From Pixels to Tokens: Otsu Thresholding for Accurate yet Confident Decoding

Anonymous ACL submission

Abstract

This paper introduces a new decoding approach, Otsu thresholding, inspired by the well-known image processing technique, which computes a threshold between low-confidence and high-confidence tokens for each generation step. The method is tested across several tasks and instruction-based models, revealing results competitive to other state-of-the-art decoding methods, while keeping output quality. Also, it is shown that tuning specific characteristics of the model, such as the pass from a softmax function of the distribution and the limitation to K highest candidates before selecting a token, can highly affect model performance. In general, this approach is a great alternative method that offers a good balance between quality, confidence, uncertainty, and diversity.

1 Introduction

In recent years, Large Language Models (LLMs) have shown great potential in generative tasks, such as text summarization and question answering (Jurafsky and Martin, 2025). According to the latest leaderboard[1], they are able to extract high accuracy output and therefore, create high quality content. However, a common problem of LLMs is the overestimation of their confidence ("they don't always know what they don't know" (Baan et al., 2023; Hu et al., 2023)), leading to outputs with hallucinations or misleading information. Ideally, a high-performance LLM should be able to generate text that is both accurate and confident, while providing a principled estimate of uncertainty, when knowledge is insufficient.

Decoding strategies can play a key role towards the aforementioned goal, and several strategies have been proposed that account for token probabilities to control the token prediction set. The choice of the decoding method for an LLM is very important, as it impacts its performance (Shi et al., 2024). However, most decoding methods fail to generate accurate and confident text, and they frequently result in non-representative prediction sets, with too few or, more often, too many low-probability tokens being considered. Such low-probability candidates correspond to regions of higher epistemic uncertainty in the model's predictive distribution. As a result, existing approaches overlook the need for fast, real-time decoding and for uncertainty estimation, leaving a critical gap in how uncertainty is represented in text generation (Shao et al., 2018; Hewitt et al., 2022; Holtzman et al., 2019). A solution to this problem could be the proposal of a new technique that efficiently distinguishes low-probability tokens from high-probability ones in a data-driven and uncertainty-aware manner.

The motivation of this paper is to propose an innovative and efficient decoding strategy that separates low-probability tokens from high-probability ones and mitigates uncertainty by suppressing low-confidence token candidates, and maintains strong generation quality. The strategy uses a very popular filtering method from image processing named Otsu thresholding. To the best of our knowledge, this is the first work to apply Otsu thresholding to probabilistic token filtering in LLM decoding. Thus, we check the impact of some features in Otsu, such as the usage of a softmax function on the logits of the tokens and the restriction of the number of candidates in the list. Finally, we estimate LLM performance with this approach and we compare the best combination to other sampling methods, finding competitive values with adaptive middle ground for certain and calibrated outputs.

To summarize, the main contributions of the paper are the following:

- We introduce Otsu thresholding as a novel uncertainty-aware decoding strategy that adaptively separates low-confidence from high-confidence token candidates, based on the

[1] https://huggingface.co/open-llm-leaderboard

model's predictive distribution.

- We analyze the proposed decoding strategy over different tasks and models.
- We experiment with the usage of a softmax function and the limitation to K candidates in the approach to understand their impact on the model output and we compare the best approach to other state-of-the-art decoding strategies to demonstrate favorable trade-offs in accuracy, confidence, uncertainty, and diversity.

2 Related Work

Outstanding studies have focused on the exploration of the impact of **decoding strategies on text uncertainty** in Large Language Models (Hashimoto et al., 2025). (Daheim et al., 2025) showed how Minimum Bayes Risk (MBR) decoding, which selects model generations according to an expected risk, can be generalized into a principled uncertainty-aware decoding method. Additionally, (He et al., 2025) proposed an uncertainty-guided adaptive decoding framework that integrates a token-level pause-then-rerank mechanism, driven by token uncertainty (Shannon entropy). AdaDec learns model-specific uncertainty thresholds and applies a lookahead-based reranking strategy, when uncertainty is high. Finally, the very recent work of (Lee et al., 2025) suggested Uncertainty-Aware Contrastive Decoding (UCD), which introduces a cumulative energy function, where uncertainty is quantified as the negative log-sum-exp over logits, and decomposed into entropy and expected logit components.

Evaluation of factuality in text generation is an ongoing challenge. (Zha et al., 2023) introduced AlignScore, a holistic metric, based on a general function of information alignment of text and its unified framework, which achieved substantial improvements over previous metrics. Additionally, (Min et al., 2023) advocated a new evaluation metric that computes factual accuracy from pieces of generated text and was used to compare the performance of different LLMs. Finally, (Bishop et al., 2023) proposed a new evaluation framework, LongDocFACTScore, for detecting human factuality, targeting specifically summarized, long documents.

Previous work has comprehensively examined **uncertainty in Natural Language Generation** (NLG) systems (Baan et al., 2023; Hu et al., 2023) and has explored strategies to address uncertainty with the goal of making LLMs more trustworthy. Firstly, (Xu et al., 2020) studied summarization decoders in both blackbox and whitebox ways by focusing on the entropy of the models' predictions and revealed that features, such as the sentence position and the syntactic distance between adjacent pairs of tokens, influence uncertainty. Moreover, (Ulmer et al., 2024) focused on token-level uncertainty and proposed a method for non-exchangeable conformal prediction, which was shown to improve text generation quality. Finally, (Fadeeva et al., 2024) introduced a token-level uncertainty method named Claim Conditioned Probability (CCP), disentangling claim-specific uncertainty from model decisions on surface forms, etc.

Diversity of LLMs is an ongoing challenge in text generation, as many tokens that could be correctly selected are ignored from the sampling methods. Studies attempt to estimate it, such as (Vijayakumar et al., 2016), which used diversity statistics by computing the number of distinct n-grams in a summary divided by the total number of tokens in it.

3 Methodology

In this work, we use Otsu thresholding as an alternative decoding strategy. Otsu thresholding is based on an automated technique from image processing (Otsu et al., 1975), where the algorithm returns a single intensity threshold t that separates pixels into two classes, background (C_0) and foreground (C_1). The goal is to find an optimal threshold that maximizes inter-class variance:

$$\sigma_b^2(argmax) = \max_{0<t<L} \sigma_b^2(t), \tag{1}$$

where L is the number of bins in the image and $\sigma_b^2(t)$ is the inter-class variance.

Let ω_0 and ω_1 be the cumulative probabilities of C_0 and C_1 respectively:

$$\omega_0(t) = \sum_{i=0}^{t-1} P(i), \tag{2}$$

$$\omega_1(t) = \sum_{i=t}^{L-1} P(i), \tag{3}$$

where $P(i|C_0)$ and $P(i|C_1)$ are the conditional probabilities of selecting the i-th pixel in C_0 and C_1.

Also, let $\mu_0(t)$ and $\mu_1(t)$ be the mean pixel intensities of C_0 and C_1 respectively:

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} iP(i)}{\omega_0(t)}, \quad (4)$$

$$\mu_1(t) = \frac{\sum_{i=t}^{L-1} iP(i)}{\omega_1(t)}. \quad (5)$$

The algorithm performs global search, where at each pass of an intensity, it updates these variables and re-computes inter-class variance.

In summary, the steps of the algorithm are the following:

1. Compute histogram and probabilities of each intensity level.
2. Initialize $\omega_0(t)$, $\mu_0(t)$, $\omega_1(t)$, and $\mu_1(t)$.
3. For each pixel intensity, update $\omega_0(t)$, $\mu_0(t)$, $\omega_1(t)$, and $\mu_1(t)$ and then, compute $\sigma_b^2(t)$.

Consequently, in our approach, the histogram is a distribution of token logits at each decoding step. Otsu computes a threshold from the distribution and distinguishes tokens with low values (C_0) from those with high values (C_1). After that, the model samples the tokens in C_1 and uniformly chooses one as the next token. From an uncertainty perspective, Otsu thresholding identifies a data-driven boundary between these regions, enabling filtering of low-confidence token candidates that are associated with hallucinated or overconfident generations. A simple comparison between the 2 types of Otsu thresholding techniques, along with an example from the dataset, can be found in Figure 1.

We experiment with different variations of the proposed decoding strategy in an attempt to understand their impact on the results and find the best combination. Specifically, we experiment with the pass of the logit distribution from a softmax function before applying Otsu to convert logits to probabilities, and we limit the candidate set to K tokens with the highest values at each decoding step to observe whether the selected option mostly originates from the first items.

4 Experimental SetUp

4.1 Datasets

The dataset that we use for our experiments is a pre-processed version[2] of Super-Natural Instructions (Wang et al., 2022), consisting of definitions, inputs, and targets for many tasks. To test the robustness and generalization of our method, we use samples from the following 3 different tasks:

- Amazon Review: Given an Amazon product review, the goal is to generate a title from it.
- Extreme Abstract: Given the abstract of a research paper, the goal is to summarize it in no more than 30 words.
- Web Questions: Given a web question, the goal is to answer it.

The above tasks don't contain the same amount of records (Amazon Review has 13,000 records, Extreme Abstract has 10,796 records, and Web Questions has 17,850 records), so we keep the first 600 records from each task for a fair comparison between them.

The prompt that we use for these tasks is:
$definition\ input\ response$,
where $definition$ and $input$ are retrieved as features from the dataset and $response$ is "Output:", "Summary:", and "Answer:" for each task respectively.

The definitions from the dataset that the above tasks have are the following:

- Amazon Review: In this task, you're given reviews from Amazon's products. Your task is to generate the Summary of the review.
- Extreme Abstract: In this task, you are given the abstract of a research paper. Your task is to generate a summary of this abstract. Your summary should not be very short, but it's better if it's not more than 30 words.
- Web Questions: A question is presented to you in this task, and your job is to write a potentially correct answer.

4.2 Large Language Models

For the scope of our research, we use two open-access, state-of-the-art, instruction-tuned LLMs, selected to cover different model architectures and design choices.

LLaMA-8B-v3.1 is an auto-regressive transformer model, trained using a combination of Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) to align

[2] https://huggingface.co/datasets/Muennighoff/natural-instructions

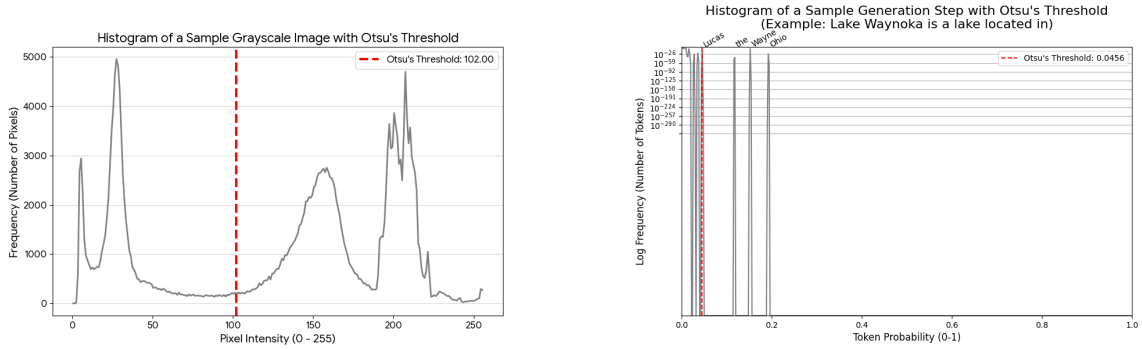

Figure 1: Example of Otsu's method applied to an image and a token generation step.

its outputs with human preferences for helpfulness and safety (Touvron et al., 2023).

Qwen-v2.5 is a high-capacity LLM that supports context lengths of up to 128,000 tokens and a maximum generation length of 8,000 tokens. It is available in 7B and 14B parameter variants and incorporates several architectural enhancements that enable stable and high-quality generation over long contexts (Team et al., 2024).

4.3 Decoding strategies

We compare Otsu decoding against 2 decoding strategies to identify those producing outputs that are both accurate and well-calibrated in terms of quality, factuality, certainty, and diversity. Specifically, we compare:

- Epsilon sampling: It filters tokens whose probability is equal or greater than an epsilon cutoff, effectively removing low probability candidates from the sampling space (Hewitt et al., 2022).
- Top-p sampling: It chooses tokens from the smallest possible set whose cumulative probability exceeds a predefined threshold p. The probability mass is then redistributed among the selected tokens (Holtzman et al., 2019).

4.4 Hyperparameter settings

Our experiments are conducted on a Google Cloud VM instance, equipped with 208 vCPUs, 1,872 GB RAM, 100 GB memory disk, and 8 NVIDIA H100 GPUs, each with 80 GB of memory. For all the text generation strategies, we set the number of generated sequences to 5, epsilon cutoff to 9^e-4, K to 50, and p to 0.75, wherever applicable. These values are selected to balance computational efficiency with reliable estimation of all the types of metrics.

4.5 Evaluation metrics

We evaluate LLM performance across the selected tasks using multiple types of metrics.

Quality metrics compare the generated prediction to a true reference. ROUGEL measures the longest common subsequence (LCS) between the 2 types of text (Lin, 2004). BERTScore leverages pre-trained contextual embeddings to match tokens in candidate and reference text using cosine similarity (Zhang et al., 2019) and provides an F1 score. As an additional step of **benchmarking** our work, we consider ROUGEL ideal for the review generation and abstract summarization task, while, for question answering, we use an F1 score that is computed over the individual words in the generated text against the true answer.

Factuality metrics assess the alignment between generated claims and ground-truth evidence. HHEM models extract a score from 0 to 1, indicating the level of hallucinations in text (Bao et al., 2024). AlignScore is another metric that produces values in the same range, using a unified information alignment function and splitting the 2 types of text to compute an average score from the maximum alignment scores of the pairs (Zha et al., 2023).

Certainty metrics quantify uncertainty at each generation step and are averaged per sample. Token certainty estimates the probability of the generated token, while token entropy computes the Shannon entropy from the probability distribution of the vocabulary (Shannon, 1948).

Diversity metrics capture lexical and semantic variation in the generated outputs. We estimate inter-generation diversity by comparing generated sequences for the same sample using BERTScore and keeping the remaining value from the subtraction to 1.

5 Results

5.1 Analysis of the features of the Otsu thresholding approach

As an initiative of understanding better the behavior of the proposed decoding strategy, we examine how specific design choices, such as the application of a softmax function to the logits and the limitation to K candidates, affect the performance of the model. We compare the combinations of these features across all the dataset categories, LLMs, and evaluation metrics, enabling analysis of their impact on quality, factuality, uncertainty, and diversity.

The results in Tables 1 and 2 reveal a clear trade-off between output diversity and the remaining evaluation metrics. Specifically, the use of a softmax function sharpens the output distribution, leading to more confident predictions, while consistently reducing diversity. In addition, restricting the analysis to the top 50 candidates yields performance that is comparable to, and, in some cases, slightly better than using all model predictions. This suggests that the most informative signal is concentrated within the top K candidates. Given that this restriction also offers improved computational efficiency, adopting a top 50 setting represents a practical and effective choice.

5.2 Comparison of Otsu thresholding to other decoding strategies

The best results from Otsu are then compared to other decoding strategies. We compare the different methods using all the dataset categories, LLMs, and evaluation metrics.

From Tables 3 and 4, it can be viewed that epsilon decoding generally increases quality and diversity metrics, but often at the cost of certainty. Top-p achieves the best values in uncertainty, but is not optimal in diversity. Across datasets and models, Otsu performs comparably to top-p and occasionally matches or exceeds it on key metrics without affecting diversity much, indicating that it is an effective, best-balanced, alternative decoding strategy.

5.3 Presentation of text generations

For a better understanding of the results, we present some examples of text generations from the experiments. In detail, we note the most certain summaries of the Extreme Abstract task that were generated from the different Otsu thresholing parameterizations and the other decoding strategies using the Qwen-7B model.

Otsu Thresholding (no softmax, all candidates)
ID: task668-f7af2fc4ebf24a6888575e5e6d566a38
Summary: The CWAE model uses a new distance metric with a closed-form solution, simplifying optimization and matching or outperforming WAE and SWAE.

Otsu Thresholding (softmax, all candidates)
ID: task668-f77032d675f041fba3261703979e2353
Summary: Proposed Prox-SGD for training NNs with nonsmooth regularization and constraints, ensuring convergence to stationary points and performing well on sparse and binary NNs. To promote a NN with specific structures, we explicitly take into consideration the nonsmooth regularization (such as L1-norm) and constraints (such as interval constraint). This is formulated as a constrained nonsmooth nonconvex optimization problem, and we propose a convergent proximal-type stochastic gradient descent (Prox-SGD) algorithm.

Otsu Thresholding (no softmax, 50 candidates)
ID: task668-725120e56d404fecaa47a0d3a94c3ef6
Summary: Morph-Net uses morphological ops for neuron computations, offering better performance and fewer params than standard nets on various datasets.

Otsu Thresholding (softmax, 50 candidates)
ID: task668-87bc07c2e6b0448e86afa5d967e41af8
Summary: The authors propose a method to minimize the policy's Lipschitz constant during training to improve robustness and efficiency in reinforcement learning with domain randomization.

Epsilon Sampling
ID: task668-7ba32c9b1af24c91a2161def01bc5a9e
Summary: This research explores the divergence minimization perspective on GAN training, comparing original and "non-saturating" variants, and develops theoretical tools for classifying f-divergences.

Top-p Sampling
ID: task668-f7af2fc4ebf24a6888575e5e6d566a38
Summary: A new generative model, Cramer-Wold AutoEncoder (CWAE), is proposed which uses Cramer-Wold distance for improved performance compared to existing methods like WAE and SWAE. CWAE simplifies optimization without compromising on quality.

Dataset	Model	Softmax Usage	Number of K	ROUGEL ↑	BERT-F1 ↑	HHEM ↑	AlignScore ↑
Amazon Review	LLaMA-8B	No	All	0.0363	0.8022	0.1721	0.0628
Amazon Review	LLaMA-8B	Yes	All	0.051	0.8019	**0.3544**	0.1058
Amazon Review	LLaMA-8B	No	50	**0.0515**	**0.8049**	0.3497	0.0973
Amazon Review	LLaMA-8B	Yes	50	0.0493	0.8019	0.349	**0.1092**
Amazon Review	Qwen-7B	No	All	**0.0518**	**0.8056**	**0.3614**	**0.0926**
Amazon Review	Qwen-7B	Yes	All	0.0358	0.8025	0.1592	0.0575
Amazon Review	Qwen-7B	No	50	0.0361	0.8031	0.1723	0.0602
Amazon Review	Qwen-7B	Yes	50	0.0359	0.8029	0.1696	0.0557
Extreme Abstract	LLaMA-8B	No	All	0.0757	**0.8243**	**0.3503**	0.0451
Extreme Abstract	LLaMA-8B	Yes	All	0.0745	0.8232	0.3383	**0.0461**
Extreme Abstract	LLaMA-8B	No	50	**0.0766**	0.8229	0.3381	0.0445
Extreme Abstract	LLaMA-8B	Yes	50	0.0733	0.8233	0.3342	0.0454
Extreme Abstract	Qwen-7B	No	All	0.0647	0.8236	0.3406	0.054
Extreme Abstract	Qwen-7B	Yes	All	0.0643	**0.8255**	**0.3542**	0.055
Extreme Abstract	Qwen-7B	No	50	**0.0653**	0.8235	0.336	**0.0556**
Extreme Abstract	Qwen-7B	Yes	50	0.0651	0.824	0.3503	0.0505
Web Questions	LLaMA-8B	No	All	0.0116	0.7728	0.407	0.043
Web Questions	LLaMA-8B	Yes	All	0.0117	0.772	**0.4086**	0.0454
Web Questions	LLaMA-8B	No	50	0.0116	**0.7736**	0.3956	0.0437
Web Questions	LLaMA-8B	Yes	50	**0.0119**	0.7723	0.4036	**0.0461**
Web Questions	Qwen-7B	No	All	**0.0124**	0.7768	0.3711	0.0429
Web Questions	Qwen-7B	Yes	All	0.0123	0.7767	**0.3852**	**0.0435**
Web Questions	Qwen-7B	No	50	0.0119	**0.7769**	0.3815	0.0412
Web Questions	Qwen-7B	Yes	50	0.0119	0.7765	0.3792	0.04

Table 1: Analysis of the features of the Otsu thresholding approach for each dataset category using all the LLMs and the quality and factuality metrics.

Dataset	Model	Softmax Usage	Number of K	Token Certainty ↑	Token Entropy ↓	Diversity Level ↑	F1-Q/A ↑
Amazon Review	LLaMA-8B	No	All	0.7759	0.5013	**0.4467**	-
Amazon Review	LLaMA-8B	Yes	All	**0.9274**	**0.1188**	0.3469	-
Amazon Review	LLaMA-8B	No	50	0.8307	0.3434	0.3878	-
Amazon Review	LLaMA-8B	Yes	50	0.9247	0.12	0.3473	-
Amazon Review	Qwen-7B	No	All	0.8253	0.3668	0.3924	-
Amazon Review	Qwen-7B	Yes	All	0.8684	0.2189	0.4286	-
Amazon Review	Qwen-7B	No	50	0.7857	0.4496	**0.4455**	-
Amazon Review	Qwen-7B	Yes	50	**0.8687**	**0.2171**	0.4274	-
Extreme Abstract	LLaMA-8B	No	All	0.8403	0.3067	**0.3555**	-
Extreme Abstract	LLaMA-8B	Yes	All	0.9083	0.1472	0.3323	-
Extreme Abstract	LLaMA-8B	No	50	0.8371	0.2981	0.3547	-
Extreme Abstract	LLaMA-8B	Yes	50	**0.9096**	**0.1464**	0.3332	-
Extreme Abstract	Qwen-7B	No	All	0.784	0.4699	**0.4112**	-
Extreme Abstract	Qwen-7B	Yes	All	0.8745	0.2073	0.3862	-
Extreme Abstract	Qwen-7B	No	50	0.7889	0.4384	0.4045	-
Extreme Abstract	Qwen-7B	Yes	50	**0.8758**	**0.2046**	0.3895	-
Web Questions	LLaMA-8B	No	All	0.8342	0.3444	**0.3835**	0.0167
Web Questions	LLaMA-8B	Yes	All	0.9098	0.1433	0.3365	**0.0171**
Web Questions	LLaMA-8B	No	50	0.8342	0.329	0.3778	0.0169
Web Questions	LLaMA-8B	Yes	50	**0.9107**	**0.1406**	0.3398	0.0169
Web Questions	Qwen-7B	No	All	0.7733	0.4917	**0.3666**	0.0173
Web Questions	Qwen-7B	Yes	All	0.867	0.2196	0.3349	**0.0176**
Web Questions	Qwen-7B	No	50	0.775	0.4692	0.3608	0.0169
Web Questions	Qwen-7B	Yes	50	**0.8676**	**0.2176**	0.3341	0.0167

Table 2: Analysis of the features of the Otsu thresholding approach for each dataset category using all the LLMs and the certainty and diversity metrics.

6 Conclusions

We proposed an innovative, uncertainty-aware decoding strategy that efficiently distinguishes low-probability from high-probability tokens. Our approach accomplished results competitive to other state-of-the-art methods, while keeping quality and diversity in text. Moreover, we showed that specific characteristics of the approach influence model output, but have a minor impact on diversity. In conclusion, the above research opens a new field of exploration in the area of probabilistic decoding using LLMs.

Dataset	Model	Method	ROUGEL ↑	BERT-F1 ↑	HHEM ↑	AlignScore ↑
Amazon Review	LLaMA-8B	Epsilon	0.0515	**0.8065**	**0.3693**	0.0942
Amazon Review	LLaMA-8B	Top-p	**0.053**	0.8032	0.3527	**0.1096**
Amazon Review	LLaMA-8B	Otsu (best)	0.0493	0.8019	0.349	0.1092
Amazon Review	Qwen-7B	Epsilon	**0.0362**	0.8027	**0.1784**	**0.0607**
Amazon Review	Qwen-7B	Top-p	0.0358	**0.8034**	0.1717	0.0601
Amazon Review	Qwen-7B	Otsu (best)	0.0359	0.8029	0.1696	0.0557
Extreme Abstract	LLaMA-8B	Epsilon	**0.0753**	0.8231	0.3422	0.0432
Extreme Abstract	LLaMA-8B	Top-p	0.0748	0.8226	**0.3501**	0.0442
Extreme Abstract	LLaMA-8B	Otsu (best)	0.0733	**0.8233**	0.3342	**0.0454**
Extreme Abstract	Qwen-7B	Epsilon	0.0642	0.8229	0.3363	**0.0528**
Extreme Abstract	Qwen-7B	Top-p	0.0643	**0.825**	0.342	0.0506
Extreme Abstract	Qwen-7B	Otsu (best)	**0.0651**	0.824	**0.3503**	0.0505
Web Questions	LLaMA-8B	Epsilon	0.0116	**0.7733**	**0.4122**	0.0431
Web Questions	LLaMA-8B	Top-p	0.0114	0.7703	0.3995	0.0459
Web Questions	LLaMA-8B	Otsu (best)	**0.0119**	0.7723	0.4036	**0.0461**
Web Questions	Qwen-7B	Epsilon	**0.0123**	**0.7769**	0.3681	**0.0433**
Web Questions	Qwen-7B	Top-p	0.0117	0.7762	0.3671	0.0406
Web Questions	Qwen-7B	Otsu (best)	0.0119	0.7765	**0.3792**	0.04

Table 3: Comparison of the best version of Otsu to the other decoding strategies for each dataset category using all the LLMs and the quality and factuality metrics.

Dataset	Model	Method	Token Certainty ↑	Token Entropy ↓	Diversity Level ↑	F1-Q/A ↑
Amazon Review	LLaMA-8B	Epsilon	0.8274	0.3468	**0.3898**	-
Amazon Review	LLaMA-8B	Top-p	0.9242	**0.1197**	0.3553	-
Amazon Review	LLaMA-8B	Otsu (best)	**0.9247**	0.12	0.3473	-
Amazon Review	Qwen-7B	Epsilon	0.7825	0.4683	**0.445**	-
Amazon Review	Qwen-7B	Top-p	0.8687	0.2186	0.427	-
Amazon Review	Qwen-7B	Otsu (best)	0.8687	**0.2171**	0.4274	-
Extreme Abstract	LLaMA-8B	Epsilon	0.8367	0.3019	**0.3561**	-
Extreme Abstract	LLaMA-8B	Top-p	**0.9117**	**0.1319**	0.3316	-
Extreme Abstract	LLaMA-8B	Otsu (best)	0.9096	0.1464	0.3332	-
Extreme Abstract	Qwen-7B	Epsilon	0.7877	0.4479	**0.4059**	-
Extreme Abstract	Qwen-7B	Top-p	0.8733	0.2065	0.3863	-
Extreme Abstract	Qwen-7B	Otsu (best)	**0.8758**	**0.2046**	0.3895	-
Web Questions	LLaMA-8B	Epsilon	0.8328	0.3326	**0.3801**	**0.0171**
Web Questions	LLaMA-8B	Top-p	**0.9164**	**0.1286**	0.3435	0.0162
Web Questions	LLaMA-8B	Otsu (best)	0.9107	0.1406	0.3398	0.0169
Web Questions	Qwen-7B	Epsilon	0.7756	0.4712	**0.3632**	0.0167
Web Questions	Qwen-7B	Top-p	0.8671	**0.2155**	0.3357	0.0162
Web Questions	Qwen-7B	Otsu (best)	**0.8676**	0.2176	0.3341	0.0167

Table 4: Comparison of the best version of Otsu to the other decoding strategies for each dataset category using all the LLMs and the certainty and diversity metrics.

Limitations

Although our work gives a great initiative for enhanced probabilistic decoding, there are still some areas that could be explored, like testing on more records, tasks, and models. In addition, more descriptive, uncertainty metrics can be incorporated into the experiments for further detection and elimination. Finally, the lack of comparison of automated metrics to human evaluation is another limitation, which could strength the paper claims, if conducted.

References

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*.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Jennifer A Bishop, Qianqian Xie, and Sophia Ananiadou. 2023. Longdocfactscore: Evaluating the factuality of long document abstractive summarisation. *arXiv preprint arXiv:2309.12455*.

Nico Daheim, Clara Meister, Thomas Möllenhoff, and Iryna Gurevych. 2025. Uncertainty-aware decoding with minimum bayes risk. *arXiv preprint arXiv:2503.05318*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Wataru Hashimoto, Hidetaka Kamigaito, and Taro Watanabe. 2025. Decoding uncertainty: The impact of decoding strategies for uncertainty estimation in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14601–14613.

Kaifeng He, Mingwei Liu, Chong Wang, Zike Li, Yanlin Wang, Xin Peng, and Zibin Zheng. 2025. Adadec: Uncertainty-guided adaptive decoding for llm-based code generation. *arXiv preprint arXiv:2506.08980*.

John Hewitt, Christopher D Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. *arXiv preprint arXiv:2210.15191*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Hakyung Lee, Subeen Park, Joowang Kim, Sungjun Lim, and Kyungwoo Song. 2025. Uncertainty-aware contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26376–26391.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Nobuyuki Otsu and 1 others. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Chenze Shao, Xilin Chen, and Yang Feng. 2018. Greedy search with probabilistic n-gram matching for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4778–4784.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of llms. *arXiv preprint arXiv:2402.06925*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dennis Ulmer, Chrysoula Zerva, and André FT Martins. 2024. Non-exchangeable conformal language generation with nearest neighbors. *arXiv preprint arXiv:2402.00707*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, and 16 others. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. *arXiv preprint arXiv:2010.07882*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.