Token-Level Density-Based Uncertainty Quantification Methods for Eliciting Truthfulness of Large Language Models

Artem Vazhentsev[2,3] **Lyudmila Rvanova**[4,6] **Ivan Lazichny**[3]
Alexander Panchenko[2,3] **Maxim Panov**[1] **Timothy Baldwin**[1,5] **Artem Shelmanov**[1]
[1]MBZUAI [2]Center for Artificial Intelligence [3]Computational Semantics Group
[4]Laboratory for Analysis and Controllable Text Generation Technologies RAS
[5]The University of Melbourne [6]Weakly-Supervised NLP Group
vazhentsev@airi.net artem.shelmanov@mbzuai.ac.ae

Abstract

Uncertainty quantification (UQ) is a prominent approach for eliciting truthful answers from large language models (LLMs). To date, information-based and consistency-based UQ have been the dominant UQ methods for text generation via LLMs. Density-based methods, despite being very effective for UQ in text classification with encoder-based models, have not been very successful with generative LLMs. In this work, we adapt Mahalanobis Distance (MD) – a well-established UQ technique in classification tasks – for text generation and introduce a new supervised UQ method. Our method extracts token embeddings from multiple layers of LLMs, computes MD scores for each token, and uses linear regression trained on these features to provide robust uncertainty scores. Through extensive experiments on eleven datasets, we demonstrate that our approach substantially improves over existing UQ methods, providing accurate and computationally efficient uncertainty scores for both sequence-level selective generation and claim-level fact-checking tasks. Our method also exhibits strong generalization to out-of-domain data, making it suitable for a wide range of LLM-based applications.

1 Introduction

Large language models (LLMs) have achieved impressive results over various tasks and applications (OpenAI et al., 2024; Dubey et al., 2024; Rivière et al., 2024). Nevertheless, even the most advanced LLMs are inevitably prone to making mistakes during text generation. Their responses often contain hallucinations or non-factual claims (Xiao and Wang, 2021; Dziri et al., 2022), posing significant challenges for LLM deployment in safety-critical domains.

Many studies have investigated methods for assessing the truthfulness of LLM responses (Manakul et al., 2023; Min et al., 2023; Chen et al., 2023; Feng et al., 2024). However, many of the proposed techniques have limited practical applicability, as they often rely on external knowledge sources or require ensembling multiple large LLMs, leading to high computational costs that make them economically unfeasible for many use cases.

One of the most promising approaches to addressing this challenge is uncertainty quantification (UQ) (Gal and Ghahramani, 2016; Shelmanov et al., 2021; Baan et al., 2023; Geng et al., 2024; Fadeeva et al., 2023). This research direction recognizes that we will never have complete information about model predictions due to the limited amount of training data and ambiguity of the tasks, and suggests general ways to estimate the reliability of predictions under different conditions. Recently, a suite of UQ methods specifically designed for text generation with LLMs has been developed (Fomicheva et al., 2020; Lin et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Duan et al., 2024). However, many of these methods are either ineffective or come with a substantial computational overhead, limiting their practicality for large-scale or real-time applications.

For text classification and regression tasks, researchers have identified several groups of techniques that maintain a balance between effectiveness and computational efficiency (Zhang et al., 2019; He et al., 2020; Xin et al., 2021; Wang et al., 2022; Vazhentsev et al., 2023a; He et al., 2024a). One such class of approach is so-called *density-based uncertainty scores* (Lee et al., 2018; van Amersfoort et al., 2020; Kotelevskii et al., 2022; Yoo et al., 2022). These methods use embeddings of instances obtained from the top layers of a classification model to fit the density of the training distribution in the latent space. The likelihood of the input data under this estimated distribution is then used for confidence estimation. This has been demonstrated to achieve excellent results in out-of-distribution detection tasks (Podolskiy et al., 2021),

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2246–2262
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

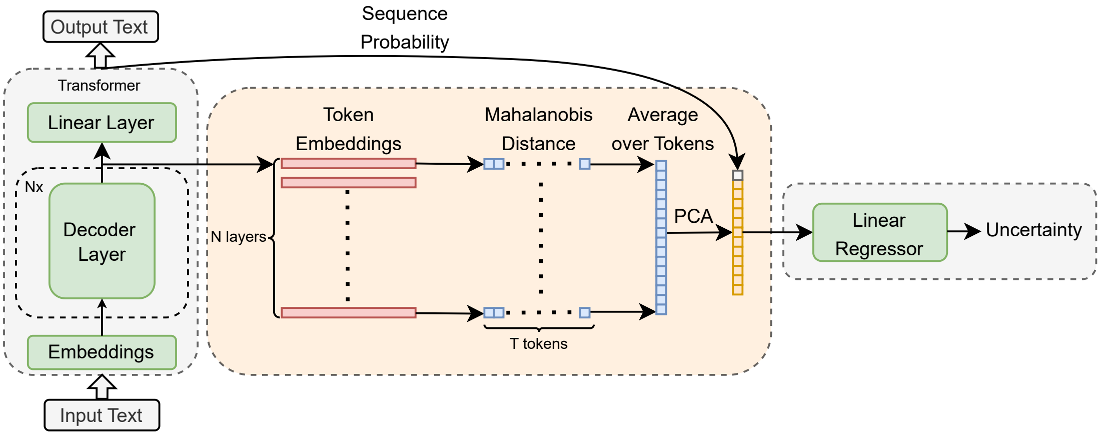

Figure 1: An illustration of the proposed method. After each decoder layer, the embeddings of each generated token are extracted. Subsequently, we compute the Mahalanobis distance for each token and layer and then average over all tokens in the generated sequence. Finally, we train a linear regression on the PCA decomposition of the calculated features with the addition of sequence probability to predict the uncertainty of the generation.

and proven to be useful for selective text classification (Vazhentsev et al., 2022, 2023a). Despite being computationally lightweight, these techniques often outperform more resource-intensive methods, such as deep ensembles (Lakshminarayanan et al., 2017) and Monte Carlo dropout (Gal and Ghahramani, 2016; Tsymbalov et al., 2018). Unfortunately, the reported performance of density-based scores for text generation so far has been notably low (Vashurin et al., 2024).

Recent work has demonstrated that the internal states of LLMs carry a lot of information about their uncertainty (Azaria and Mitchell, 2023; Chen et al., 2024; He et al., 2024b; CH-Wang et al., 2024; Vazhentsev et al., 2024). These techniques train a supplementary model on top of the activations of LLM internal layers. However, they often rely on simplistic features and fail to incorporate more advanced, well-established density-based UQ methods, limiting their ability to capture uncertainty.

In this work, we address this gap by adapting density-based techniques for the UQ of LLMs and propose a new supervised method based on density-based features. Specifically, we adapt one of the most robust methods for UQ in the classification tasks, namely Mahalanobis Distance (MD; Lee et al. (2018)), and train a linear regression on top of the MD scores from various layers of the LLM. These features are supplemented with a probability of the generated sequence. Figure 1 illustrates the scheme of the proposed supervised UQ method. Our extensive experimental evaluation demonstrates that the proposed method provides substantial improvement over the state of the art.

Our key **contributions** are as follows.

- We conduct a comprehensive investigation of density-based UQ methods for LLMs. While previous research (Vashurin et al., 2024) has indicated that sequence-level density-based methods are ineffective, we propose a token-level adaptation of MD that is on par with or better than state-of-the-art UQ techniques.
- We propose a new computationally efficient supervised method for UQ in LLMs using layer-wise density-based scores as features to improve uncertainty estimation without sacrificing the performance.
- We conduct a vast empirical investigation that demonstrates the effectiveness of the proposed methods for sequence-level selective classification across eleven datasets and claim-level fact-checking.

2 Related Work

Many effective UQ methods, such as deep ensembles (Lakshminarayanan et al., 2017) and Monte Carlo (MC) dropout (Gal and Ghahramani, 2016), require sampling multiple predictions from a model, which leads to substantial computational and memory overheads. A key challenge in UQ is developing techniques that balance effectiveness with computational efficiency. Among the most promising approaches in this regard are density-based methods (Lee et al., 2018; Liu et al., 2020; van Amersfoort et al., 2020; Kotelevskii et al., 2022; Yoo et al., 2022). These methods leverage latent representations of instances to model the training data distribution, then estimate how likely a new instance belongs to that distribution. Lee et al. (2018) propose to use Mahalanobis Distance (MD) as a measure of uncertainty for out-of-distribution detection in computer vision tasks. Podolskiy et al. (2021) adapt MD to out-of-distribution in text classification tasks. Vazhentsev et al. (2022, 2023b) show that it also provides high performance in selective text classification.

For LLMs, Fomicheva et al. (2020) and Kuhn et al. (2023) proposed UQ methods that sample multiple predictions and leverage their diversity. In the context of black-box LLMs, where we have no access to the logits or embeddings of a model, Fomicheva et al. (2020) propose the use of lexical dissimilarity of sampled texts as a measure of uncertainty. Lin et al. (2023) leverage a similarity matrix between responses for deeper analysis of the diversity of the sampled generations. Some methods also combine sampling diversity measures with the probability of each generation (Kuhn et al., 2023; Duan et al., 2024; Nikitin et al., 2024; Cheng and Vlachos, 2024; Chen et al., 2024; Vashurin et al., 2025).

Recently, it was demonstrated that MD is an efficient approach for out-of-distribution detection in sequence-to-sequence models (Vazhentsev et al., 2023b; Ren et al., 2023; Darrin et al., 2023). However, for selective generation tasks, density-based methods so far have substantially underperformed compared to trivial baselines (Vashurin et al., 2024).

Supervised methods are another research direction for UQ of LLMs. Azaria and Mitchell (2023) demonstrate that the internal states of the model contain information about uncertainty, and propose to train a multi-layer perceptron over the hidden LLM representation to predict the truthfulness of the model responses. He et al. (2024b) enhance this idea by training a deep neural network with recurrent and convolutional layers. Furthermore, this method uses embeddings from all layers and incorporates features based on the probability and the dynamics of the generated tokens through layers. In contrast to these methods, we employ a simple linear model, but focus on more accurate feature extraction from internal layers, using well-established density-based UQ methods.

3 Background on Density-Based Methods

Recently, Mahalanobis distance (MD) and Robust Density Estimation (RDE) were adapted (Vazhentsev et al., 2023b; Ren et al., 2023) to the text generation task by considering the marginal distribution of the training dataset.

Following the assumption of a Gaussian distribution of training instance representations, the MD method (Lee et al., 2018) calculates a centroid of the training data μ and the empirical covariance matrix Σ. For a given instance $\mathbf{x}$, the uncertainty score is defined as the Mahalanobis distance:

$$U^{\mathrm{MD}}(\mathbf{x}, l) = (h_l(\mathbf{x}) - \mu)^T \Sigma^{-1} (h_l(\mathbf{x}) - \mu),$$

where $h_l(\mathbf{x})$ is a hidden representation extracted from the layer l.

RDE (Yoo et al., 2022) operates within the reduced dimensionality of $h_l(\mathbf{x})$ via the kernel PCA decomposition. To ensure the robustness of the covariance matrix, it uses the Minimum Covariance Determinant estimate (Rousseeuw, 1984). Finally, the uncertainty score is computed as the Mahalanobis distance in the reduced dimensionality.

Ren et al. (2023) proposed a modification of MD – Relative Mahalanobis Distance (RMD). It takes into account a background contrastive MD score. The score aims to assess how close the test instance is to the in-domain training data compared to the background data. The uncertainty score based on RMD is given by the following equation:

$$U^{\mathrm{RMD}}(\mathbf{x}, l) = U^{\mathrm{MD}}(\mathbf{x}, l) - U_0^{\mathrm{MD}}(\mathbf{x}, l),$$

where $U_0^{\mathrm{MD}}(\mathbf{x}, l)$ is a Mahalanobis distance computed with the centroid μ^0 and the empirical covariance matrix Σ^0 calculated using the background dataset, such as C4 (Raffel et al., 2020).

For the sequence-to-sequence tasks, it was proposed to use the last encoder and decoder layer for extracting hidden representation of the model (Vazhentsev et al., 2023b; Ren et al., 2023). In contrast, recent research (Azaria and Mitchell, 2023; Chen et al., 2024) indicates that the middle layers of the model may be more suitable for decoder-only models.

4 Proposed Method: Token-Level Mahalanobis Distance

To define the method, we assume access to training data consisting of a set of prompts paired with

LLM responses, each accompanied by an assessment of its correctness. The assessment can be based on ground truth answers (as in tasks like question-answering, machine translation, or summarization) or through alternative means, such as human annotation or another big LLM.

4.1 Layer-Wise Uncertainty Score

Embedding extraction. First of all, we need to extract embeddings of instances in the training dataset. We note that previous works use sequence-level embeddings, which are essentially an average of token-level embeddings. Recent studies (Azaria and Mitchell, 2023; Chen et al., 2024) note that sequence embeddings might be useless for UQ with LLMs and propose to use embeddings of the last or the first generated token, as they encode useful information for estimating the truthfulness of the entire generation. We acknowledge that this property may not always hold, as the informative tokens are likely to vary depending on the specific task. In our method, we first compute individual token-level uncertainty scores and then aggregate them into a sequence-level score.

Embedding selection. To construct a covariance matrix and centroid for MD, a model training set is required. However, unlike standard text classification tasks, where training sets are typically limited and accessible during the development of an ML-based application, the pre-training data for general-purpose LLMs is extremely large and usually not publicly available. Moreover, even if this data were available, LLM performance on it would likely be not homogeneous and could be low for certain tasks. Therefore, to construct the parameters for MD, we propose selecting a subset of token embeddings from high-quality LLM responses.

From the responses generated in the training set, we select a subset of token embeddings that correspond to responses that meet a defined correctness criterion. Let $\mathcal{T}$ be a training set of input prompts and $|\mathcal{T}| = N_{|\mathcal{T}|}$. For each prompt $\mathbf{x}^j \in \mathcal{T}$, the model generates a response as a sequence $\tilde{\mathbf{y}}^j = \mathbf{t}_1^j, \dots, \mathbf{t}_{N_j}^j$, where N_j is a length of the j-th generation and $\mathbf{t}_i^j, i \in [1, \dots, N_j]$ is an i-th token in the response. We define a set of selected tokens as $\mathcal{D} = \{\mathbf{t}_i^j \colon \mathcal{Q}(\tilde{\mathbf{y}}^j) > \tau, i \in [1, \dots, N_j], j \in [1, \dots, N_{|\mathcal{T}|}]\}$, where $\mathcal{Q}(\cdot)$ is a quality metric and τ is a given threshold. Then $\mathcal{E}_l = \{h_l(\mathbf{t}) \colon \mathbf{t} \in \mathcal{D}\}$ is the set of selected token embeddings. The correctness criterion helps filter out low-quality responses. Depending on the dataset in the experiment, exact match and AlignScore are employed as quality metrics. The correctness criterion used for token selection is described in Section 5.1.

Layer-wise scores. For each layer $l = 1, \dots, L$ of the model, we compute the covariance matrix $\Sigma_{\mathcal{E}_l}$ and the centroid $\mu_{\mathcal{E}_l}$ using the set of selected token embeddings $\mathcal{E}_l$. For each token from the generated sequence $\tilde{\mathbf{y}}^k = \mathbf{t}_1^k, \dots, \mathbf{t}_{N_k}^k$, we compute the layer-wise MD as follows:

$$U^{\mathrm{MD}}(\mathbf{t}_i^k, l) = \big(h_l(\mathbf{t}_i^k) - \mu_{\mathcal{E}_l}\big)^T \Sigma_{\mathcal{E}_l}^{-1} \big(h_l(\mathbf{t}_i^k) - \mu_{\mathcal{E}_l}\big).$$

For the token-level RMD, we additionally compute the background covariance matrix Σ_l^0 and the background centroid μ_l^0 using the embeddings of all generated tokens for the input prompts from some background dataset.

Finally, the uncertainty score of the entire generated sequence $\tilde{\mathbf{y}}^k$ is the *Average Token-level Mahalanobis Distances (ATMD)* over $\mathbf{t}_i^k, i = 1, \dots, N_k$ (for RMD, we designate it as ATRMD).

4.2 Linear Regression on Layer-Wise Scores

The ATMD and ATRMD scores can be computed on various layers. Azaria and Mitchell (2023) indicate that the best-performing layer may vary depending on the generation task. To effectively integrate information from multiple layers, we propose training a regression model on top of the layer-wise scores.

For a generation $\tilde{\mathbf{y}}^k$, we construct a vector of features based on ATMD or ATRMD: $f^*(\tilde{\mathbf{y}}^k) = [U^*(\tilde{\mathbf{y}}^k, 1), \dots, U^*(\tilde{\mathbf{y}}^k, L)]$ (we use $*$ instead of ATMD or ATRMD). To learn the uncertainty of the generation, we define target variables as negative values of a quality metric for generations $\tilde{\mathbf{y}}^k$: $\mathbf{q}^k = -\mathcal{Q}(\tilde{\mathbf{y}}^k)$. We note that the features $f^*(\tilde{\mathbf{y}}^k)$ might be highly correlated with each other (a multicollinearity problem; Shrestha (2020)), which makes linear models to overfit (Chan et al., 2022). To make our features more robust, we use top $N = 10$ components from the PCA decomposition of feature vectors: $\tilde{f}^*(\tilde{\mathbf{y}}^k) = \mathrm{PCA}_N\big(f^*(\tilde{\mathbf{y}}^k)\big)$.

We train the machine learning model $G(\cdot)$ to predict an uncertainty score as follows:

1. Split the entire training dataset $\mathcal{T}$ into two parts $\mathcal{T}_1$ and $\mathcal{T}_2$.
2. Using $\mathcal{T}_1$, construct $\mathcal{E}_l, l \in [1, \dots, L]$ and fit layer-wise covariance matrices and centroids. ATRMD also fits layer-wise background covariance matrices and background centroids.

3. For each generation $\tilde{\mathbf{y}}^k, k = 1, \ldots, |\mathcal{T}_2|$ for the prompts from $\mathcal{T}_2$, compute features $\tilde{f}^*(\tilde{\mathbf{y}}^k)$ and targets $\mathbf{q}^k$.
4. Train the machine learning model $G^*(\cdot)$ to predict the targets $\mathbf{q}^k$ using the features $\tilde{f}^*(\tilde{\mathbf{y}}^k)$, $k = 1, \ldots, |\mathcal{T}_2|$. In our work, we use linear regression models as $G^*(\cdot)$.
5. Re-estimate layer-wise parameters of the distribution using the entire training dataset $\mathcal{T}$.

Finally, the supervised uncertainty score for a test generation $\tilde{\mathbf{y}}^k$ based on token-level MD or RMD, namely SATMD or SATRMD is:

$$U^{\mathrm{S}^*}(\tilde{\mathbf{y}}^k) = G^*\big(\tilde{f}^*(\tilde{\mathbf{y}}^k)\big).$$

Following He et al. (2024b), we also experiment with adding the sequence probability $P(\tilde{\mathbf{y}}^k \mid \mathbf{x}^k)$ as an additional feature to the features vector: $\tilde{f^*}_{prob}(\tilde{\mathbf{y}}^k) = [\tilde{f}^*(\tilde{\mathbf{y}}^k); P(\tilde{\mathbf{y}}^k \mid \mathbf{x}^k)]$, and get

$$U^{\mathrm{S}^*\text{+MSP}}(\tilde{\mathbf{y}}^k) = G^*\big(\tilde{f^*}_{prob}(\tilde{\mathbf{y}}^k)\big).$$

4.3 Hybrid Score

In addition, we explore Hybrid Uncertainty Quantification (HUQ; Vazhentsev et al. (2023a)), which empirically combines multiple uncertainty scores. Using HUQ, we combine sequence probability $U_1(\tilde{\mathbf{y}}^k) = 1 - P(\tilde{\mathbf{y}}^k \mid \mathbf{x}^k)$ and the proposed SATMD or SATRMD scores: $U_2(\tilde{\mathbf{y}}^k) = U^{\mathrm{S}^*}(\tilde{\mathbf{y}}^k)$. The hyperparameters of HUQ are tuned on the $\mathcal{T}_2$ dataset. A detailed description of the HUQ method is given in Appendix B.

5 Experiments

5.1 Experimental Setup

For the experimental evaluation, we employ the LM-Polygraph framework (Fadeeva et al., 2023; Vashurin et al., 2024). We consider two tasks: (1) sequence-level selective generation (Ren et al., 2023), in which we can "reject" untruthful generations based on provided uncertainty; (2) claim-level fact-checking (Fadeeva et al., 2024), where we aim to identify nonfactual claims in long generations, consisting of several claims.

Metrics. To evaluate the quality of UQ methods on the selective generation task, we use the standard Prediction Rejection Ratio (PPR) metric (Malinin and Gales, 2021; Vashurin et al., 2024). This metric measures the correctness of the ranking of generations based on uncertainty relative to a specified quality metric. PRR computes the area under the Prediction Rejection (PR) curve, which is constructed by sequentially rejecting the most uncertain generation and calculating the average quality for all stored generations at each possible threshold. Subsequently, this area is normalized by scaling between the PR curve for the random selection and oracle selection. A higher value of the PRR corresponds to a better quality of selective generation. Following previous work (Vashurin et al., 2024), we use ROUGE-L, Accuracy, and AlignScore (Zha et al., 2023) as text generation quality metrics.

For claim-level fact-checking, we follow previous work (Fadeeva et al., 2024) and consider this task as a binary classification problem. We utilize the ROC-AUC and PR-AUC metrics, where nonfactual claims represent a positive class.

Models. For selective generation, we experiment with two state-of-the-art LLMs in their size: Llama@8b v3.1 (Dubey et al., 2024) and Gemma 9b v2 (Rivière et al., 2024). For fact-checking, we utilize Mistral 7b v0.1 Instruct (Jiang et al., 2023). The inference hyperparameters are presented in Table 9 in Appendix F.

Datasets. We consider several text generation tasks, including text summarization (TS), question-answering (QA) with long free-form answers, QA based on reading comprehension, QA with short free-form answers, and multiple-choice QA. Dataset statistics are presented in Table 8 in Appendix E. For TS, we utilize XSum (Narayan et al., 2018), SamSum (Gliwa et al., 2019), and the CNN/-DailyMail (See et al., 2017) dataset. For QA with long free-form answers, we use PubMedQA (Jin et al., 2019), MedQUAD (Abacha and Demner-Fushman, 2019), TruthfulQA (Lin et al., 2022), and GSM8k (Cobbe et al., 2021). For reading comprehension, we use CoQA (Reddy et al., 2019) and SciQ (Welbl et al., 2017). For QA with short free-form answers, we use TriviaQA (Joshi et al., 2017). The last three datasets represent the common benchmarks for evaluating UQ methods in previous work (Kuhn et al., 2023; Duan et al., 2024; Lin et al., 2023). For multiple-choice QA, we utilize MMLU (Hendrycks et al., 2021), which is a common dataset for evaluating LLMs.

UQ baselines. In an experimental evaluation, we compare the proposed methods against several UQ baselines, including trivial yet robust information-based methods such as Maximum Sequence Probability (MSP) and Perplexity (Fomicheva et al.,

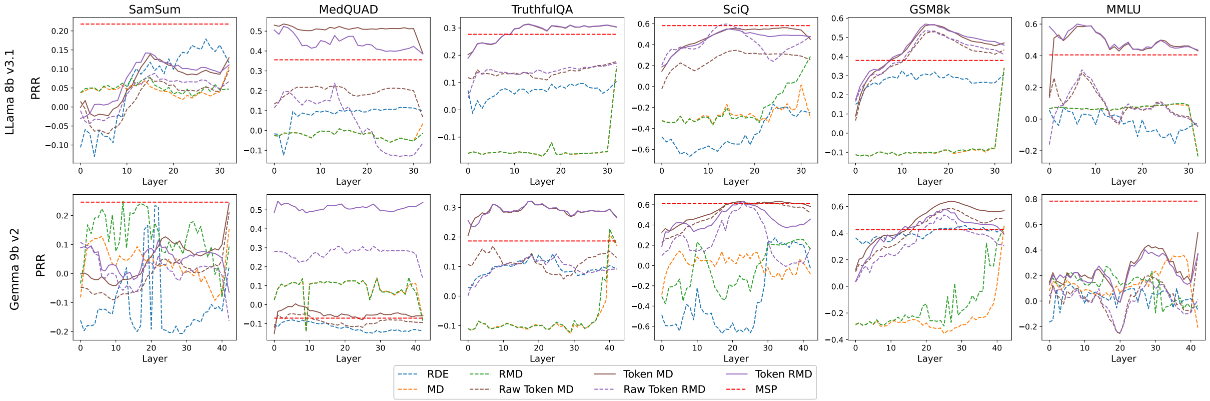

Figure 2: Performance of embeddings from various layers in density-based scores. PRR↑ for density-based methods computed using embeddings from various layers of Llama 8b v3.1 (upper row) and Gemma 9b v2 (lower row) models. Raw ATMD/ATRMD denotes a corresponding method without selecting embeddings using the correctness criterion. Higher values indicate better results.

UQ Method	XSUM		SamSum		CNN		PubMedQA		MedQUAD		TruthfulQA	CoQA	SciQ	TriviaQA	GSM8k	MMLU	Mean Rank
	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Accuracy	Accuracy	
SATMD	.189	.062	.369	.232	.059	.030	.217	.183	.471	.552	.230	.310	.385	.293	.587	.623	4.75
SATRMD	.389	.181	.338	.282	**.138**	-.004	.351	**.213**	.459	.559	.251	.324	.542	.394	.606	.504	3.69
HUQ-SATMD	-.326	-.123	.441	.249	.055	.030	.088	.128	.417	.465	.289	.450	.577	.683	.540	.552	4.56
HUQ-SATRMD	**.395**	**.187**	**.486**	.297	.126	.048	.351	.211	.386	.506	.308	.450	**.653**	.646	.592	.609	2.94
SATMD+MSP	.234	.086	.377	**.420**	.094	**.074**	.371	.203	**.493**	.527	**.361**	**.466**	.178	**.708**	.618	**.836**	**2.50**
SATRMD+MSP	.372	.179	.383	.408	.135	.016	**.372**	.202	.466	**.575**	.353	.419	.542	.702	**.642**	.816	2.56

Table 1: Performance of various versions of the proposed supervised methods. PRR↑ for Llama 8b v3.1 model for various tasks for the considered sequence-level aggregation methods. Warmer color indicates better results.

2020), and consistency-based methods considered state-of-the-art for LLMs (Vashurin et al., 2024): Lexical Similarity based on ROUGE-L (Fomicheva et al., 2020), black-box methods (DegMat, Eccentricity, EigValLaplacian; Lin et al. (2023)), Semantic Entropy (Kuhn et al., 2023), and Shifting Attention to Relevance (SAR; Duan et al. (2024)). Furthermore, to ensure the robustness of the proposed methods, the suite of baselines in our experiments also includes methods that utilize model internal states: Factoscope (He et al., 2024b), SAPLMA (Azaria and Mitchell, 2023), and EigenScore (Chen et al., 2024). The first two are supervised methods, while EigenScore is unsupervised. Following the previous works (Azaria and Mitchell, 2023; Chen et al., 2024), we use embeddings from the middle layer of the model for the latter two methods. For the methods that require sampling, we sample five generations for each input text.

Configuration of ATMD/ATRMD. In in-domain experimental evaluation on the SciQ, CoQA, TriviaQA, MMLU, and GSM8k datasets, we select token embeddings used to construct the covariance matrix and centroids for ATMD and ATRMD from generations that are fully accurate according to the exact match criterion. On other datasets, we utilize generations with AlignScore greater than 0.3. Raw ATMD/ATRMD denotes a corresponding method without selecting embeddings using the correctness criterion.

5.2 Results

Layer-wise comparison of density-based methods. Figure 2 presents the layer-wise comparison of various sequence-level density-based approaches for selective generation for the Llama 8b v3.1 and Gemma 9b v2 models. These results demonstrate the presence of robust patterns across the majority of datasets and models.

Consistent with the findings of Vashurin et al. (2024), we observe that in most cases, density-based methods that use sequence-level embeddings (MD, RMD, and RDE) yield PRR scores that are close to or below zero, indicating performance comparable to random selection. Only for GSM8k, these methods provide meaningful uncertainty scores, but they still do not outperform the basic MSP baseline. Furthermore, we see that using sequence-level embeddings derived from internal layers does not improve the performance of density-based methods; they usually perform better when using embeddings from the last layer.

MD that uses token-level embeddings performs consistently better than the MD based on sequence-level embeddings for all datasets except SamSum,

UQ Method	XSUM		SamSum		CNN		PubMedQA		MedQUAD		TruthfulQA	CoQA	SciQ	TriviaQA	GSM8k	MMLU	Mean Rank
	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Accuracy	Accuracy	
Maximum Sequence Probability	-.343	-.128	.452	.218	.021	.096	-.155	-.011	.297	.356	.277	.450	.582	.687	.380	.405	7.81
Perplexity	-.384	-.108	.080	.308	**.150**	**.242**	.215	-.044	.425	.438	.178	.450	.197	.689	.259	.308	7.38
DegMat NLI Score Entail.	.017	.093	.250	.226	.084	.144	.064	.058	.066	.162	.156	.420	.446	**.714**	.357	.224	7.19
Eccentricity NLI Score Entail.	-.036	.013	.160	.117	.028	.050	.033	.021	.070	.060	.122	.409	.444	.654	.403	.312	10.00
EigValLaplacian NLI Score Entail.	.016	.099	.251	.224	.087	.143	.054	.054	.056	.160	.152	.444	.398	.669	.335	.344	7.69
Lexical Similarity ROUGE-L	.071	.066	.320	.209	.123	.122	.144	.041	.252	.132	.008	.403	.360	.621	.467	.311	7.81
SAR	.040	.044	.300	.217	.120	.154	.122	.032	.286	.192	.105	.465	.440	.710	.455	.284	6.50
Semantic Entropy	.041	.012	.311	.206	.077	.096	.064	.034	.075	.007	.171	.416	.466	.669	.424	.220	8.38
SentenceSAR	-.085	-.032	.264	.215	.055	.091	-.000	.006	.015	.033	.185	**.472**	.543	.703	.151	.343	8.75
Factoscope	.032	-.029	.034	-.024	.007	.004	-.035	.001	.358	.428	.017	.242	.316	-.046	.048	.727	11.19
EigenScore	.041	.029	.196	.150	.040	.045	.074	.027	.050	.043	.023	.402	.373	.619	.430	.196	10.44
SAPLMA	.144	.129	.243	.313	.126	.179	.240	.155	.407	.490	.112	.082	.388	.522	.598	.481	5.50
HUQ-SATRMD	**.395**	**.187**	**.486**	.297	.126	.048	.351	**.211**	.386	.506	.308	.450	**.653**	.646	.592	.609	3.44
SATRMD+MSP	.372	.179	.383	**.408**	.135	.016	**.372**	.202	**.466**	**.575**	**.353**	.419	.542	.702	**.642**	**.816**	**2.94**

Table 2: Main results on selective generation tasks. PRR↑ for Llama 8b v3.1 model for various tasks for the considered sequence-level methods. Warmer color indicates better results.

where all methods perform similarly to each other. Moreover, density-based methods that compute MD using token-level embeddings from internal layers outperform those that rely on embeddings from the top layers. While for SamSum and MMLU with the Gemma 9b v2 model, ATMD achieves the best performance using the last layer embeddings, for all other cases the best performance is achieved by using embeddings from the middle layers. This finding is consistent with previous research (Azaria and Mitchell, 2023; Chen et al., 2024).

Using the selection of correct generations from the training dataset for fitting the covariance matrix and centroid is key to achieving good performance of the methods based on token-level MDs. ATMD/ATRMD consistently outperform raw ATMD/ATRMD. The highest difference was observed on the MedQUAD and TruthfulQA datasets, where using selection improved PRR by 0.2-0.3.

Comparison of sequence-level aggregations. Tables 1 and 5 in Appendix A.1 present the comparison of various sequence-level supervised approaches for selective generation for the Llama 8b v3.1 and Gemma 9b v2 models. The results demonstrate that SATMD and SATRMD provide stable and robust performance, which is often superior or equal to the performance of the MD/RMD using the embeddings from the best layer. As anticipated, the incorporation of MSP as an additional feature or combining it using HUQ significantly improved the performance of SATMD/SATRMD. While on average by mean rank, the best performance across various datasets was achieved by SATMD+MSP, for XSum and SciQ, HUQ-SATRMD exhibited a slight improvement. It is also noteworthy that using RMD led to a consistent improvement in performance compared to the original MD for all variants of the supervised method.

Main results on the selective generation tasks. The main results on the selective generation tasks for the Llama 8b v3.1 and Gemma 9b v2 models are presented in Tables 2 and 6 in Appendix A.2. In the summarization task, our supervised methods outperform all the baselines on XSum and SamSum. HUQ-SATRMD achieves the best performance on the XSum and SamSum datasets in terms of PRR-ROUGE-L. For SamSum, the SATRMD+MSP method significantly outperforms other methods in terms of PRR-AlignScore. For the CNN dataset, the proposed methods demonstrate the second-best results in terms of PRR-ROUGE-L, but they substantially fall behind unsupervised UQ techniques in terms of PRR-AlignScore.

In the QA tasks with long answers (PubMedQA, MedQUAD, TruthfulQA, and GSM8k), SATRMD+MSP consistently demonstrates the best performance, with a notable margin over best supervised and unsupervised techniques, while HUQ-SATRMD ranks second. In the reading comprehension task, HUQ-SATRMD is the best-performing method. Meanwhile, on the MMLU dataset, SATRMD+MSP is the most effective method, significantly outperforming HUQ-SATRMD and other state-of-the-art baselines.

Considering QA with short answers on CoQA, we observe that HUQ-SATRMD performs on par with the MSP baseline, while on the SciQ dataset performs the best with a large margin. On TriviaQA, SATRMD+MSP outperforms the MSP baseline, underperforming only sampling-based methods that require much more computation time.

Overall, we can conclude that HUQ-SATRMD is the most effective method for summarization and reading comprehension tasks, where it significantly outperforms state-of-the-art unsupervised UQ methods. For all other QA datasets, including those with long answers and tasks requiring internal knowledge, the best performance is demonstrated by SATRMD+MSP.

UQ Method	SamSum		MedQUAD		TruthfulQA	SciQ	MMLU	Mean Rank
	ROUGE-L	AlignScore	ROUGE-L	AlignScore	AlignScore	AlignScore	Accuracy	
Maximum Sequence Probability	**.452**	.218	.297	**.356**	.277	.582	.405	2.29
DegMat NLI Score Entail.	.250	.226	.066	.162	.156	.446	.224	5.00
SAR	.300	.217	.286	.192	.105	.440	.284	4.71
Semantic Entropy	.311	.206	.075	.007	.171	.466	.220	5.29
Factoscope	.061	.037	.084	.101	.045	.420	.071	7.00
SAPLMA	.241	.075	.079	.129	.009	.091	-.097	7.00
HUQ-SATRMD	.413	.229	.293	.355	**.283**	**.644**	**.770**	**1.71**
SATRMD+MSP	.207	**.314**	**.304**	.304	.122	.598	.681	3.00

Table 3: Out-of-domain generalization. PRR↑ for Llama 8b v3.1 for selective generation tasks for the considered sequence-level methods in the out-of-domain setting. Warmer color indicates better results.

UQ Method	Mistral 7b	
	ROC-AUC	PR-AUC
Maximum Claim Probability	.620	.271
P(True)	.638	.276
CCP	.716	.388
SAPLMA	.489	.166
SATRMD	.647	.275
HUQ-SATRMD	**.750**	.410
SATRMD+CCP	.739	**.414**

Table 4: Results in fact-checking. ROC-AUC↑ and PR-AUC↑ for the Mistral 7b v0.1 Instruct model for fact-checking for the considered claim-level methods. Warmer color indicates better results.

Out-of-domain generalization. Table 3 presents a comparison of various sequence-level methods for the selective generation task for the Llama 8b v3.1 model in the out-of-domain settings. We train and evaluate supervised methods using a leave-one-out approach: train on all datasets except one, which is left for testing. For each evaluation dataset, the training set is composed of 400 instances sampled from each of the remaining datasets. We use the negative AlignScore generation quality metric as a target for all considered datasets in this setting.

The performance of supervised methods in the out-of-domain setting shows a significant decline compared to the in-domain setting. Despite this, HUQ-SATRMD achieves the best results according to the mean rank, outperforming unsupervised state-of-the-art methods across the majority of datasets and metrics, except MSP on SamSum and MedQUAD in terms of PRR-ROUGE-L. Notably, when testing on the MMLU dataset, the training data consists of texts from summarization tasks and free-form QA, which differ significantly from the multiple-choice QA format used in MMLU. Nevertheless, the strong performance on MMLU demonstrates the potential of our supervised method HUQ-SATRMD for broad generalization.

Other supervised methods, including SATRMD+MSP and the baselines, show significantly poorer results in the out-of-domain setting. SATRMD+MSP underperforms MSP on several datasets, including SamSum, MedQUAD, and TruthfulQA. SAPLMA and Factoscope are not able to provide meaningful uncertainty scores, lagging significantly behind unsupervised UQ methods.

Fact-checking results. Table 4 presents a comparison of various claim-level methods for the fact-checking task using the Mistral 7b v0.1 Instruct model. The baseline supervised method SAPLMA performs similarly to a random choice. Our method SATRMD provides meaningful uncertainty scores slightly outperforming Maximum Claim Probability (MCP). We note that CCP, like MCP, is also based on the probabilities derived from the model output but demonstrates better performance than MCP. Consequently, we combine CCP with SATRMD to provide more effective claim-level fact-checking. The results demonstrate that HUQ-SATRMD achieves the best results in terms of ROC-AUC, outperforming CCP by 3.4%, while in terms of PR-AUC, SATRMD+CPP is the best, outperforming CCP by 2.6%. These results demonstrate that the proposed SATRMD methods are effective not only for sequence-level uncertainty quantification but also for estimating uncertainty on the claim level.

Impact of training data size. Figure 5 in Appendix A.3 illustrates the dependency of the performance of supervised UQ methods on the size of the training data. As expected, the optimal results on all datasets are achieved when the maximum number of training instances is used. Nevertheless, for all datasets, except SamSum and MedQUAD, the results obtained with 200-500 training instances are only slightly lower than with 2,000-5,000 instances. Furthermore, even with fewer than 200 training instances for MedQUAD, GSM8k, and MMLU, HUQ and SATRMD+MSP are able to substantially outperform the MSP method. These results demonstrate the robustness of the proposed methods with respect to the size of the training dataset.

Impact of the correctness threshold. Figure 3 presents the dependence of the performance of the SATRMD+MSP and HUQ-SATRMD methods on the correctness threshold used for the embedding selection for computing the centroid and covariance matrix for MD.

The results demonstrate that the proposed methods are generally not sensitive to the correctness threshold and consistently show high performance. However, for the MedQUAD dataset, we can see

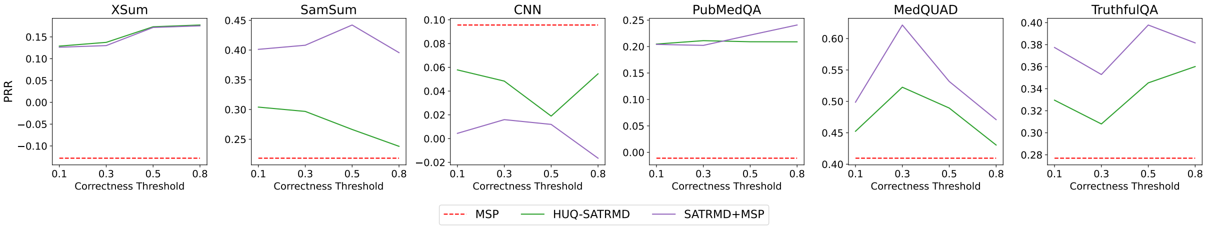

Figure 3: Dependency of PRR↑ of the SATRMD+MSP and HUQ-SATRMD methods on the correctness threshold for the embedding selection for the centroid and covariance matrix for MD for the Llama 8b v3.1 model. Higher values indicate better results.

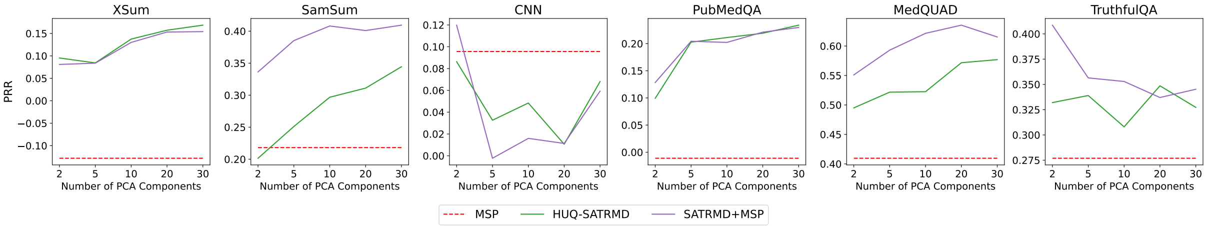

Figure 4: Dependency of PRR↑ of the SATRMD+MSP and HUQ-SATRMD methods on the number of the PCA components for the features of linear regression for the Llama 8b v3.1 model. Higher values indicate better results.

the results with a threshold of 0.3 are significantly better than those with other thresholds. Specifically, lower thresholds (e.g., 0.1) result in selecting the embeddings corresponding to incorrect instances, while higher thresholds (e.g., 0.8) exclude some embeddings associated with correct instances. Both scenarios result in suboptimal estimation of the centroid and covariance matrix, leading to a slight degradation in overall performance.

Impact of the number of PCA components. Figure 4 illustrates the impact of the number of PCA components used for the features of linear regression on the performance of SATRMD+MSP and HUQ-SATRMD methods. The best performance is achieved with 10 or 20 components for most datasets. Only for CNN and TruthfulQA, using just 2 components yields slightly better results than using more. Overall, these results indicate that our choice of 10 components is well-balanced on average across multiple datasets. We also observe that results with more than five PCA components remain stable across all datasets, showing minimal variation. Therefore, methods based on RMD are not sensitive to the precise choice of the number of PCA components.

Computational efficiency. Table 7 in Appendix C summarizes the average runtime per instance for each UQ method, along with the percentage overhead compared to standard LLM inference. State-of-the-art UQ methods that require sampling from the LLM multiple times (Semantic Entropy, SAR, Lexical Similarity) introduce a huge computational overhead (315-700%). In contrast, the proposed methods HUQ-SATRMD and SATRMD+MSP introduce minimal overhead (5.3-7.6%), which makes them much more practical.

6 Conclusion

We have introduced a series of new supervised UQ methods based on layer-wise features derived from the Mahalanobis distance. We show that calculating MD over token-level embeddings yields much better results than previous attempts that leverage sequence-level embeddings. Training a linear regression on top of the layer-wise scores allows us to produce even better uncertainty scores and outperform the state-of-the-art supervised and unsupervised UQ methods in selective classification across eleven datasets and in claim-level fact-checking. We also show that the proposed methods are computationally efficient and have the potential for generalization, which makes them useful in real-world LLM-based applications.

In future work, we aim to improve the generalization capabilities of the supervised UQ methods on out-of-domain data by investigating new features and a more robust training pipeline.

Limitations

Our approach is supervised, which means that its performance depends on the quality and size of the data available for supervision. We evaluated the robustness of the approach to dataset variation, which demonstrates that the method does not significantly degrade its quality compared to the target dataset. Nevertheless, we observe certain performance drops; thus, the resulting UQ method should be used with care beyond the supervision domain.

We did not test our method on very large LLMs, such as LLaMA 3 70b, as we were limited to using 7-9b models due to constraints in our computational resources.

Ethical Considerations

In our study, we focused on open-source LLMs and datasets that are not designed to produce harmful content. However, LLMs can still generate potentially harmful texts that may impact various groups. Uncertainty quantification techniques offer a way to enhance the reliability of neural networks and can even be used to detect harmful outputs, though this is not our focus.

Although our proposed method shows substantial performance improvements, it may sometimes incorrectly flag safe and accurate generated text as having high uncertainty. While we explicitly benchmarked the method on robustness to the task change, its applicability across various tasks remains limited.

Acknowledgments

We thank the anonymous reviewers for their valuable suggestions, which significantly contributed to improving this paper.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4401–4420, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jireh Chan, Steven Leow, Khean Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen. 2022. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, 10:1283.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023. RainProof: An umbrella to shield text generator from out-of-distribution data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5857, Singapore. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

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. 2023. LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

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, New York, New York, USA. PMLR.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Jianfeng He, Linlin Yu, Shuo Lei, Chang-Tien Lu, and Feng Chen. 2024a. Uncertainty estimation on sequential labeling via uncertainty transmission. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2823–2835, Mexico City, Mexico. Association for Computational Linguistics.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and Chang-Tien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8362–8372. Association for Computational Linguistics.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024b. LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10218–10230, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. 2022. Nonparametric uncertainty quantification for single deterministic neural network. In *Advances in Neural Information Processing Systems*.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, volume 31, pages 7167–7177.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512. Curran Associates, Inc.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany

Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, and Luke Metz et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13675–13682.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, 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, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin 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-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Peter J Rousseeuw. 1984. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Noora Shrestha. 2020. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 8:39–42.

Evgenii Tsymbalov, Maxim Panov, and Alexander Shapeev. 2018. Dropout-based active learning for regression. In *Analysis of Images, Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia, July 5–7, 2018*, pages 247–258.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9690–9700. PMLR.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *arXiv preprint arXiv:2406.15627*.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Cocoa: A generalized approach to uncertainty quantification by integrating confidence and consistency of llm outputs. *arXiv preprint arXiv:2502.04964*.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2024. Unconditional truthfulness: Learning conditional dependency for uncertainty quantification of large language models. *arXiv preprint arXiv:2408.10692*.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023a. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023b. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. In *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)*, pages 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.

A Additional Experimental Results

A.1 Comparison of Sequence-Level Aggregations

UQ Method	XSUM		SamSum		CNN		PubMedQA		MedQUAD		TruthfulQA	CoQA	SciQ	TriviaQA	GSM8k	MMLU	Mean Rank
	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Accuracy	Accuracy	
SATMD	-.071	.048	.298	.304	.032	<u>.093</u>	-.018	.076	-.357	-.027	<u>.294</u>	.371	.572	.571	<u>.682</u>	.657	4.25
SATRMD	**.444**	.239	.193	.282	<u>.034</u>	.063	.541	<u>.299</u>	.610	.479	.293	.326	.588	.450	.550	.537	3.94
HUQ-SATMD	-.251	-.055	<u>.391</u>	<u>.332</u>	**.060**	.085	-.527	-.225	-.388	-.071	.245	**.525**	<u>.626</u>	.750	.677	**.784**	3.75
HUQ-SATRMD	<u>.441</u>	<u>.253</u>	**.397**	.317	<u>.034</u>	.063	.515	.296	.558	<u>.496</u>	.258	<u>.515</u>	**.635**	.750	.526	.759	<u>3.06</u>
SATMD+MSP	-.157	.048	.316	.331	.014	**.101**	**.581**	.297	<u>.672</u>	.353	.171	.495	.547	**.794**	**.690**	.320	3.44
SATRMD+MSP	.407	**.273**	.328	**.362**	-.048	.058	<u>.576</u>	**.304**	**.711**	**.528**	**.368**	.475	.607	<u>.790</u>	.669	<u>.769</u>	**2.56**

Table 5: Performance of various versions of the proposed supervised methods. PRR↑ for Gemma 9b v2 model for various tasks for the considered sequence-level aggregation methods. Warmer color indicates better results.

A.2 Selective Generation Results

UQ Method	XSUM		SamSum		CNN		PubMedQA		MedQUAD		TruthfulQA	CoQA	SciQ	TriviaQA	GSM8k	MMLU	Mean Rank
	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	ROUGE-L	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Accuracy	Accuracy	
Maximum Sequence Probability	-.345	-.139	<u>.381</u>	.246	-.025	.053	-.526	-.214	-.389	-.071	.187	**.527**	.614	.772	.425	**.784**	8.81
Perplexity	-.361	-.191	.206	.298	.078	.010	<u>.557</u>	.200	<u>.757</u>	.321	.171	<u>.517</u>	.178	<u>.779</u>	.225	<u>.771</u>	6.81
DegMat NLI Score Entail.	.058	.090	.169	.268	.025	.104	.019	.011	-.061	.111	.201	.421	.516	.759	.520	.443	7.69
Eccentricity NLI Score Entail.	-.026	.065	.082	.157	-.033	.000	-.012	-.021	-.216	-.005	.147	.459	.526	.713	.504	.587	10.25
EigValLaplacian NLI Score Entail.	.052	.085	.174	.264	.026	.106	.021	.012	-.200	.040	.198	.448	.510	.746	.493	.572	7.88
Lexical Similarity ROUGE-L	.122	.053	.284	.267	**.085**	.107	.047	.009	.310	.066	.038	.448	.495	.731	.537	.581	7.00
SAR	.099	.054	.243	.282	.055	<u>.117</u>	.080	.003	.078	.063	.139	.472	.492	.776	.558	.702	6.38
Semantic Entropy	.099	.070	.272	.261	.078	**.128**	-.001	.011	-.136	.000	.015	.460	.507	.727	.544	.675	7.19
SentenceSAR	-.043	-.017	.186	.172	.021	.076	-.079	-.034	-.263	-.020	.151	.512	<u>.624</u>	.768	.324	.712	9.25
Factoscope	-.023	-.027	.105	.097	-.062	.042	-.044	-.000	.334	.345	-.069	.308	.548	-.040	.089	.425	11.12
EigenScore	.096	-.006	.147	.138	.040	.111	-.050	-.033	-.222	-.017	.080	.444	.494	.693	.382	.444	10.50
SAPLMA	.221	.194	.257	**.375**	<u>.079</u>	.075	.357	.221	**.765**	.249	**.460**	.069	.531	.667	<u>.667</u>	.541	5.19
HUQ-SATRMD	**.441**	<u>.253</u>	**.397**	.317	.034	.063	.515	<u>.296</u>	.558	<u>.496</u>	.258	.515	**.635**	.750	.526	.759	<u>3.62</u>
SATRMD+MSP	<u>.407</u>	**.273**	.328	<u>.362</u>	-.048	.058	**.576**	**.304**	.711	**.528**	<u>.368</u>	.475	.607	**.790**	**.669**	.769	**3.31**

Table 6: Main results on selective generation tasks. PRR↑ for Gemma 9b v2 model for various tasks for the considered sequence-level methods. Warmer color indicates better results.

A.3 Dependency on the Size of the Training Dataset

Figure 5 presents the results when varying the size of the training dataset for the supervised methods. We train the linear regression model on the training datasets of size: 100, 200, 500, 1000, 2000, and additionally on a training dataset of 5000 instances for SciQ and MMLU. Since the TruthfulQA dataset consists of only 817 instances, of which we use 409 instances as the test subset, we train linear regression on the training datasets of sizes: 100, 200, and 408.

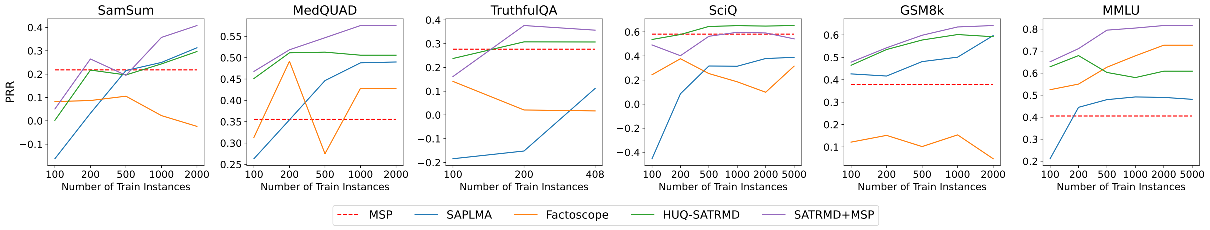

Figure 5: Dependency of PRR↑ of the supervised methods on the size of the training dataset for the Llama 8b v3.1 model. Higher values indicate better results.

B Hybrid Uncertainty Quantification

We combine the sequence probability $U_1(\tilde{\mathbf{y}}^k) = 1 - P(\tilde{\mathbf{y}}^k \mid \mathbf{x}^k)$ with the SATMD and SATRMD methods $U_2(\tilde{\mathbf{y}}^k) = U^{\mathrm{S}^*}(\tilde{\mathbf{y}}^k)$. For a given $\mathcal{T}_1$ and $\mathcal{T}_2$ from Section 4.2, and trained SATRMD method, we fit HUQ hyperparameters on the $\mathcal{T}_2$ set.

Following Vazhentsev et al. (2023a), we define the set of in-distribution instances from $\mathcal{T}_2$ as follows: $\mathcal{T}_{\text{ID}} = \{\mathbf{x} \in \mathcal{T}_2 \colon U_2(\mathbf{x}) \leq \delta_{\min}\}$. We define the set of arbitrary in-distribution instances $\mathcal{X}_{\text{ID}} = \{\mathbf{x} \colon U_2(\mathbf{x}) \leq \delta_{\min}\}$ and ambiguous in-distribution instances $\mathcal{X}_{\text{IDA}} = \{\mathbf{x} \in \mathcal{X}_{\text{ID}} \colon U_1(\mathbf{x}) > \delta_{\max}\}$ using $\delta_{\min}, \delta_{\max}$ are thresholds selected on the $\mathcal{T}_2$ dataset.

To make different uncertainty scores comparable, we define a ranking function $R(\mathbf{u}, \mathfrak{D})$ as a rank of $\mathbf{u}$ over a sorted dataset $\mathfrak{D}$, where $\mathbf{u}_1 > \mathbf{u}_2$ implies $R(\mathbf{u}_1, \mathfrak{D}) > R(\mathbf{u}_2, \mathfrak{D})$. We compute the total uncertainty $U_{\mathrm{T}}(\mathbf{x})$ as a linear combination $U_{\mathrm{T}}(\mathbf{x}) = (1-\alpha)R(U_2(\mathbf{x}), \mathcal{T}_2) + \alpha R(U_1(\mathbf{x}), \mathcal{T}_2)$, where α is a hyperparameter selected on the $\mathcal{T}_2$ dataset. As a result, we define HUQ as follows:

$$U_{\mathrm{HUQ}}(\mathbf{x}) = \begin{cases} R(U_1(\mathbf{x}), \mathcal{T}_{\mathrm{ID}}), \forall \mathbf{x} \in \mathcal{X}_{\mathrm{ID}} \setminus \mathcal{X}_{\mathrm{AID}}, \\ R(U_1(\mathbf{x}), \mathcal{T}_2), \forall \mathbf{x} \in \mathcal{X}_{\mathrm{AID}}, \\ U_{\mathrm{T}}(\mathbf{x}), \forall \mathbf{x} \notin \mathcal{X}_{\mathrm{ID}}. \end{cases}$$

C Computational Efficiency

UQ Method	Runtime per batch	Overhead
MSP	2.10 ± 1.31	-
DegMat NLI Score Entail.	9.47 ± 3.41	350%
Lexical Similarity ROUGE-L	8.69 ± 3.31	315%
Semantic Entropy	9.47 ± 3.41	350%
SAR	16.89 ± 6.85	700%
SAPLMA	2.10 ± 1.31	**0.04**%
Factoscope	8.45 ± 5.92	300%
HUQ-SATRMD	2.21 ± 1.36	5.30%
SATRMD+MSP	2.26 ± 1.38	7.61 %

Table 7: The evaluation of the inference runtime of UQ methods measured on all test instances from all datasets with predictions from Llama 8b v3.1. The best results are in bold. The second best results are underlined.

D Computational Resources

All experiments were conducted on a cluster with 6 NVIDIA H100 GPUs. The total time for all conducted experiments for all models across all datasets is approximately 400 GPU hours.

E Dataset Statistics

Task	Dataset	N-shot	Train texts for STMD	Evaluation texts
Text Summarization	CNN/DailyMail	0	2,000	2,000
	XSum	0	2,000	2,000
	SamSum	0	2,000	819
QA Long answer	PubMedQA	0	2,000	2,000
	MedQUAD	5	2,000	2,000
	TruthfulQA	5	408	409
	GSM8k	5	2,000	1,319
QA Short answer	SciQ	0	5,000	1,000
	CoQA	all preceding questions	5,000	2,000
	TriviaQA	5	5,000	2,000
MCQA	MMLU	5	5,000	2,000

Table 8: The statistics of the datasets used for evaluation.

F Inference Hyperparameters

Dataset	Task	Max Input Length	Generation Length	Temperature	Top-p	Do Sample	Beams	Repetition Penalty
XSum	TS	-	56	1.0	1.0	False	1	1
SamSum			128					
CNN			128					
PubMedQA	QA Long answer		128					
MedQUAD			128					
TruthfulQA			128					
GSM8k			256					
CoQA	QA Short answer		20					
SciQ			20					
TriviQA			20					
MMLU	MCQA		3					

Table 9: Text generation hyperparameters for all LLMs used in the experiments.