Unconditional Truthfulness: Learning Conditional Dependency for Uncertainty Quantification of Large Language Models

Anonymous ACL submission

Abstract

Uncertainty quantification (UQ) has emerged as a promising approach for detecting hallucinations and low-quality output of Large Language Models (LLMs). However, obtaining proper uncertainty scores is complicated by the conditional dependency between the generation steps of an autoregressive LLM, because it is hard to model it explicitly. Here, we propose to learn this dependency from attention-based features. In particular, we train a regression model that leverages LLM attention maps, probabilities on the current generation step, and recurrently computed uncertainty scores from previously generated tokens. To mitigate the overfitting due to "teacher forcing" in the recurrent features, we also suggest a two-staged training procedure. Our experimental evaluation on ten datasets and three LLMs shows that the proposed method is highly effective for selective generation, achieving substantial improvements over rivaling unsupervised and supervised approaches.

1 Introduction

Uncertainty quantification (UQ) (Gal and Ghahramani, 2016; Baan et al., 2023; Geng et al., 2024; Fadeeva et al., 2023) is of growing interest in the Natural Language Processing (NLP) community for dealing with Large Language Models (LLMs) hallucinations (Fadeeva et al., 2024) and low-quality generations (Malinin and Gales, 2021) in an efficient manner. For example, high uncertainty could serve as an indicator that the LLM generation should be discarded as potentially harmful or misleading. This approach is known in the literature as selective generation (Baan et al., 2023).

There are many approaches for detecting hallucinations and low-quality outputs of LLMs (Manakul et al., 2023; Min et al., 2023; Chen et al., 2023). However, many of them leverage external knowledge sources or a second LLM. Knowledge sources are generally patchy in coverage, while censoring the outputs of a small LLM using a bigger one has a high computational cost and is impractical. We argue that LLMs inherently contain information about the limitations of their own knowledge, and that there should be an efficient way to access this information, which can enable LLM-based applications that are both safe and practical.

While for general classification and regression tasks there is a well-developed battery of UQ techniques (Zhang et al., 2019; He et al., 2020; Xin et al., 2021; Wang et al., 2022; Vazhentsev et al., 2023; He et al., 2024a), for text generation tasks, UQ is much more complicated. The complexity is multifold: (1) there is an infinite number of possible generations, which complicates the normalization of the uncertainty scores, (2) in the general case, there are an infinite number of correct answers, (3) decisions are generally based on imprecise sampling and inference algorithms such as beam search, (4) there is not one, but multiple tokens, and the uncertainty of these predictions needs to be aggregated, and finally, (5) the predictions at each generation step are not conditionally independent (Zhang et al., 2023).

This last problem is the focus of the present work. During generation, LLMs condition on the previously generated tokens. Thus, if an LLM has hallucinated and generated an incorrect claim at the beginning or in the middle of the sequence, all subsequently generated claims might also be incorrect. Even if the first claim was generated with high uncertainty, this is not taken into account during the subsequent generation process. This means that while the first error could be recognized as having high uncertainty, all subsequent errors will be overlooked because the generation process conditioned on this error will be very confident.

We note that the attention between the generated tokens provides information about the conditional dependency between the generation steps. Previously, there have been several attempts to

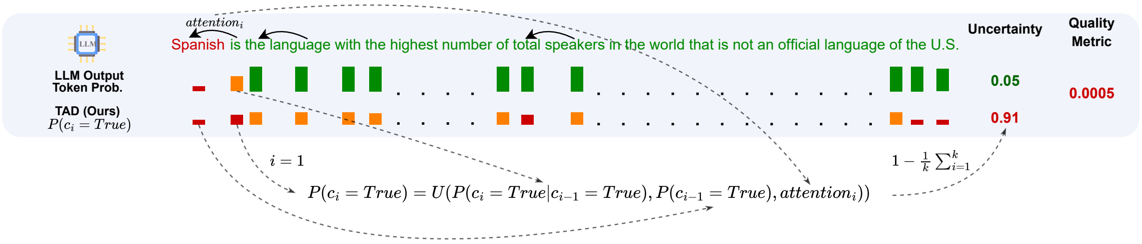
attention$_i$
Spanish is the language with the highest number of total speakers in the world that is not an official language of the U.S.
Uncertainty
Quality Metric
LLM
LLM Output Token Prob.
TAD (Ours) $P(c_i = True)$
0.05
0.0005
0.91
$i = 1$
$1 - \frac{1}{k}\sum_{i=1}^{k}$
$P(c_i = True) = U(P(c_i = True|c_{i-1} = True), P(c_{i-1} = True), attention_i))$

Figure 1: An illustration of the proposed method TAD. The figure shows the generated tokens, the uncertainty scores for the generated sequence, and the probabilities assigned by an LLM and by TAD (represented with bars). The output is generated by Gemma 7b for the question *What is the language with the highest number of total speakers in the world that is not an official language of the U.S.?* The LLM starts by generating a token *Spanish* that leads to the erroneous answer. The probabilities estimated by the LLM are high for all tokens except for the first one, which makes the uncertainty scores based on raw probabilities misleadingly low. On the contrary, TAD takes into account uncertainty from the previous step using a trainable model $U(\cdot)$ based on attention, resulting in a high overall uncertainty for the generated answer.

suggest heuristic approaches to model this dependency (Zhang et al., 2023). We argue that the particular algorithmic function would be too difficult to engineer, and thus we propose to learn this dependency from data instead.

For this purpose, we generate a training dataset with a target variable, representing the quality score of the generated text according to some ground truth annotation, and train a regression model that leverages LLM attention maps, probabilities on the current generation step, and recurrently computed uncertainty scores from previously generated tokens. To mitigate the overfitting due to "teacher forcing" in the recurrent features, we suggest a two-staged training procedure where ground truth annotations for previously calculated scores are replaced by outputs from the intermediate model obtained in the first training step. We call the proposed approach *trainable attention-based dependency* (*TAD*). Figure 1 illustrates the idea of the method on the real output of an LLM.

The contributions of this work are as follows:

- We develop a new data-driven supervised approach to uncertainty quantification that leverages features based on attention maps, probabilities on the current generation step, and recurrently computed uncertainty scores from previously generated tokens.
- We show that both attention and recurrent features are essential for achieving high performance in UQ, and two step training procedure is necessary to avoid overfitting.
- We conduct vast empirical investigation in selective generation and show that the proposed approach outperforms previous unsupervised and supervised UQ methods across nine datasets and three LLMs.

2 Related Work

With the advent of LLMs, UQ has become a very timely research problem in NLP. As previously mentioned, this area not only offers promising practical benefits, but it also presents several intriguing research challenges. The majority of the methods for UQ of LLM generations has been unsupervised, with few recently-proposed supervised methods.

Unsupervised UQ methods. Several methods adapt information-based UQ techniques by aggregating the logits of the generated tokens in various ways. Fomicheva et al. (2020) experimented with perplexity and mean token entropy for MT quality estimation. Takayama and Arase (2019) adapted point-wise mutual information (PMI), and van der Poel et al. (2022) extended this approach to conditional PMI. The advantages of these techniques are their simplicity, usually minimal computational overhead, and robust performance. A well-known approach to UQ in general is ensembling (Lakshminarayanan et al., 2017) and Monte Carlo (MC) dropout (Gal and Ghahramani, 2016). Malinin and Gales (2021) and Fomicheva et al. (2020) adapted it to sequence generation problems. In this category, lexical similarity (Fomicheva et al., 2020) is a very competitive baseline that can be applied to black-box models (without any access to logits or internal model representations).

The problem of multiple correct generations was explicitly addressed in (Kuhn et al., 2023; Nikitin et al., 2024; Cheng and Vlachos, 2024; Zhang et al., 2024) and in a series of black-box generation methods (Lin et al., 2023). The main idea is to sample multiple generations from an LLM, extract semantically equivalent clusters, and analyze the diversity of the generated meanings instead of the surface forms. Chen et al. (2024) proposed evaluating the consistency of the multiple generations in the embedding space using their hidden states.

Fadeeva et al. (2024) addressed the problem of multiple sources of uncertainty present in the LLM's probability distribution that are irrelevant for hallucination detection. In addition to dealing with multiple correct generations, they also suggested mitigating the influence of the uncertainty related to the type of generated claims.

Zhang et al. (2023) and Duan et al. (2024) highlighted that not all tokens should contribute to the uncertainty score, proposing heuristics to select the relevant tokens. Zhang et al. (2023) also modeled the conditional dependencies between the generation steps by penalizing the uncertainty scores based on the uncertainties of the previously generated tokens and the max-pooled attention to the previous tokens.

Overall, most previous work on UQ has not addressed the conditional dependency between the predictions, or has addressed it using heuristics. We argue that the conditional dependency is an important aspect of UQ for text generation tasks and we propose a data-driven approach to it. We also note that techniques based on sampling multiple answers from LLMs usually introduce prohibitive computational overhead. We argue that for UQ methods to be practical, they should also be computationally efficient.

Supervised UQ methods. Supervised regression-based confidence estimators are well-known for classification problems, primarily from computer vision (Lahlou et al., 2022; Park and Blei, 2024). Their key benefit is computational efficiency.

A handful of papers applied them to text generation tasks. Lu et al. (2022) proposed to train a regression head of a model to predict confidence. They noted that the probability distribution of a language model is poorly calibrated and cannot be used directly to spot low-quality translations. They modified the model architecture and the loss function, restricting this approach to fine-tuning language models only for Machine Translation (MT) and making it unsuitable for general-purpose LLMs. In a similar vein, Azaria and Mitchell (2023) approached the task of UQ by training a multi-layer perceptron (MLP) on the activations of the internal layers of LLMs to classify true vs. false statements. They demonstrated that it outperformed other supervised baselines and few-shot prompting of the LLM itself. However, the reliance on forced decoding limits the real-world applicability for hallucination detection in unrestricted generation cases.

Several studies enhanced this method by refining the model architecture and the training procedure. Su et al. (2024) combined the hidden state of the last token with the average hidden state of the sequence, while CH-Wang et al. (2024) introduced a trainable attention layer over token embeddings and used linear regression on top of the MLP's predictions based on embeddings from various layers. He et al. (2024b) proposed to combine multiple deep learning models trained on diverse features extracted from hidden states. Chuang et al. (2024) suggested training the linear classifier using features derived from attention matrices. A key limitation of these methods is that they can only provide veracity scores for the entire generated text.

Unlike previous methods, we focus on modeling conditional dependency between generation steps using attention and leverage recurrently computed uncertainty scores from previous generation steps. Our method is also flexible as it can be applied at different levels: to the entire text, to a sub-sequence, or to individual tokens.

3 Trainable Attention-Based Conditional Dependency

In this section, we present our approach to learning the conditional dependency between the generation steps and our UQ method based on it.

3.1 Motivation

When an LLM generates a sequence of tokens t_i, it provides us a conditional probability distribution $p(t_i \mid t_{<i})$. This essentially means that the LLM considers that everything generated so far is correct, which might not be the case. In practice, we would like to somehow propagate the uncertainty from the previous generation steps.

For the sake of simplicity, let us assume that only the uncertainty from the previous tokens is propagated to the current generation step. This assumption can be expressed as follows: $p(t_i \mid t_{<i}) \simeq p(t_i \mid t_{i-1})$. Let us further consider that we have trained an LLM that generates only tokens

that are true ('T') or false ('F'). The probability of the token t_i being 'T' is given by the conditional probability $p(t_i \mid t_{i-1}) = p(t_i = \mathrm{T} \mid t_{i-1} = \mathrm{T})$. Assume we already have some tokens $t_1, t_2, \ldots, t_n$ and a prompt x. At each step, the LLM provides us $p(t_1 = \mathrm{T} \mid x), p(t_2 = \mathrm{T} \mid t_1 = \mathrm{T}), \ldots, p(t_n = \mathrm{T} \mid t_{n-1} = \mathrm{T})$.

These probability distributions are conditionally dependent on the previously generated tokens. However, to estimate the correctness of some token t_i, we need to obtain an *unconditional probability* $p(t_i) = p(t_i = \mathrm{T})$. Let us expand $p(t_i = \mathrm{T})$ according to the law of total probability and express it using conditional probability:

$$p(t_i = \mathrm{T}) = p(t_i = \mathrm{T} \mid t_{i-1} = \mathrm{T}) \cdot p(t_{i-1} = \mathrm{T}) \\ + p(t_i = \mathrm{T} \mid t_{i-1} = \mathrm{F}) \cdot \big(1 - p(t_{i-1} = \mathrm{T})\big).$$

In this formula, $p(t_i = \mathrm{T} \mid t_{i-1} = \mathrm{T})$ is what the LLM provides during the current generation step in accordance with the specified assumptions, and $p(t_{i-1} = \mathrm{T})$ is recurrently calculated based on the previous generation step. We still do not know the remaining term: $p(t_i = \mathrm{T} \mid t_{i-1} = \mathrm{F})$. This simplistic example shows that in order to obtain a reliable uncertainty estimate, we cannot rely solely on the probability distribution provided by the LLM and we also need to model the conditional dependency of the generation steps. It also makes it explicit that we need some recurrence in the computation of the token-level uncertainty scores.

Attention weights commonly reflect the degree of conditional dependency between the generation steps. However, obtaining a direct expression that would accurately approximate the conditional dependency between the generation steps is challenging. The assumptions in our simplistic example do not hold in real LLMs, and thus the predictions on each step depend on multiple previous tokens in a complicated fashion. We suggest learning this dependency in a supervised way from attention. In particular, we propose a feature set for training token-level uncertainty scores U consisting of the attention weights Att_i, the token probabilities from the LLM on the current step $p(t_i \mid t_{<i})$, and the recurrently calculated uncertainty scores on the previous steps $U_{<i}$:

$$U(t_i \mid t_{<i}) = U\big(Att_i, p(t_i \mid t_{<i}), U_{<i}\big). \qquad (1)$$

3.2 Implementation

We implement the proposed method for token-level UQ and aggregate token-level scores into a score for the entire sequence.

Obtaining unconditional probability. In order to obtain the surrogate $\hat{p}(t_i)$ for the unconditional probability $p(t_i)$ for a generated token $t_i \in \tilde{y}$ during the training phase, we compute the similarity between the generated answer $\tilde{y}$ and the ground truth y:

$$\tilde{U} = 1 - \mathrm{sim}(\tilde{y}, y). \qquad (2)$$

For generating the targets, we use task-specific similarity measures, such as Accuracy, COMET (Rei et al., 2020), and AlignScore (Zha et al., 2023).

Generating training data for TAD. We generate the training data for TAD using the original textual training dataset in the following way:

1. For the input prompt x_k, we use an LLM to generate a text $\tilde{y}_k = t_1 t_2 \ldots t_{n_k}$ of some length n_k and token probabilities $p(t_i \mid x_k, t_{<i})$.
2. For the first generated token t_1 in each text, we define its unconditional probability estimate as a target variable $\hat{p}_k(t_1) = 1 - \tilde{U}^k = \mathrm{sim}(\tilde{y}_k, y_k)$ according to equation (2).
3. For each generated token t_i, $i = 2, \ldots, n_k$:

   (a) For the predefined number of preceding tokens N, we construct a feature vector z_i^k. The feature vector z_i^k includes: the conditional probabilities $p(t_i \mid x_k, t_{<i})$ and, for $l = 1, \ldots, \min\{N, i-1\}$, $p(t_{i-l} \mid x_k, t_{<i-l})$, the unconditional probabilities' estimates from the previous steps $\hat{p}_k(t_{i-l})$, and the attention weights $a_{i,i-l}$ from the $(i-l)$-th token to the i-th token from all layers and heads. If $N > i-1$, we pad the feature vector with zeros to ensure they have the same length.

   (b) We compute the unconditional probability estimates $\hat{p}(t_i)$. On the first iteration of learning it is done according to equation (2): $\hat{p}_k(t_i) = \mathrm{sim}(\tilde{y}_k, y_k)$. On the next iterations of learning it is done via the learned function $\hat{p}_k(t_i) = 1 - U(z_i^k)$.

As a result, for each instance in the training dataset and for each iteration of learning, we generate a sequence of target variables $\tilde{U}_i^k = 1 - \mathrm{sim}(\tilde{y}_k, y_k)$ and corresponding feature vectors $z_i^k, k = 1, \ldots, K, i = 2, \ldots, n_k$. We use this dataset to train the model U.

Model for U and its training procedure. The training procedure involves using the ground truth estimates of the unconditional probabilities from the previous steps as features. This introduces a form of *teacher forcing*, which can lead to overfitting. To address this issue, we perform the training procedure twice. In the second time, instead of ground-truth estimates, we leverage the predictions of the function U trained on the first iteration. This two-step training approach enables us to leverage the conditional dependency of the current step on the previous ones when computing the uncertainty score. Our experiments show that it is essential for achieving good performance.

We experiment with two regression models for TAD: linear regression (LinReg) and a multi-layer perceptron (MLP). The hyper-parameters of the regressors are obtained using cross-validation with five folds on the training dataset. We select the optimal values of the hyperparameters based on the best average PRR. The optimal values are used to train the regression model on the full training set. The selected hyper-parameters values for the TAD modules are presented in Appendix C.1.

Inference procedure. During inference, we obtain predictions from the LLM as always, but we also extract features from the attention outputs. For the first generated token t_1, its unconditional probability is defined as $p(t_1) = p(t_1 \mid x_k)$. For each subsequent token, the function U computes the predictions recursively, leveraging the attentions, the conditional probabilities, and the unconditional probabilities predicted for the preceding tokens. Moreover, the token-level scores are aggregated into a score for the entire generation.

4 Experiments and Evaluation

4.1 Experimental Setup

For the experimental evaluation, we use the LM-Polygraph framework (Fadeeva et al., 2023). We focus on the task of selective generation (Ren et al., 2023) where we "reject" generated sequences due to low quality based on uncertainty scores. Rejecting means that we do not use the model output, and the corresponding queries are processed differently, e.g., they could be further reprocessed manually.

Evaluation measures. Following previous work on UQ in text generation (Malinin and Gales, 2021; Vashurin et al., 2024), we compare UQ methods using the Prediction Rejection Ratio (PRR) metric. PRR quantifies how well an uncertainty score can identify and reject low-quality predictions according to some quality measure. The PRR scores are normalized to the range $[0, 1]$ by linearly scaling the area under the PR curve between the values obtained with random selection (corresponding to 0) and oracle selection (corresponding to 1). Higher PRR values indicate better quality of the selective generation. We use Accuracy, COMET (Rei et al., 2020), and AlignScore (Zha et al., 2023) as generation quality measures.

Datasets. We consider five text generation tasks: text summarization (TS), machine translation (MT), Question Answering (QA) with long free-form answers, QA with free-form short answers, and multiple-choice QA. Statistics about the datasets are provided in Table 15 in Appendix D. For TS, we experiment with CNN/DailyMail (See et al., 2017) and SamSum (Gliwa et al., 2019). For the long answer QA task, we use MedQUAD (Abacha and Demner-Fushman, 2019), which consists of real medical questions, TruthfulQA (Lin et al., 2022), which consists of questions that some people would answer incorrectly due to a false belief or a misconception, and GSM8k (Cobbe et al., 2021) with a grade school math questions. For the QA task with short answers, we follow previous work on UQ (Kuhn et al., 2023; Duan et al., 2024; Lin et al., 2023) and we use three datasets: SciQ (Welbl et al., 2017), CoQA (Reddy et al., 2019), and TriviaQA (Joshi et al., 2017). For multiple-choice QA, we use MMLU (Hendrycks et al., 2021), a widely used benchmark for evaluating LLMs. For MT, we use WMT19 (Barrault et al., 2019), focusing on translations from German to English.

LLMs. We experiment with three LLMs: LLaMA-3.1 8b (Dubey et al., 2024), Gemma-2 9b (Rivière et al., 2024), and Qwen-2.5 7b (Yang et al., 2024). The values of the inference hyper-parameters are given in Table 14 in Appendix C.2.

UQ baselines. The set of baselines includes Maximum Sequence Probability (MSP) and Perplexity (Fomicheva et al., 2020), which are considered simple yet strong and robust baselines for selective generation across various tasks (Fadeeva et al., 2023). We also compare our method to unsupervised techniques considered to be state-of-the-art: 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), hallucination detection with a stronger focus (Focus; Zhang

UQ Method	SamSum	CNN	WMT19	MedQUAD	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k	Mean	Mean
	AlignScore	AlignScore	Comet	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR	Rank
MSP	.298	.157	.569	.356	.277	.450	.582	.687	.444	.380	.420	6.90
Perplexity	.029	-.116	.460	.438	.178	.450	.202	.689	.374	.259	.296	12.30
CCP	.287	.101	.453	.321	.176	.385	.364	.712	.261	.408	.347	10.50
Simple Focus	.230	.101	.553	.381	.262	.475	.540	.703	.413	.381	.404	7.20
Focus	.144	-.002	.501	.460	.213	.345	.456	.621	.155	.402	.330	12.40
Lexical Similarity Rouge-L	.073	.074	.455	.153	.029	.428	.555	.613	.313	.452	.315	12.30
EigenScore	-.002	.094	.468	.047	.033	.412	.541	.591	.154	.385	.272	14.40
EVL NLI Score entail.	.111	.056	.366	.133	.134	.458	.527	.684	.304	.359	.313	12.70
Ecc. NLI Score entail.	.020	.003	.406	.099	.127	.434	.541	.632	.322	.399	.298	13.30
DegMat NLI Score entail.	.112	.062	.388	.138	.134	.453	.542	.703	.279	.385	.320	11.30
Semantic Entropy	.089	.056	.524	.027	.051	.423	.527	.660	.223	.465	.305	13.20
SAR	.121	.081	.508	.219	.078	.458	.545	.697	.299	.471	.348	9.20
LUQ	.153	.058	.258	.107	.099	.428	.499	.692	.267	.289	.285	13.50
Factoscope	.067	.086	.218	.236	.164	.049	.386	.460	.703	.108	.248	14.30
SAPLMA	.284	.073	.574	.429	.146	.039	.425	.535	.492	.508	.350	10.30
MIND	.217	.162	.494	.583	.385	.381	.589	.632	**.813**	.607	.486	6.50
Sheeps	.292	.179	.554	.552	.464	.500	.487	.709	.796	.659	.519	4.20
LookBackLens	**.459**	**.233**	**.615**	.579	.386	.441	.594	.631	.774	.619	.533	4.10
TAD	.431	.215	.612	**.662**	**.565**	**.509**	**.644**	**.737**	.806	**.682**	**.586**	**1.40**

Table 1: PRR↑ of UQ methods for the Llama-3.1 8b model. Warmer color indicates better results. The best method is in **bold**, the second best is underlined.

et al. (2023)), claim-conditioned probability (CCP; Fadeeva et al. (2024)), Shifting Attention to Relevance (SAR; Duan et al. (2024)), EigenScore (Chen et al., 2024), and long-text uncertainty quantification (LUQ; Zhang et al. (2024)). For sampling-based methods, we generate five samples.

The suite of baselines also includes state-of-the-art supervised methods that use hidden states or attention weights: Factoscope (He et al., 2024b), SAPLMA (Azaria and Mitchell, 2023), MIND (Su et al., 2024), Sheeps (CH-Wang et al., 2024), and LookBackLens (Chuang et al., 2024).

4.2 Main Results

Fine-grained comparison to the baselines. Tables 1, 5 and 6 in Appendix A.1 present the results for LLaMa-3.1 8b, Gemma-2 9b, and Qwen-2.5 7b, respectively.

The results demonstrate that, across all summarization and translation datasets, both LookBackLens and TAD outperform state-of-the-art methods by a substantial margin. For Llama, LookBackLens achieves slightly better results than TAD, but TAD outperforms LookBackLens on the CNN dataset when using Gemma and on the WMT19 dataset with Qwen.

For QA involving long answers (e.g., MedQUAD, TruthfulQA, and GSM8k), TAD demonstrates substantial improvements over the baselines across all considered models. For example, in the experiment with LLaMA-3.1 8b on TruthfulQA, TAD outperforms the second-best baseline, Sheeps, by 0.101 of PRR. On the MedQUAD dataset, TAD achieves an improvement of 0.079 in PRR over the second-best baseline, and on GSM8k, it improves PRR by 0.023.

For QA with short answers (CoQA, SciQ, and TriviaQA), TAD generally exhibits notable improvements over the baseline methods in the majority of cases. The only exception is the case of the SciQ dataset, where LookBackLens is marginally better for Gemma-2 9b and Qwen-2.5 7b. On TriviaQA, when using the Gemma-2 9b model, TAD performs on par with sampling-based methods, while other supervised methods fall behind simple baselines by a margin.

Finally, for the MMLU dataset, TAD also notably outperforms state-of-the-art methods for both Gemma-2 9b and Qwen-2.5 7b. However, for LLaMA-3.1 8b, TAD slightly underperforms compared to MIND.

Summarizing, our findings indicate that certain UQ methods, such as LookBackLens and Sheeps, can achieve top performance in specific experimental settings. However, TAD demonstrates the most consistent and robust performance across all eleven tasks, never ranking below the second-best method. In contrast, other supervised methods occasionally underperform, sometimes even falling below simple baseline such as MSP.

Overall results. Table 2 presents the mean rank of each method aggregated over all datasets for each model separately. The lower rank is better. The column *Mean Rank* shows the mean rank of the ranks across all models. Figure 2 additionally summarizes all experimental setups. Each cell presents a win rate for a method from a row compared to a method from a column. The aggregated results emphasize the significance of the performance im-

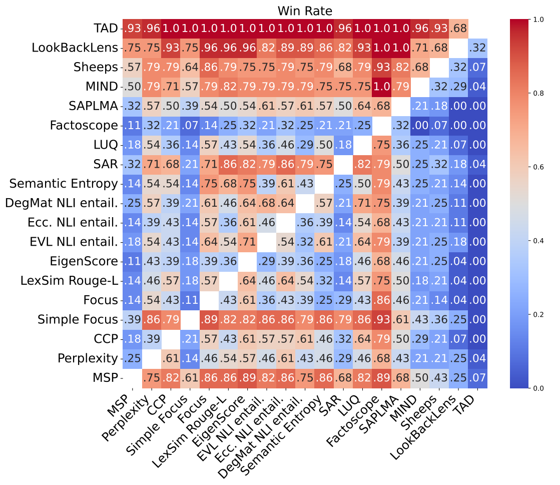

Figure 2: Summary of 30 experimental setups with various models and datasets. Each cell in the diagram presents the fraction of experiments where a method from a row outperforms a method from a column. Warmer colors indicate better results.

UQ Method	Llama-3.1 8b	Gemma-2 9b	Qwen-2.5 7b	Mean Rank
MSP	6.90	6.44	6.44	4.00
Perplexity	12.30	11.56	11.78	12.17
CCP	10.50	11.56	12.78	12.50
Simple Focus	7.20	6.89	6.44	4.67
Focus	12.40	11.44	14.89	14.00
Lexical Similarity Rouge-L	12.30	12.33	11.56	12.50
EigenScore	14.40	14.44	11.78	16.50
EVL NLI Score entail.	12.70	11.22	11.00	10.83
Ecc. NLI Score entail.	13.30	12.56	12.56	15.67
DegMat NLI Score entail.	11.30	11.67	11.00	10.83
Semantic Entropy	13.20	9.78	11.22	11.00
SAR	9.20	8.44	8.11	6.67
LUQ	13.50	13.00	11.89	16.00
Factoscope	14.30	16.22	16.78	18.67
SAPLMA	10.30	10.11	12.89	11.33
MIND	6.50	7.22	7.33	5.00
Sheeps	4.20	9.00	6.44	4.67
LookBackLens	<u>4.10</u>	<u>4.56</u>	<u>3.33</u>	<u>2.00</u>
TAD	**1.40**	**1.56**	**1.78**	**1.00**

Table 2: Mean ranks of UQ methods aggregated over all datasets for each LLM separately (the lower the better). The column *Mean Rank* corresponds to the mean rank of the ranks across all LLMs. The best method is in **bold**, the second best is <u>underlined</u>.

provements of the proposed method. Despite some baselines showing good results in particular cases, they usually are quite unstable, resulting in poor overall ranking. TAD demonstrates more robust improvements across multiple tasks and LLMs, making it a better choice overall.

Generalization to out-of-domain datasets. Table 3 compares the results of the supervised methods trained on all training datasets except for one that represents the out-of-domain dataset for testing. This setting evaluates the out-of-domain generalization capabilities of the supervised techniques. The results show that all considered supervised methods substantially degrade compared to their in-domain performance and in many cases, underperform MSP. Nevertheless, TAD demonstrates strong out-of-domain performance on the CoQA, SciQ, MMLU, and GSM8k datasets. On the MMLU dataset, TAD outperforms the MSP method, but falls behind the Sheeps method.

UQ Method	CNN	WMT19	MedQUAD	CoQA	SciQ	MMLU	GSM8k	Mean
	AlignScore	Comet	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR
MSP	**.157**	**.569**	**.356**	<u>.450</u>	<u>.582</u>	.444	<u>.380</u>	**.420**
Factoscope	.023	.131	.166	.007	.129	-.022	-.082	.050
SAPLMA	.021	-.003	.137	.012	.151	-.034	.073	.051
MIND	<u>.048</u>	<u>.258</u>	.095	.171	.222	.415	.335	.220
Sheeps	-.021	.059	.044	.201	.364	**.624**	.348	.231
LookBackLens	-.032	.069	.061	.111	.331	.224	.261	.147
TAD	.003	.192	<u>.336</u>	**.461**	**.601**	<u>.489</u>	**.391**	<u>.353</u>

Table 3: PRR↑ for Llama 8b v3.1 model for various tasks for the considered supervised sequence-level methods trained on the general dataset. Warmer colors indicate better results. The best method is in **bold**, and the second best one is <u>underlined</u>.

These findings indicate that supervised UQ methods are generally effective only in in-domain experimental setups. However, in specific scenarios, the TAD method demonstrates the ability to achieve generalization. More details about these experiments are presented in Appendix A.2.

4.3 Ablation Studies

Comparison of features. Table 9 in Appendix A.3 presents the ablation experiment with different features for the regression model in the TAD method. For *TAD (probs.)*, we only use probabilities along with predictions from the preceding tokens $p(t_{i-k} = \mathrm{T})$ for $k = 1, \ldots, N$. For *TAD (attention)*, we use attention weights on the N preceding tokens without probabilities. The results show that *TAD (probs.)* provides meaningful but relatively low performance. *TAD (attention)* demonstrates substantial improvements, underscoring the importance of using the attentions in the TAD method. The final version, *TAD (attention+probs.)*, which combines both attention weights and probabilities, achieves slight but consistent performance gains. This indicates the potential benefit of taking into account the information about the uncertainty from the previous tokens.

Impact of the two-step training procedure. Table 10 in Appendix A.3 presents the ablation experiment comparing one-step vs. two-step training procedures for the TAD method. The results show that the two-step procedure substantially outperforms the single-step training. These results highlight the importance of avoiding the teacher forcing, where the ground truth labels from previous tokens are used as features for linear regression. By using the training process with two steps, the model can avoid overfitting on the ground truth information, and achieve better performance.

Regression models and aggregation approaches. Detailed results with various regression models and aggregation approaches are presented in Table 7. The optimal values of the hyper-parameters of TAD for all experimental setups are presented in Tables 11 to 13 in Appendix C.1 for LLaMA-3.1 8b, Gemma-2 9b, and Qwen-2.5 7b, respectively.

We compared two strategies for aggregating the token-level TAD scores: (*i*) the mean of the scores and (*ii*) the sum of the log scores inspired by perplexity. For the majority of the considered settings, the mean of the probabilities yielded the best results. However, for QA with short answers, the sum of the log probabilities performed slightly better.

We can see that the difference between MLP and LinReg is minimal. On average, TAD with LinReg outperforms TAD with MLP by 0.029 in PRR. Therefore, for simplicity, we use LinReg as a regression method for TAD.

Impact of the number of previous tokens. Table 8 presents experiments with different numbers of preceding tokens used in TAD. The results show that using ten preceding tokens generally yields better performance compared to using only 1–2 tokens across all datasets, except for SamSum.

Impact of the attention layers. Figure 3 in Appendix A.3 presents the normalized average weights of linear regression for different attention layers in the TAD method. We can see similar patterns across various tasks, revealing that the most important layers are typically the middle ones, which is consistent with observations in previous work (Azaria and Mitchell, 2023; Chen et al., 2024). Additionally, we note that for the majority of the tasks, the first and the last attention layers play a crucial role.

4.4 Computational Efficiency

In order to demonstrate the computational efficiency of TAD, we compare its runtime to other UQ methods. The runtime is based on experiments conducted using a single 80GB H100 GPU, as detailed in Table 1. The inference is implemented as a single-batch model call for all tokens in the output text.

Table 4 presents the average runtime per text instance for each UQ method, along with the percentage overhead over the standard LLM inference with MSP. As we can see, many state-of-the-art UQ methods such as (black-box, Semantic Entropy, and SAR) introduce huge computational overhead (400-600%) because they need to perform sampling from the LLM multiple times. In contrast, all supervised methods introduce minimal overhead. In particular, TAD introduces only 5% overhead, which makes it a highly practical and efficient choice for uncertainty quantification.

UQ Method	**Runtime per batch**	**Overhead**
MSP	1.30±0.62	-
DegMat NLI Score Entail.	6.86±2.28	430 %
Lexical Similarity ROUGE-L	6.72±2.24	420%
Semantic Entropy	6.86±2.28	430%
SAR	8.83±2.94	580%
Factoscope	3.30±2.13	150%
SAPLMA	1.30±0.62	**0.06%**
MIND	1.30±0.62	0.10%
Sheeps	1.50±0.97	15%
LookBackLens	1.30±0.62	0.08%
TAD	1.37±0.68	5%

Table 4: 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**, and the second best results are underlined.

5 Conclusion and Future Work

We have presented a new uncertainty quantification method based on learning conditional dependencies between the predictions made on multiple generation steps. The method relies on attention to construct features for learning this functional dependency and leverages this dependency to alter the uncertainty of the subsequent generation steps. This yields improved results in selective generation tasks, especially when the LLM output is long. Our experimental study shows that our proposed technique usually outperforms other state-of-the-art UQ methods (such as SAR) resulting in the best overall performance across three LLMs and nine datasets. TAD does not introduce much computational overhead due to the simplicity of the regression model (linear regression), which makes it a potentially practical choice for LLM-based applications.

In future work, we aim to apply the suggested method to quantifying the uncertainty of retrieval-augmented LLMs. TAD potentially could be used to take into account the credibility of the retrieved evidence.

Limitations

In the motivation of our approach, we assumed a strict Markov chain property between the generation steps. However, in reality, this property does not hold as the current generation step usually depends on multiple previous steps. This limitation of our method could be addressed by estimating the conditional dependency between multiple previous steps, e.g., by using a Transformer layer instead of the linear regressor. Nevertheless, our current implementation, based on the Markov assumption, already yields strong results, and thus we leave investigation of more complex modifications for future work.

We also did not test our method on extra large LLMs such as LLaMA 3 70b. We only used 7-9b models due to limitations in our available computational resources.

Ethical Considerations

Responsible use. In our work, we considered open-weights LLMs and datasets not aimed at harmful content. However, LLMs may generate potentially damaging texts for various groups of people. Uncertainty quantification techniques can help create more reliable use of neural networks. Moreover, they can be applied to detecting harmful generation, but this is not our intention.

Limited applicability. Moreover, despite that our proposed method demonstrates sizable performance improvements, it can still mistakenly highlight correct and not dangerous generated text with high uncertainty in some cases. Thus, as with other uncertainty quantification methods, it has limited applicability.

Human subject considerations. All annotators in our study provided informed consent, were fully aware of the study's objectives, and had the right to withdraw at any time.

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*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. 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*.

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.

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*.

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*, pages 9367–9385, Bangkok, Thailand. 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, Singapore. 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 ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372.

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.

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.

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*. OpenReview.net.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2022. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*.

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.

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*.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning confidence for transformer-based neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

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.

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*.

Yookoon Park and David Blei. 2024. Density uncertainty layers for reliable uncertainty estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 163–171. PMLR.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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*.

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.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

Junya Takayama and Yuki Arase. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138. Association for Computational Linguistics.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965. Association for Computational Linguistics.

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*.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. 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.

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.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. 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 with other UQ Methods

Here, we present the main results for Gemma and Qwen.

UQ Method	SamSum	CNN	WMT19	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k	Mean	Mean
	AlignScore	AlignScore	Comet	AlignScore	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR	Rank
MSP	.370	.061	.588	.187	.527	.614	.772	.771	.425	.479	6.44
Perplexity	.008	-.036	.480	.171	.517	.178	.779	.756	.225	.342	11.56
CCP	.266	.031	.432	.102	.448	.450	.769	.678	.482	.406	11.56
Simple Focus	.308	.066	.578	.178	.543	.583	.770	.755	.436	.469	6.89
Focus	.110	-.040	.494	.198	.446	.528	.721	.721	.419	.400	11.44
Lexical Similarity Rouge-L	.077	.071	.458	-.002	.453	.453	.751	.587	.544	.377	12.33
EigenScore	.134	.085	.368	-.144	.456	.452	.701	.473	.355	.320	14.44
EVL NLI Score entail.	.143	.089	.373	.035	.469	.464	.750	.606	.486	.380	11.22
Ecc. NLI Score entail.	.073	.047	.393	-.020	.487	.478	.742	.609	.512	.369	12.56
DegMat NLI Score entail.	.147	.090	.381	.034	.427	.466	.762	.465	.514	.365	11.67
Semantic Entropy	.181	.078	.521	-.039	.490	.473	.744	.673	.546	.407	9.78
SAR	.107	.087	.491	.069	.496	.472	**.781**	.690	.545	.415	8.44
LUQ	.104	.114	.261	.140	.411	.430	.755	.503	.451	.352	13.00
Factoscope	.090	.063	.088	-.093	-.056	.480	.289	.542	.084	.165	16.22
SAPLMA	.318	.019	.600	.375	-.005	.535	.601	.535	.604	.398	10.11
MIND	.292	.098	.608	.511	.345	.524	.528	.782	.702	.488	7.22
Sheeps	.304	.080	.638	.397	.358	.439	.551	.733	.756	.473	9.00
LookBackLens	**.475**	.194	**.672**	.481	.465	**.666**	.685	.750	.712	.567	4.56
TAD	.462	**.219**	.643	**.575**	**.555**	.641	.773	**.812**	**.769**	**.605**	**1.56**

Table 5: PRR↑ for Gemma 9b v2 model for various tasks for the considered sequence-level methods. Warmer color indicates better results. The best method is in **bold**, the second best is underlined.

UQ Method	SamSum	CNN	WMT19	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k	Mean	Mean
	AlignScore	AlignScore	Comet	AlignScore	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR	Rank
MSP	.394	.148	.582	.210	.490	.661	.706	.508	.455	.461	6.44
Perplexity	.114	.047	.503	.232	.461	.447	.717	.310	.536	.374	11.78
CCP	.374	.117	.455	.131	.398	.422	.707	.197	.470	.363	12.78
Simple Focus	.317	.093	.570	.250	.513	.639	.718	.449	.490	.449	6.44
Focus	.156	.056	.503	.196	.356	.500	.643	-.351	.474	.282	14.89
Lexical Similarity Rouge-L	.244	.059	.485	.151	.400	.553	.653	.381	.683	.401	11.56
EigenScore	.050	.054	.489	.089	.426	.643	.643	.364	.709	.385	11.78
EVL NLI Score entail.	.206	.091	.383	.270	.468	.595	.675	.290	.572	.394	11.00
Ecc. NLI Score entail.	.186	.036	.439	.216	.401	.598	.648	.342	.590	.384	12.56
DegMat NLI Score entail.	.214	.091	.418	.263	.419	.546	.699	.319	.593	.396	11.00
Semantic Entropy	.262	.081	.514	.189	.458	.589	.674	.252	.564	.398	11.22
SAR	.238	.076	.515	.224	.475	.634	.707	.333	.708	.435	8.11
LUQ	.123	.075	.314	.278	.423	.543	.682	.321	.607	.374	11.89
Factoscope	.064	.016	.134	.038	.205	.447	.467	.821	-.368	.203	16.78
SAPLMA	.283	.030	.416	.316	-.035	.442	.519	.432	.643	.338	12.89
MIND	.316	.124	.308	.369	.489	.640	.639	.890	.783	.506	7.33
Sheeps	.395	**.180**	.515	.387	.380	.429	.704	.900	**.837**	.525	6.44
LookBackLens	**.445**	.159	.571	.398	.434	**.703**	.708	.848	.753	.558	3.33
TAD	.434	.140	**.607**	**.468**	**.515**	.648	**.728**	**.904**	.825	**.585**	**1.78**

Table 6: PRR↑ for Qwen 7b v2.5 model for various tasks for the considered sequence-level methods. Warmer color indicates better results. The best method is in **bold**, the second best is underlined.

A.2 Generalization Experiments

In this experiment, we examine how our approach can be generalized on the unseen datasets. For each dataset, we create a general training dataset by using 300 samples from the training datasets from each of the eleven other datasets used in the experiments. Thus, we evaluate TAD that is not trained on the target dataset. We conduct experiments on one dataset from each task: CNN, WMT19, MedQUAD, CoQA, SciQ, MMLU, and GSM8k. We compare the results with the baseline MSP method. Overall, we can see that the TAD method can be generalized on unseen datasets and outperform all other baselines in specific settings.

A.3 Ablation Studies

Here, we present ablation studies for various numbers of the preceding tokens, different features, and the impact of various layers for the TAD method.

UQ Method	Aggregation	SamSum	CNN	WMT19	MedQUAD	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k	Mean	Mean
		AlignScore	AlignScore	Comet	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR	Rank
TAD (LinReg)	$\frac{1}{K}\sum_{k=1}^{K} p_k$	**.431**	.215	**.612**	**.662**	**.565**	**.543**	.542	.757	.806	**.682**	**.581**	**1.80**
TAD (LinReg)	$\sum_{k=1}^{K} \log p_k$	.348	**.245**	.462	.307	.450	.509	**.644**	.737	.816	.605	.512	2.80
TAD (MLP)	$\frac{1}{K}\sum_{k=1}^{K} p_k$	.402	.208	.602	.591	.482	.526	.491	**.764**	.814	.645	.552	2.40
TAD (MLP)	$\sum_{k=1}^{K} \log p_k$	.375	.239	.397	.222	.461	.482	.626	.746	**.818**	.522	.489	3.00

Table 7: Comparison of various considered regression models and aggregation strategies for TAD (PRR↑, Llama 8b v3.1 model). Warmer colors indicate better results.The best method is in **bold**, the second best is underlined.

UQ Method	SamSum	CNN	WMT19	MedQUAD	TruthfulQA	GSM8k	Mean
	AlignScore	AlignScore	Comet	AlignScore	AlignScore	Acc.	PRR
TAD (1 tokens)	**.228**	.425	.602	.570	.519	.659	.501
TAD (2 tokens)	.224	.424	.606	.596	.537	.679	.511
TAD (5 tokens)	.219	.397	**.618**	.628	.556	**.687**	.517
TAD (10 tokens)	.215	**.431**	.612	**.662**	**.565**	.682	**.528**

Table 8: PRR↑ for Llama 8b v3.1 model for various tasks for the various number of preceding tokens for the TAD method. Warmer color indicates better results. The best method is in **bold**, the second best is underlined.

UQ Method	SamSum	CNN	WMT19	MedQUAD	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k	Mean
	AlignScore	AlignScore	Comet	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR
TAD (probs.)	.178	.086	.411	.437	.270	.444	.567	.683	.668	.374	.412
TAD (attention)	.426	.212	.611	**.670**	**.566**	.480	.632	.712	.804	.673	.579
TAD (attention+probs.)	**.431**	**.215**	**.612**	.662	.565	**.509**	**.644**	**.737**	**.806**	**.682**	**.586**

Table 9: PRR↑ for Llama 8b v3.1 model for various tasks for different features for the TAD method. Warmer color indicates better results. The best method is in **bold**, the second best is underlined.

UQ Method	SamSum	CNN	WMT19	MedQUAD	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k	Mean
	AlignScore	AlignScore	Comet	AlignScore	AlignScore	AlignScore	AlignScore	AlignScore	Acc.	Acc.	PRR
TAD (1 step)	.107	.043	.281	.057	.168	.421	.499	.677	.397	.285	.294
TAD (2 step)	**.431**	**.215**	**.612**	**.662**	**.565**	**.509**	**.644**	**.737**	**.806**	**.682**	**.586**

Table 10: PRR↑ for Llama 8b v3.1 model for various tasks for the different number of learning steps for the TAD method. Warmer color indicates better results. The best method is in **bold**.

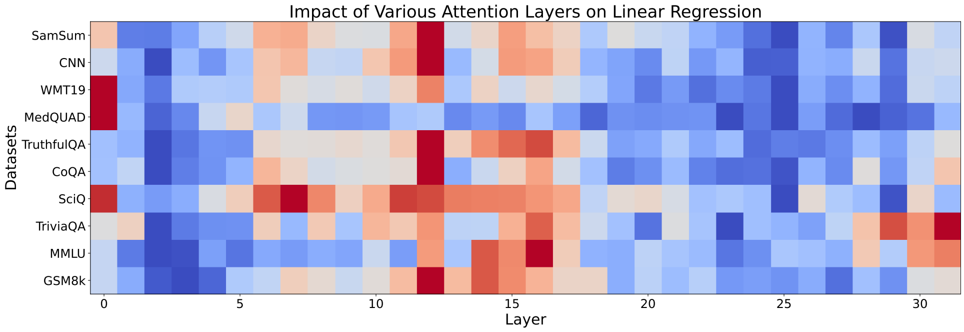

Figure 3: Normalized average weights of linear regression for different attention layers in the TAD method across the considered datasets. Warmer color indicates a higher impact on the TAD performance.

B Computational Resources and Efficiency

All experiments were conducted on a single NVIDIA H100 GPU. On average, training a single model across all datasets took over 750 GPU hours, while inference on the test set took 260 GPU hours.

C Hyperparameters

C.1 Optimal Hyperparameters for TAD

The optimal hyperparameters for TAD for various considered regression models and different aggregation strategies are presented in Tables 11 to 13 for Llama-3.1 8b, Gemma-2 9b, and Qwen-2.5 7b models respectively. These hyperparameters are obtained using cross-validation with five folds using the training dataset. We train a regression model on $k - 1$ folds of the training dataset and estimate uncertainty on the remaining fold. The optimal hyperparameters are selected according to the best average PRR for AlignScore. Finally, we use these hyperparameters to train the regression model on the entire training set.

The hyperparameter grid for the linear regression is the following:
L2 regularization: [1e+1, 1, 1e-1, 1e-2, 1e-3, 1e-4].

The hyperparameter grid for the MLP is the following:
Num. of layers: [2, 4];
Num. of epochs: [10, 20, 30];
Learning rate: [1e-5, 3e-5, 5e-5];
Batch size: [64, 128].

UQ Method	Aggregation	SamSum	CNN	WMT19	MedQUAD	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k
TAD (MLP)	$\frac{1}{K}\sum_{k=1}^{K} p_k$	4, 30, 1e-05, 0, 128	4, 30, 3e-05, 0, 128	4, 30, 3e-05, 0, 128	4, 30, 1e-05, 0, 128	4, 30, 5e-05, 0, 128	4, 30, 3e-05, 0, 64	2, 30, 5e-05, 0, 128	4, 30, 3e-05, 0, 128	4, 30, 5e-05, 0, 128	4, 30, 1e-05, 0, 128
TAD (MLP)	$\sum_{k=1}^{K} \log p_k$	4, 30, 5e-05, 0, 64	4, 30, 1e-05, 0, 128	2, 20, 5e-05, 0, 64	4, 30, 5e-05, 0, 64	4, 30, 5e-05, 0, 64	2, 30, 5e-05, 0, 64	4, 30, 5e-05, 0, 128	4, 30, 3e-05, 0, 128	4, 30, 5e-05, 0, 64	4, 30, 3e-05, 0, 128
TAD (LinReg)	$\frac{1}{K}\sum_{k=1}^{K} p_k$	1	10.0	1	0.01	1	1	0.001	10.0	1	10.0
TAD (LinReg)	$\sum_{k=1}^{K} \log p_k$	1	1	0.0001	0.001	0.1	10.0	10.0	1	1	0.01

Table 11: Optimal values of the hyper-parameters for the TAD methods for the Llama 8b v3.1 model.

UQ Method	SamSum	CNN	WMT19	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k
TAD (LinReg)	10.0	10.0	0.0001	1	10.0	1	10.0	10.0	1

Table 12: Optimal values of the hyper-parameters for the final configuration of the TAD method for the Gemma 9b v2 model.

UQ Method	SamSum	CNN	WMT19	TruthfulQA	CoQA	SciQ	TriviaQA	MMLU	GSM8k
TAD (LinReg)	0.1	0.01	1	0.01	10.0	10.0	10.0	1	1

Table 13: Optimal values of the hyper-parameters for the final configuration of the TAD method for the Qwen 7b v2.5 model.

C.2 LLM Generation Hyperparameters

Dataset	Task	Max Input Length	Generation Length	Temperature	Top-p	Do Sample	Beams	Repetition Penalty
SamSum	TS	-	128	1.0	1.0	False	1	1
CNN			128					
WMT19	MT		107					
MedQUAD	QA Long answer		128					
TruthfulQA			128					
GSM8k			256					
CoQA	QA Short answer		20					
SciQ			20					
TriviQA			20					
MMLU	MCQA		3					

Table 14: Values of the text generation hyper-parameters for all LLMs used in our experiments.

D Dataset Statistics

Task	Dataset	N-shot	Train texts for TAD	Evaluation texts
Text Summarization	CNN/DailyMail	0	2,000	2,000
	SamSum	0	2,000	819
MT	WMT19 De-En	0	2,000	2,000
QA Long answer	MedQUAD	5	700	2,000
	TruthfulQA	5	408	409
	GSM8k	5	700	1,319
QA Short answer	SciQ	0	2,000	1,000
	CoQA	all preceding questions	2,000	2,000
	TriviaQA	5	2,000	2,000
MCQA	MMLU	5	2,000	2,000

Table 15: Statistics about the datasets used for evaluation.