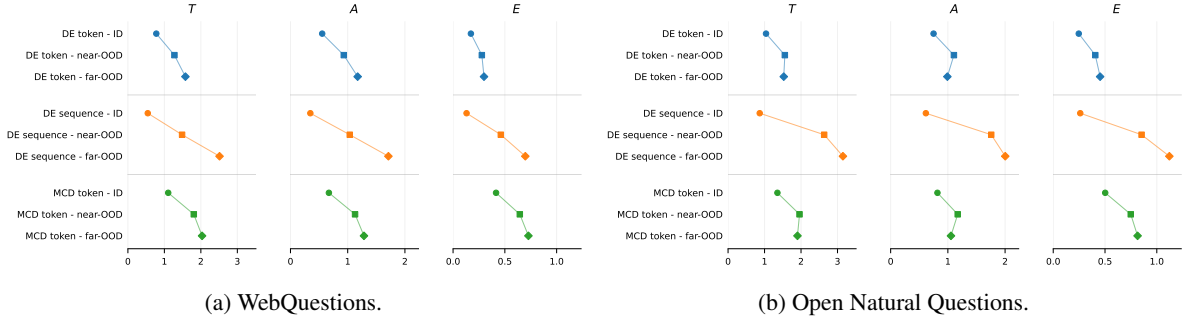


Figure 2: Question-level uncertainty profiles across train-test overlap partitions. Panels show total uncertainty T , aleatoric uncertainty A , and epistemic uncertainty E for the token-level Deep Ensemble, sequence-level Deep Ensemble, and token-level MC Dropout. Markers denote the median within the ID, near-OOD, and far-OOD partitions.

MCD on NQ-Open.

The accompanying decline in EM accuracy is consistent with [Lewis et al. \(2021\)](#), who found that QA performance depends strongly on train-test overlap. Our results show that this overlap structure is also reflected in predictive uncertainty. However, because aleatoric uncertainty increases alongside epistemic uncertainty, the decomposition does not isolate the association with weaker train-test overlap to the epistemic component.

Because this train-test overlap analysis is observational rather than interventional, the partitions contain different questions and therefore do not isolate the effect of overlap. That is, the lower-overlap questions may also be intrinsically more difficult or differ in answer frequency and ambiguity. Moreover, the labels reflect overlap with the fine-tuning set rather than knowledge acquired during pretraining, so the train-test overlap annotations do not perfectly reflect the model’s knowledge state. Complementary experiments in Appendix B perturb conditions associated with increased aleatoric and epistemic uncertainty and examine whether the estimated components respond in the expected directions through interventions.

5.2 OOD Detection

The OOD detection results show that predictive uncertainty contains useful information about train-test overlap. Far-OOD questions were generally easier to distinguish from ID questions than near-OOD questions, suggesting that uncertainty becomes a more informative detection signal as overlap with the training data weakens. The sequence-level Deep Ensemble performed strongest overall, in particular on distinguishing ID questions from Far-OOD questions.

The relative usefulness of the uncertainty components depended on the dataset. On NQ-Open, epistemic uncertainty achieved the highest AUROC for both OOD contrasts among the estimators. On WQ, however, total and aleatoric uncertainty were often comparable to epistemic uncertainty. Epistemic uncertainty therefore provides a useful signal of weaker train-test overlap, but does not consistently outperform total predictive uncertainty as an OOD detection score.

Aleatoric uncertainty obtaining comparable results to epistemic uncertainty is consistent with the partition-level results, where median aleatoric uncertainty also changed as overlap weakened. More broadly, comparable performance of aleatoric and epistemic uncertainty in OOD detection has also been observed in previous disentanglement studies ([Mucsányi et al., 2024](#); [de Jong et al., 2026](#)).

6 Conclusion

This paper investigated how total, aleatoric, and epistemic uncertainty respond to progressively weaker train-test overlap in closed-book question answering, and how well these uncertainty estimates distinguish overlap-based OOD questions. Using token-level uncertainty estimates from Deep Ensembles and Monte Carlo Dropout and sequence-level uncertainty estimates from Deep Ensembles, we evaluated these questions on WebQuestions and Open Natural Questions.

Weaker train-test overlap was associated with higher epistemic and total uncertainty, but aleatoric uncertainty also tends to increase. The expected epistemic response was therefore observed, but was not specific to the epistemic component. The estimates distinguished the overlap partitions moderately well, with Far-OOD questions generally

Dataset	Level	Estimator	AUC ID vs. Near-OOD			AUC ID vs. Far-OOD		
			T	A	E	T	A	E
WebQuestions	Token	Baseline	0.684	–	–	0.749	–	–
		MC Dropout	0.703	0.688	0.712	0.767	0.751	0.775
		Deep Ensemble	<u>0.715</u>	0.708	0.695	0.781	0.778	0.745
	Sequence	Baseline	0.681	–	–	0.775	–	–
		Deep Ensemble	0.716	0.712	0.710	0.814	0.808	<u>0.812</u>
NQ-Open	Token	Baseline	0.685	–	–	0.683	–	–
		MC Dropout	0.691	0.669	0.703	0.693	0.638	0.770
		Deep Ensemble	0.709	0.693	0.735	0.717	0.686	0.787
	Sequence	Baseline	0.700	–	–	0.741	–	–
		Deep Ensemble	<u>0.740</u>	0.733	0.744	<u>0.803</u>	0.783	0.833

Table 3: Out-of-distribution detection performance measured by AUROC ($\uparrow$), with ID (question overlap) questions treated as the negative class and Near-OOD (answer overlap only) or Far-OOD (no overlap) questions as the positive class. Total uncertainty T , aleatoric uncertainty A , and epistemic uncertainty E are evaluated as detection scores. For the deterministic baselines, only T is reported. Within each dataset and comparison, best results are shown in bold and pink, and second-best results are underlined and yellow.

easier to distinguish from ID than Near-OOD questions and with the sequence-level Deep Ensemble performing strongest overall. However, the relative performance of epistemic, aleatoric, and total uncertainty was dataset-dependent, and epistemic uncertainty did not consistently provide the best OOD detection score.

Future work should test the generality of these findings across additional pretrained architectures, including decoder-only language models. More direct annotation of uncertainty sources could also strengthen the evaluation. For example, AmbigNQ provides ambiguity annotations for questions from NQ-Open (Min et al., 2020), which could be used to test whether aleatoric uncertainty specifically reflects question ambiguity and to separate this effect from train-test overlap. Extending the evaluation from short-answer CBQA to free-form natural language generation would further test the uncertainty decomposition when valid outputs are highly diverse and token-level uncertainties must be aggregated across long, variable-length sequences.

Limitations

The experiments use a single pretrained model architecture, T5-Large-SSM, and two short-answer CBQA datasets. Although multiple uncertainty estimators and prediction levels are evaluated, the observed magnitudes and estimator-specific responses may not generalize to other language models, tasks, or uncertainty decompositions.

The results are reported as point estimates. Although DE uses multiple independently trained members and MCD uses multiple stochastic forward passes, these are used to construct a single uncertainty estimate, rather than as independent repetitions of the complete experiment. As a result, modest differences between estimators or uncertainty scores should be interpreted descriptively. Repeating the complete training pipeline and reporting confidence intervals would provide stronger evidence for the stability of the results.

Additional Results

The appendix contains additional results that did not fit into the main paper. This section provides short summaries. Section B presents additional uncertainty disentanglement experiments, mainly using soft labels and reduction in training set size, which affect aleatoric and epistemic uncertainty correspondingly. Section C performs and evaluates selective prediction using the total uncertainty and its disentanglement components, which assesses the use of uncertainty to filter incorrect answers, with the model abstaining on uncertain predictions.

Code

The code is available on Github at <https://github.com/dries02/cbqa-ood/>.

References

- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.
- Ivo Pascal de Jong, Andreea Ioana Sburlea, Matthia Sabatelli, and Matias Valdenegro-Toro. 2026. [Measuring orthogonality as the blind-spot of uncertainty disentanglement](#). *Preprint*, arXiv:2408.12175.
- Yarin Gal and Zoubin Ghahramani. 2016. [Dropout as a Bayesian approximation: Representing model uncertainty in deep learning](#). In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR.
- Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. [Bias-reduced uncertainty estimation for deep neural classifiers](#). In *International Conference on Learning Representations*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. [REALM: Retrieval-augmented language model pre-training](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Transactions on Information Systems*, 43(2).
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations*.
- 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. 2019. [Natural Questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). *Advances in Neural Information Processing Systems*, 30:6402–6413.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. [Latent retrieval for weakly supervised open domain question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096. Association for Computational Linguistics.
- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. [Question and answer test-train overlap in open-domain question answering datasets](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008. Association for Computational Linguistics.
- Andrey Malinin and Mark Gales. 2021. [Uncertainty estimation in autoregressive structured prediction](#). In *International Conference on Learning Representations*.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797. Association for Computational Linguistics.
- Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. 2024. [Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 50972–51038.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. Association for Computational Linguistics.
- Jeremias Traub, Till J Bungert, Carsten T Lüth, Michael Baumgartner, Klaus H Maier-Hein, Lena Maier-Hein, and Paul F Jäger. 2024. [Overcoming common flaws in the evaluation of selective classification systems](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 2323–2347.

A Hyperparameters

All models were trained on an NVIDIA GeForce RTX 5090 GPU with 32 GB VRAM. The training and inference settings used throughout the experiments are summarized below.

A.1 Training

Hyperparameter	Value
Pretrained checkpoint	T5-Large-SSM
Optimizer	AdamW
Initial learning rate	3×10^{-4}
Weight decay	0.01
Learning-rate schedule	Linear warm-up and decay
Warm-up duration	10%
Gradient clipping	Maximum norm of 1.0
Dropout probability	0.1 (NQ-Open), 0.2 (WQ)
Number of epochs	10
Batch size	64 (NQ-Open), 16 (WQ)
Maximum input length	32 tokens
Maximum target length	32 tokens
Question prefix	question:
Checkpoint retained	Final training epoch

Table 4: Training and architectural hyperparameters.

Weight decay was not applied to normalization parameters, the shared token-embedding matrix, or the relative attention-bias parameters. The learning rate was increased linearly during the first 10% of optimization steps and subsequently decayed linearly to zero. Models were evaluated on the development set after each epoch for monitoring, but no early stopping or development-based checkpoint selection was applied.

A.2 Uncertainty estimation and decoding

Setting	Value
Deep Ensemble members	5
MC Dropout forward passes	30
Token-level decoding	Greedy
Sequence-level sampling method	Multinomial sampling
Sequence-level temperature	1
Sequence samples per member	50
Maximum generation length	32 tokens
Token uncertainty threshold ε_{tok}	0.1

Table 5: Uncertainty estimation and decoding settings.

For MC Dropout, the same dropout probability is used for inference as during training.

B Disentanglement Experiments

Following de Jong et al. (2026), we perform two complementary interventions to examine whether the estimated aleatoric and epistemic components

respond in their intended directions. Soft-target training exposes the model to multiple valid answers and is intended primarily to increase aleatoric uncertainty, whereas reducing the amount of training data is intended to primarily increase epistemic uncertainty.

B.1 Soft-Target Training

To determine whether aleatoric uncertainty responds to variation in valid answers, we modify the training targets for questions with multiple reference answers. Under the standard ‘hard-target’ training procedure, as described in Section 3.3, each question is paired with a single deterministic reference. In the soft-target setting, we retain up to five reference answers per question, so that questions with many annotations do not excessively reduce repeated exposure to each complete answer sequence. Each time a training example (x, Y) is encountered, one reference $y^{(S)} = (y_1^{(S)}, \dots, y_{L_S}^{(S)}) \in Y$ is sampled uniformly and supplies the teacher-forced decoder prefix. At position i , we consider the references that share this prefix:

$$R_i^{(S)} = \left\{ k \mid y_{<i}^{(k)} = y_{<i}^{(S)} \right\}.$$

Because only the sampled reference supplies the decoder path, including many references would reduce how often the later tokens of each answer sequence are observed during training. Let

$$V_i^{(S)} = \left\{ y_i^{(k)} \mid k \in R_i^{(S)} \right\}$$

denote the set of distinct next-token continuations observed after that prefix. The target distribution assigns equal probability to these continuations:

$$q_i^{(S)}(v) = \begin{cases} 1/|V_i^{(S)}|, & v \in V_i^{(S)}, \\ 0, & \text{otherwise.} \end{cases}$$

For the sampled reference path, let

$$p_i^{(S)}(v) = p(v \mid x, y_{<i}^{(S)}), \quad v \in \mathcal{V},$$

denote the model’s predictive distribution at position i over its vocabulary $\mathcal{V}$. The soft-target loss along this path is then

$$\mathcal{L}_{\text{soft}}(x, Y; S) = \sum_{i=1}^{L_S} D_{\text{KL}} \left(q_i^{(S)} \parallel p_i^{(S)} \right).$$

For each test question i and $U \in \{T, A, E\}$, we compute the paired change

$$\Delta U_i = U_i^{\text{soft}} - U_i^{\text{hard}}$$

and report its median and the proportion of positive changes. The paired changes are analyzed separately for single-reference and multi-reference test questions. The multi-reference subset is the primary condition of interest, because the hard and soft-target objectives differ directly on training examples with multiple references. The single-reference subset provides a comparison condition for changes that may arise indirectly from training the model under a different objective.

B.1.1 Results

Figure 3 shows positive median changes in total and aleatoric uncertainty across all estimators and both datasets after the soft-target intervention. These changes are consistently larger for questions with multiple reference answers, with the strongest responses occurring on WQ. By contrast, the epistemic response is smaller and estimator-dependent. Token-level DE estimates of epistemic uncertainty remain approximately unchanged, token-level MCD estimates decrease, and sequence-level DE estimates increase.

Figure 4 illustrates the response to soft-target training for a multi-reference question. Both models generate the correct answer ‘*United States of America*’. At the first decoding position, soft-target training increases total uncertainty from 1.75 to 3.21, and aleatoric uncertainty from 1.09 to 2.83, consistent with uncertainty over which valid country to generate. Once ‘*United*’ identifies the selected answer, later tokens become nearly deterministic. Epistemic uncertainty also changes, however, illustrating that the intervention is not fully selective.

B.2 Reduced Training Data

We manipulate the amount of training data available to the model to determine whether epistemic uncertainty responds to data availability. Because NQ-Open is substantially larger than WQ, the fractions are chosen separately for each dataset to produce meaningful reductions while retaining sufficient examples for training. For WQ, the low-data, intermediate-data, and full-data conditions contain 10%, 30%, and 100% of the training set, respectively. For NQ-Open, the corresponding conditions

contain 1%, 10%, and 100%. Within each dataset, the conditions form nested subsets.

The full-data condition serves as the reference. To control for optimization budget, all conditions use the same number of optimizer steps as the corresponding full-data model. Following de Jong et al. (2026), we scale the number of epochs accordingly to match the number of optimizer steps across conditions.

For each reduced-data fraction f , test question i , and uncertainty quantity $U \in \{T, A, E\}$, we define the paired change relative to the full-data condition as

$$\Delta U_i^{(f)} = U_i^{(f)} - U_i^{(100\%)}. \quad (1)$$

For each contrast, we report the median paired change and the proportion of questions for which $\Delta U_i^{(f)} > 0$. We additionally examine whether the response becomes stronger from the intermediate-data to the low-data condition.

B.2.1 Results

Figure 5 shows that reducing the training-set size generally increases epistemic uncertainty E while decreasing aleatoric uncertainty A . The median change in E is positive for all configurations except the 30% WQ sequence-level DE, where it is approximately zero. These responses are typically stronger under the more severe data reductions. The 10% WQ models and 1% NQ-Open models tend to show larger increases in E and larger decreases in A than their corresponding intermediate-data models trained on 30% and 10% of the original training set, respectively.

Despite the increase in E , total uncertainty T usually decreases or remains approximately unchanged because the reduction in A is larger than the increase in E . The intervention therefore produces the expected increase in epistemic uncertainty, but the effect is not specific to the epistemic component.

Figure 6 shows the mixed response to reduced training data at the token-level. At the first output position, epistemic uncertainty increases while aleatoric uncertainty decreases. Because the decrease in A_i is larger than the increase in E_i , total uncertainty also decreases. Once *President* has been generated, the remaining tokens carry near-zero total uncertainty.

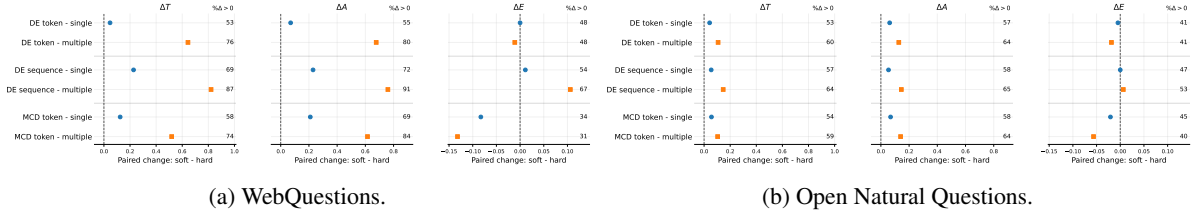


Figure 3: Paired question-level changes from hard- to soft-target training in total uncertainty ΔT , aleatoric uncertainty ΔA , and epistemic uncertainty ΔE . Rows correspond to the token-level Deep Ensemble, sequence-level Deep Ensemble, and token-level MC Dropout, with results separated between questions with one and multiple reference answers. Points denote median paired changes, percentages report the proportion of positive changes, and dashed vertical lines indicate zero change.

Q: What countries does NAFTA include?

Reference answers: Canada, United States of America, Mexico

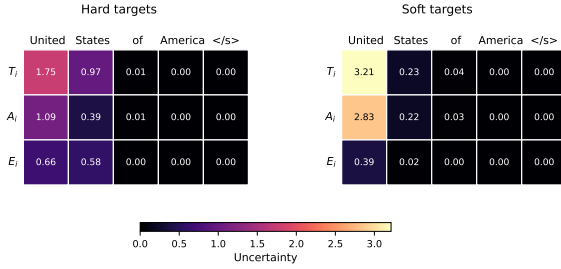


Figure 4: Token-level Deep Ensemble uncertainty for the WebQuestions question ‘What countries does NAFTA include?’ under hard- and soft-target training. Both models generate the same correct answer, ‘United States of America’, which is one of three reference answers.

C Selective Prediction

Selective prediction evaluates whether an uncertainty score can identify incorrect answers by allowing the model to abstain on its most uncertain predictions. Coverage is the fraction of questions that are answered, whereas selective risk is the error rate among those answered questions.

C.1 Methods

Suppose we have a data set $D = \{(x_i, Y_i)\}_{i=1}^N$ with $|D| = N$. Given an uncertainty score $u(x)$, we order the examples in ascending order of uncertainty. For the k most certain examples, let S_k denote the retained set and $c_k = k/N$ its coverage, where $k \in \{1, \dots, N\}$. The selective risk at coverage c_k is

$$R(c_k) = \frac{1}{k} \sum_{(x, Y) \in S_k} [1 - \text{EM}(x, Y)].$$

The risk-coverage curve plots $R(c_k)$ against c_k . We summarize selective prediction performance using

three metrics, for which lower values indicate better performance. The area under the risk-coverage curve (AURC) is given by

$$\text{AURC} = \int_0^1 R(c) dc = \frac{1}{N} \sum_{k=1}^N R(c_k).$$

AURC depends both on the ordering induced by the uncertainty score and on the underlying question answering performance. We therefore also report excess AURC (Geifman et al., 2019),

$$\text{E-AURC} = \text{AURC} - \text{AURC}^*,$$

where AURC^* is the minimum achievable AURC under an oracle ordering in which all correct predictions precede all incorrect predictions.

Finally, we report the area under the generalized risk-coverage curve (AUGRC) proposed by Traub et al. (2024). Generalized risk is the fraction of all test examples that are both retained and answered incorrectly:

$$\begin{aligned} G(c_k) &= \frac{1}{N} \sum_{(x, Y) \in S_k} [1 - \text{EM}(x, Y)] \\ &= \frac{k}{N} \cdot \frac{1}{k} \sum_{(x, Y) \in S_k} [1 - \text{EM}(x, Y)] \\ &= c_k R(c_k). \end{aligned}$$

AUGRC averages generalized risk over the attainable coverage levels:

$$\text{AUGRC} = \int_0^1 G(c) dc = \frac{1}{N} \sum_{k=1}^N G(c_k).$$

By weighting selective risk by coverage, AUGRC limits the influence of individual errors among the model’s most confident predictions (Traub et al., 2024). This is particularly useful in closed-book

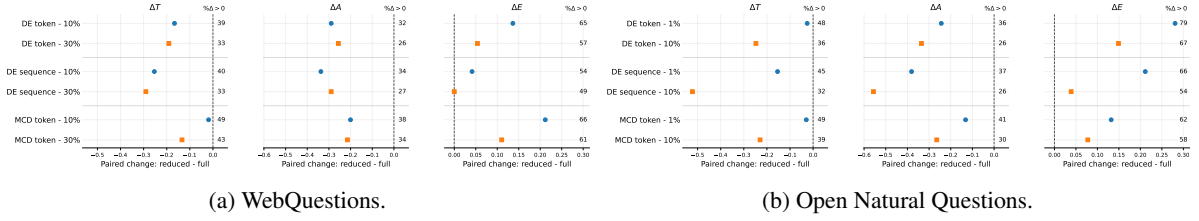


Figure 5: Paired question-level changes in total uncertainty ΔT , aleatoric uncertainty ΔA , and epistemic uncertainty ΔE for models trained on reduced data relative to the corresponding full-data model. Rows correspond to the token-level Deep Ensemble, sequence-level Deep Ensemble, and token-level MC Dropout, with results shown for the 10% and 30% WebQuestions subsets and the 1% and 10% NQ-Open subsets. Points denote median paired changes, percentages report the proportion of positive changes, and dashed vertical lines indicate zero change.

Q: What is Warren G. Harding most known for?
Reference answer: President of the United States

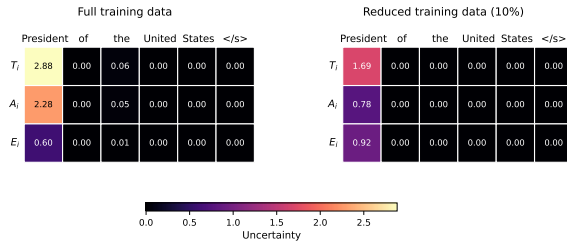


Figure 6: Token-level Deep Ensemble uncertainty for the WebQuestions question ‘What is Warren G. Harding most known for?’ after training on the full dataset and a 10% subset. Both models generate the same correct answer, ‘President of the United States’, which is the only reference answer.

question answering, where exact-match labels can be noisy because the reference set may not contain every valid answer, or there may be annotation errors.

C.2 Results

Figure 7 shows the risk-coverage curves obtained using total uncertainty as the uncertainty score. The sequence-level DE generally achieves the lowest selective risk on both datasets. The curves are closer together on WQ. Table 6 reports AURC, E-AURC, and AUGRC for total, aleatoric, and epistemic uncertainty. The three metrics show a similar overall pattern, with the sequence-level DE performing strongest on both datasets. Using total uncertainty, it obtains an AURC of 0.487, an E-AURC of 0.123, and an AUGRC of 0.290 on WQ. The corresponding values on NQ-Open are 0.473, 0.102, and 0.290.

This overall ordering partly reflects differences in QA performance. The sequence-level DE has the highest exact-match score on both datasets, at 27.9% on WQ and 27.5% on NQ-Open, whereas

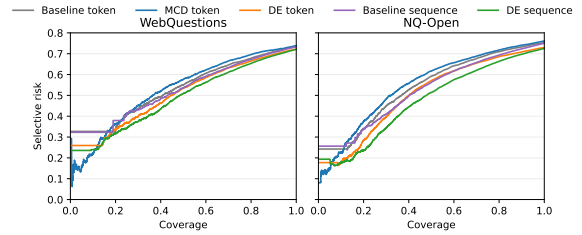


Figure 7: Risk-coverage curves on WebQuestions and Open Natural Questions, using total uncertainty T as the uncertainty score. The vertical axis is limited to $[0, 0.8]$ for readability.

MCD has the lowest, at 26.2% and 23.9%, respectively. Since selective risk at full coverage equals the overall exact-match error, these differences affect the risk-coverage curves and the resulting AURC and AUGRC values. Within each estimator, the differences between T , A , and E are generally small. On WQ, T obtains the lowest values for both DE configurations, while A and T tie for MCD. On NQ-Open, E obtains the lowest values or ties with T , while A generally obtains higher values. Overall, the results do not show a consistent selective-prediction advantage for either uncertainty component over total uncertainty.

Overall, total uncertainty provides a useful ranking of prediction errors, as rejecting answers with high T reduces selective risk across much of the coverage range. However, using A or E separately does not consistently improve over T , while differences between estimators partly reflect their underlying QA performance.

Dataset	Level	Estimator	AURC ($\downarrow$)			E-AURC ($\downarrow$)			AUGRC ($\downarrow$)		
			T	A	E	T	A	E	T	A	E
WQ	Token	Baseline	0.533	–	–	0.148	–	–	0.307	–	–
		MC Dropout	0.528	0.528	0.546	0.141	0.141	0.158	0.313	0.313	0.319
		Deep Ensemble	0.506	0.509	0.519	0.139	0.142	0.152	0.297	0.298	0.305
	Sequence	Baseline	0.527	–	–	0.146	–	–	0.303	–	–
		Deep Ensemble	0.487	0.491	0.490	0.123	0.126	0.125	0.290	0.291	0.291
	NQ-Open	Token	Baseline	0.540	–	–	0.133	–	–	0.319	–
MC Dropout			0.550	0.566	0.547	0.130	0.147	0.128	0.326	0.332	0.326
Deep Ensemble			0.505	0.514	0.502	0.131	0.140	0.127	0.305	0.309	0.304
Sequence		Baseline	0.525	–	–	0.121	–	–	0.310	–	–
		Deep Ensemble	0.473	0.478	0.473	0.102	0.107	0.102	0.290	0.292	0.290

Table 6: Selective-prediction performance on WebQuestions and Open Natural Questions measured by AURC, E-AURC, and AUGRC (all $\downarrow$), using total uncertainty T , aleatoric uncertainty A , and epistemic uncertainty E as uncertainty scores. For each stochastic estimator, the best-performing uncertainty score(s) for each evaluation metric are shown in bold and pink. Ties are highlighted jointly. For the deterministic baseline, only T is reported.