

From Pixels to Tokens: Adaptive LLM Decoding and Reliability Analysis with Otsu Thresholding

Eleni Partalidou^{1,2}, Tatiana Passali¹, Chrysoula Zerva^{3,4,5},
Grigorios Tsoumakas^{1,2}, Sophia Ananiadou^{2,6,7}

¹Aristotle University of Thessaloniki, ²Archimedes, Athena Research Center, Greece,
³Instituto Superior Técnico, Universidade de Lisboa, ⁴Instituto de Telecomunicações, Portugal,
⁵ELLIS Unit Lisbon, ⁶The University of Manchester ⁷ELLIS Unit Manchester

Correspondence: epartala@csd.auth.gr

Abstract

This paper investigates the relationship between generation-time uncertainty and output reliability in probabilistic Large Language Model (LLM) decoding. Specifically, we assess the correlation between several factuality, quality and certainty metrics to infer existing patterns across state-of-the-art decoding strategies, diverse tasks and instruction-tuned models. We also propose a novel, image processing-inspired decoding approach, Otsu Thresholding, that performs adaptive, parameter-free threshold selection directly from the next-token predictive confidence distribution. The proposed Otsu-based variants achieve competitive quality and factuality across tasks and models, while exhibiting concentrated predictive distributions comparable to strong truncation-based baselines. Furthermore, we employ a selective filtering setup to analyze the impact of removing highly uncertain instances on the quality and factuality metrics, ultimately revealing that filtering based on averaged token-level uncertainty metrics can be practically useful for identifying less reliable generations, with filtering frequently improving the quality or factuality of the retained outputs.

1 Introduction

Large Language Models (LLMs) have shown great potential in generative tasks, such as text summarization, story generation, and question-answering (Min et al., 2025; Dzhubaeva et al., 2025; Hu et al., 2025). According to the latest HuggingFace leaderboard¹, they consistently generate high-quality content. Despite their strong performance, a common problem of LLMs is the overestimation of their confidence (“they don’t always know what they don’t know”), leading to outputs with hallucinations or misleading information (Baan et al., 2023; Hu et al., 2023). A reliable LLM should be able to generate

text that is both accurate and confident, while providing a principled estimate of uncertainty when knowledge is insufficient, and thus the output is prone to errors or inconsistencies.

Decoding strategies can play a key role towards the aforementioned goal (Shi et al., 2024) and several strategies have been proposed that account for token probabilities under varying assumptions, in order to control the token prediction set (Fan et al., 2018; Holtzman et al., 2020; Hewitt et al., 2022; Ravfogel et al., 2023).

Beyond determining which tokens are considered during generation, the predictive distribution at each decoding step also provides a direct view of the model’s uncertainty. Token probabilities and related uncertainty measures, such as predictive entropy, can therefore potentially provide useful signals about the reliability of the generated output. However, it remains unclear to what extent such readily available token-level signals, when aggregated across a generated sequence, reflect the actual quality and factuality of the resulting text. Establishing this connection is important both for understanding the behaviour of different decoding strategies and for identifying uncertain generations that may warrant abstention or additional verification.

At the same time, most decoding methods rely on fixed, predefined hyperparameters, like the value of k in top- k sampling or the probability threshold p in nucleus sampling. These fixed truncation rules are applied uniformly across contexts and cannot adapt to changes in the shape or uncertainty of the predictive distribution. This motivates the development of fast, adaptive, and uncertainty-aware decoding strategies that dynamically determine token selection thresholds, based on the structure of the predictive distribution. Figure 1 illustrates how different sampling-based decoding methods select candidate tokens under different predictive distributions.

To address this, we frame token filtering

¹<https://huggingface.co/open-llm-leaderboard>

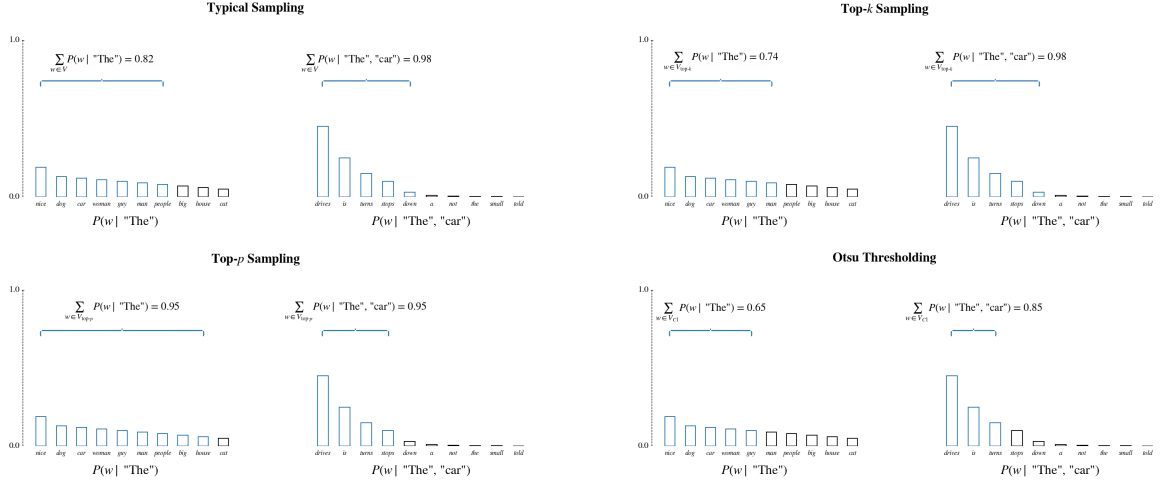


Figure 1: Examples of candidate selection from sampling-based decoding methods (generated using: <https://huggingface.co/blog/how-to-generate>).

as an *adaptive thresholding problem*. Instead of fixed hyperparameters, we apply a classic image processing technique, Otsu Thresholding (Otsu, 1979), as a parameter-free mechanism for dynamically determining the token-selection threshold at each decoding step. The motivation lies in that despite the typical long-tailed LLM predictive distributions, probability mass is usually concentrated in a small subset of candidate tokens before dropping into a low-probability tail, implying a natural boundary between low- and high-confidence tokens. To the best of our knowledge, this is the first application of Otsu Thresholding to token filtering in LLM decoding.

We evaluate the proposed approach alongside widely used sampling-based decoding strategies across multiple tasks and instruction-tuned LLMs. Crucially, our evaluation considers not only conventional measures of output quality and factuality, but also uncertainty signals derived during generation. We investigate whether token-level certainty and uncertainty measures, aggregated over the generated sequence, are systematically associated with output quality and factuality across models, tasks, and decoding strategies. Finally, we examine their practical utility through selective filtering, asking whether filtering the most uncertain generations can consistently improve the quality and factuality of the retained outputs. The main contributions of the paper are outlined as follows:

- We systematically investigate the relationship between generation-time uncertainty and output quality and factuality, across multiple tasks, instruction-tuned LLMs, and decoding

strategies, using complementary evaluation metrics and correlation analysis.

- We introduce Otsu Thresholding as a novel, parameter-free, adaptive decoding strategy that determines token-selection thresholds directly from the predictive distribution, and show that it achieves competitive performance across tasks and models.
- We assess the practical utility of decoding-time uncertainty for selective filtering, showing how removing the most uncertain generations affects the quality and factuality of the retained outputs.

2 Related Work

Evaluating open-ended language generation requires **balancing output quality, factuality, and certainty**. Traditional quality metrics primarily capture lexical or semantic similarity to reference outputs, but high-quality or fluent generations may still contain unsupported or factually incorrect information (Fabbri et al., 2022). Uncertainty estimation therefore provides a complementary perspective on generation reliability, with prior work examining uncertainty at the token, sequence, and semantic levels (Baan et al., 2023; Hu et al., 2023).

Sampling-based decoding methods control the candidate set from which the next token is generated. Top- k sampling restricts generation to the k most probable tokens (Fan et al., 2018), while nucleus sampling selects the smallest set whose cumulative probability exceeds a threshold p (Holtzman et al., 2020). Epsilon sampling removes tokens below a fixed probability cutoff (Hewitt et al.,

2022), and conformal nucleus sampling calibrates the candidate set according to the predictive distribution (Ravfogel et al., 2023). These approaches show that candidate-set selection substantially affects generation behaviour, but generally rely on predefined or calibrated truncation rules.

Factuality evaluation is an ongoing challenge. Zhu et al. (2023) introduced DIASUMFACT, for fine-grained factual error detection at the sentence-level for dialogue summarization. Zha et al. (2023) introduced AlignScore, which evaluates factual consistency, measuring information alignment between generated text and its evidence. Tang et al. (2024) introduced MiniCheck as an efficient LLM-based factuality evaluator, while Wan et al. (2025) proposed FaStFact for scalable long-form factuality evaluation.

A growing body of work has examined whether **generation-time uncertainty** can provide a useful signal of output correctness or quality. Xu et al. (2020) studied predictive entropy in summarization and linked generation uncertainty to token-level behaviour, including whether models copy or generate tokens. Moving toward output-level reliability, Kuhn et al. (2023) introduced semantic entropy, which groups semantically equivalent generations and was shown to predict question-answering accuracy. More directly related to the relationship between **uncertainty and generation reliability**, Huang et al. (2024) proposed rank-calibration, based on the principle that higher uncertainty should correspond, on average, to lower generation quality. Probability-based signals have likewise been studied as indicators of factual reliability. Azaria and Mitchell (2023) found that sentence probability is related to truthfulness, although it is also affected by factors such as sequence length and token frequency. Mahaut et al. (2024) showed that factual-confidence estimators vary substantially in reliability across models and semantically equivalent inputs, while Fadeeva et al. (2024) proposed *Claim Conditioned Probability* for estimating uncertainty at the claim level. Similarly, Cha and Lee (2024) showed that faithful and unfaithful generations can exhibit distinguishable probability and uncertainty distributions.

Uncertainty is also affected by the decoding procedure itself. Hashimoto et al. (2025) showed that decoding strategies influence both generation quality and uncertainty estimation, while Fastowski et al. (2025) studied factual robustness through token-distribution entropy and sensitivity to decod-

ing temperature. Together, these findings suggest that generation-time uncertainty can be informative, but its reliability depends on the metric, task, model, and decoding strategy.

Building on this work, we systematically examine whether simple uncertainty signals available directly during decoding are associated with both output quality and factuality across multiple tasks, LLMs, and sampling-based decoding strategies. We further evaluate their practical utility through uncertainty-based selective filtering and introduce Otsu Thresholding as an adaptive, parameter-free method for determining the candidate token set from the predictive distribution.

3 Methodology

3.1 Quality, Factuality, and Uncertainty Analysis

We evaluate generated outputs along three complementary dimensions: *quality*, *factuality*, and *uncertainty*. Quality measures how closely a generated output matches the corresponding reference, while factuality measures whether the information expressed in the generation is supported by the available reference or evidence. The specific metrics used for both dimensions are described in Section 4.4.

In parallel, we extract uncertainty information directly from the model’s predictive distribution during generation. Let $\mathbf{p}_t = p(\cdot \mid y_{<t}, x)$ denote the predictive distribution over the vocabulary at decoding step t , and let y_t denote the generated token. We use the probability assigned to the generated token, $p_t(y_t)$, as a token-level certainty signal. We additionally compute the predictive entropy of the full token distribution,

$$H_t = - \sum_{v \in V} p_t(v) \log p_t(v), \quad (1)$$

which captures the uncertainty of the model’s next-token prediction at decoding step t .

For a generated sequence y of length T , these token-level quantities are aggregated by averaging across decoding steps:

$$C(y) = \frac{1}{T} \sum_{t=1}^T p_t(y_t), \quad H(y) = \frac{1}{T} \sum_{t=1}^T H_t. \quad (2)$$

Here, $C(y)$ represents sequence-level certainty, while $H(y)$ represents the average predictive un-

certainty encountered during generation. We additionally define probability-based uncertainty as $U(y) = 1 - C(y)$, such that larger values indicate greater uncertainty.

We further consider semantic entropy (Kuhn et al., 2023) as a sequence-level uncertainty measure. For each input, we sample five candidate generations and group semantically equivalent responses into clusters representing distinct meanings. The probability mass of each cluster is obtained by aggregating the sequence-level probabilities of its members, and semantic entropy is computed over the resulting distribution:

$$H_{\text{sem}}(x) = - \sum_{c \in \mathcal{C}(x)} P(c | x) \log P(c | x), \quad (3)$$

where $\mathcal{C}(x)$ is the set of semantic clusters for input x . This captures uncertainty over distinct meanings rather than surface-level variations of the same response.

Finally, we record the average number of tokens retained as candidates during decoding,

$$W(y) = \frac{1}{T} \sum_{t=1}^T |\mathcal{S}_t|, \quad (4)$$

where $\mathcal{S}_t$ denotes the candidate set available for sampling at step t . We treat $W(y)$ as a *candidate-set statistic*, rather than a direct uncertainty measure, since its value is jointly determined by the shape of the predictive distribution and the decoding strategy.

Correlation analysis. To assess whether uncertainty observed during generation is informative about the reliability of the resulting text, we compute Pearson correlations between the uncertainty-related measures and the corresponding quality and factuality scores. Correlations are computed across examples for each combination of dataset, model, and decoding strategy. Higher sequence-level certainty is expected to correlate positively with quality and factuality, whereas higher predictive or semantic entropy is expected to correlate negatively with these measures. We additionally analyze the relationship between candidate-set size and output performance to investigate whether the breadth of the sampling space provides useful information about generation reliability.

Selective filtering. We further evaluate whether generation-time uncertainty can be used to identify less reliable outputs through selective filtering. For each experimental setting, generations are ranked by their average predictive entropy $H(y)$. We progressively remove the 5% to 50% most uncertain generations and recompute the quality and factuality metrics over the retained outputs. This analysis tests not only whether uncertainty correlates with generation performance, but also whether predictive entropy can serve as a practical signal for selectively retaining outputs with higher quality and factual reliability.

Eventually, both the uncertainty analysis and our proposed decoding strategy operate on the same core object: the next-token predictive distribution p_t . While this section measures the certainty of the distribution post-hoc to evaluate reliability, Otsu Thresholding leverages its variance structure dynamically at each step to separate informative candidate tokens from low-probability tail noise.

3.2 Otsu Thresholding

Otsu’s method, a classic image-thresholding method, partitions a normalized histogram of the pixel intensities into two classes, background (C_0) and foreground (C_1), by determining a single intensity threshold t that maximizes inter-class variance, thereby maximizing the separation between the two classes.

Let L denote the number of histogram bins and $P(i)$ the probability associated with the i -th intensity level, such that:

$$\sum_{i=0}^{L-1} P(i) = 1. \quad (5)$$

For a candidate threshold t , the two classes are $C_0 = \{0, \dots, t-1\}$ and $C_1 = \{t, \dots, L-1\}$, with corresponding probability masses

$$\omega_0(t) = \sum_{i=0}^{t-1} P(i), \quad \omega_1(t) = \sum_{i=t}^{L-1} P(i), \quad (6)$$

and class means

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} iP(i)}{\omega_0(t)}, \quad \mu_1(t) = \frac{\sum_{i=t}^{L-1} iP(i)}{\omega_1(t)}. \quad (7)$$

For every candidate threshold t , Otsu’s method computes the between-class variance as:

$$\sigma_b^2(t) = \omega_0(t)\omega_1(t) (\mu_0(t) - \mu_1(t))^2. \quad (8)$$

The optimal threshold, t^* , can thus be obtained by exhaustively evaluating all candidate thresholds and selecting

$$t^* = \arg \max_{0 < t < L} \sigma_b^2(t). \quad (9)$$

Thus, Otsu thresholding determines the partition directly from the distribution rather than relying on a manually specified cutoff.

3.3 Otsu Thresholding for Text Generation

We adapt Otsu Thresholding to next-token prediction by treating candidate selection as an adaptive thresholding problem. At each decoding step, probability mass is typically concentrated on a relatively small subset of the candidates from the vocabulary, while most tokens receive substantially lower probabilities, suggesting a natural separation. Rather than imposing a manually selected cutoff, we leverage Otsu Thresholding to identify this separation directly from the probability distribution itself and account for context- and model-dependent distributional differences. Hence, the decoding process can be interpreted as a segmentation problem, where the objective is to distinguish informative candidate tokens from background noise in the vocabulary distribution.

The probability distribution at each decoding step is treated analogously to the intensity distribution used in image thresholding, upon which Otsu Thresholding is applied to identify an adaptive boundary, separating low-probability tokens (C_{low}) from high-probability tokens (C_{high}).

More specifically, the proposed Otsu-based decoding strategy operates as follows: at each decoding step s , we obtain the logits over the vocabulary V from the language model and convert them into token probabilities, using the softmax function:

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{|V|} e^{z_j}}, \quad (10)$$

where z_i is the logit associated with token v_i and $|V|$ denotes the vocabulary size. Given the probability distribution over the vocabulary at decoding step s , we apply Otsu Thresholding to determine the optimal threshold t^* , which maximizes the between-class variance between low-probability and high-probability token groups. The resulting threshold defines the candidate token set:

$$C_{high} = \{v_i \in V \mid P_i \geq t^*\}, \quad (11)$$

where C_{high} denotes the subset of candidate tokens retained for sampling. The probability distribution is then renormalized over C_{high} and the next token is sampled multinomially from this reduced candidate set. Unlike fixed truncation strategies, the threshold and resulting candidate-set size adapt to the predictive distribution at each decoding step: concentrated distributions generally retain fewer candidates, while flatter distributions retain broader sets.

We evaluate two variants of the method. *Otsu-all* computes the threshold over the full vocabulary, whereas *Otsu-k* first restricts the distribution to the top k candidates before thresholding. For the latter, we use $k = 50$ in all experiments.

4 Experimental Setup

To systematically evaluate the proposed approach, our experimental setup spans four main axes: dataset tasks, language models, decoding strategies, and evaluation metrics.

4.1 Datasets

The dataset collection that we use for our experiments is a preprocessed version of Super-Natural Instructions (Wang et al., 2022), consisting of definitions, inputs, and targets for different conditional generation tasks. To evaluate the decoding methods across different generation settings, we consider four tasks spanning summarization, story generation, and question answering:

- **Amazon Review** (summarization): generate a summary of a given Amazon product review.
- **Extreme Abstract** (summarization): summarize a research abstract in less than 30 words.
- **Counterfactual** (story generation): generate a revised story ending that is consistent with a modified context.
- **Web Questions** (question answering): generate an answer to a given web question.

Dataset	Total	Validation	Test
Amazon Review	13,000	333	3,000
Extreme Abstract	10,796	333	3,000
Counterfactual	12,800	333	3,000
Web Questions	17,850	333	3,000

Table 1: Statistics of each dataset category.

Metric	Typical	Top- k	Top- p	Epsilon	Conformal	Otsu (all)	Otsu (k)
ROUGE-L	5	4.5	3.15	4	5.5	3.1	2.25
BERT-F1	4.95	3.6	2.4	3.7	5.75	3	2.05
F1-Q/A	4.6	4.6	2.8	5	5.6	3	1.2
Mean Rank (Quality)	4.85	4.2333	2.7833	4.2333	5.6167	3.0333	1.8333
HHEM	4.75	4.7	3.15	3.85	5	3.3	3.15
AlignScore	4.4	3.65	3.2	3.7	5.6	3.5	3.8
MiniCheck	4.65	4.2	3.45	4.2	6	2.2	3.2
Mean Rank (Factuality)	4.6	4.1833	3.2667	3.9167	5.5333	3	3.3833
Certainty	5.65	5.25	1.8	4.4	6.65	2.05	2.05
Entropy	5.95	5.2	1.8	4.1	6.6	1.9	2.2
Semantic Entropy	4.45	4.6	2.7647	4.4737	4.9	3.1053	2.7895
Candidate-set Size	6.8	5	1.3	4	6.2	2.3	2.4
Mean Rank (Uncertainty-related)	5.7125	5.0125	1.9162	4.2434	6.0875	2.3388	2.3599

Table 2: Average rank of the decoding methods across all the datasets and LLMs for each evaluation type.

Table 1 presents the number of records each dataset category contains and the number of records that were used for validation and inference.

4.2 Large Language Models

We evaluate five open-access, state-of-the-art, instruction-tuned LLMs, spanning different model architectures, number of parameters, and design choices. Particularly, we use the 8B and 70B parameter versions of LLaMA-v3 (Touvron et al., 2023), the Mistral-2501² 24B version, and the 8B and 14B parameter versions of Qwen-v3 (Team et al., 2024). Additional details are presented in Appendix A.

4.3 Decoding Strategies

To identify which strategies produce accurate, well-calibrated outputs across quality, factuality, and certainty, we evaluate Otsu Thresholding with five additional methods: **typical multinomial sampling**, **top- k sampling**, **top- p (nucleus) sampling**, **epsilon (ϵ) sampling**, and **conformal nucleus sampling**. For the calibration step in conformal sampling, we use the validation set for each dataset.

Unless otherwise specified, generation uses a temperature of 0.7 and five sampled sequences per input. We set $k = 50$ for top- k sampling, $p = 0.80$ for top- p sampling, and $\epsilon = 9 \times 10^{-4}$ for epsilon sampling. For conformal nucleus sampling, the calibration step uses the validation subset of each task. Full generation and computational settings are provided in Appendix C.

²<https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501>

4.4 Evaluation Metrics

We evaluate generated outputs using complementary measures of *quality* and *factuality*, together with the uncertainty measures and candidate-set statistic defined in Section 3.1.

For **quality**, we use ROUGE-L that measures the longest common subsequence between the two texts (Lin, 2004) and the BERTScore F1 that leverages pre-trained contextual embeddings to match tokens in candidate and reference text, using cosine similarity (Zhang et al., 2020). For the Web Questions, we also report token-level F1 against the reference answer. For **factuality**, we use HHEM (Bao et al., 2024), AlignScore (Zha et al., 2023), and MiniCheck (Tang et al., 2024). These metrics provide complementary estimates of whether the generated content is supported by the corresponding reference or evidence.

For **uncertainty**, we report sequence-level certainty $C(y)$, average predictive entropy $H(y)$, and semantic entropy $H_{\text{sem}}(x)$, which is computed, according to the work of Sen et al. (2025). Average predictive entropy is additionally used as the ranking criterion in the selective-filtering experiments. We report average candidate-set size $W(y)$ as a candidate-set statistic.

5 Results

Below, we present our experimental results, which reflect baseline hyperparameter selection and not an optimization from extensive tuning.

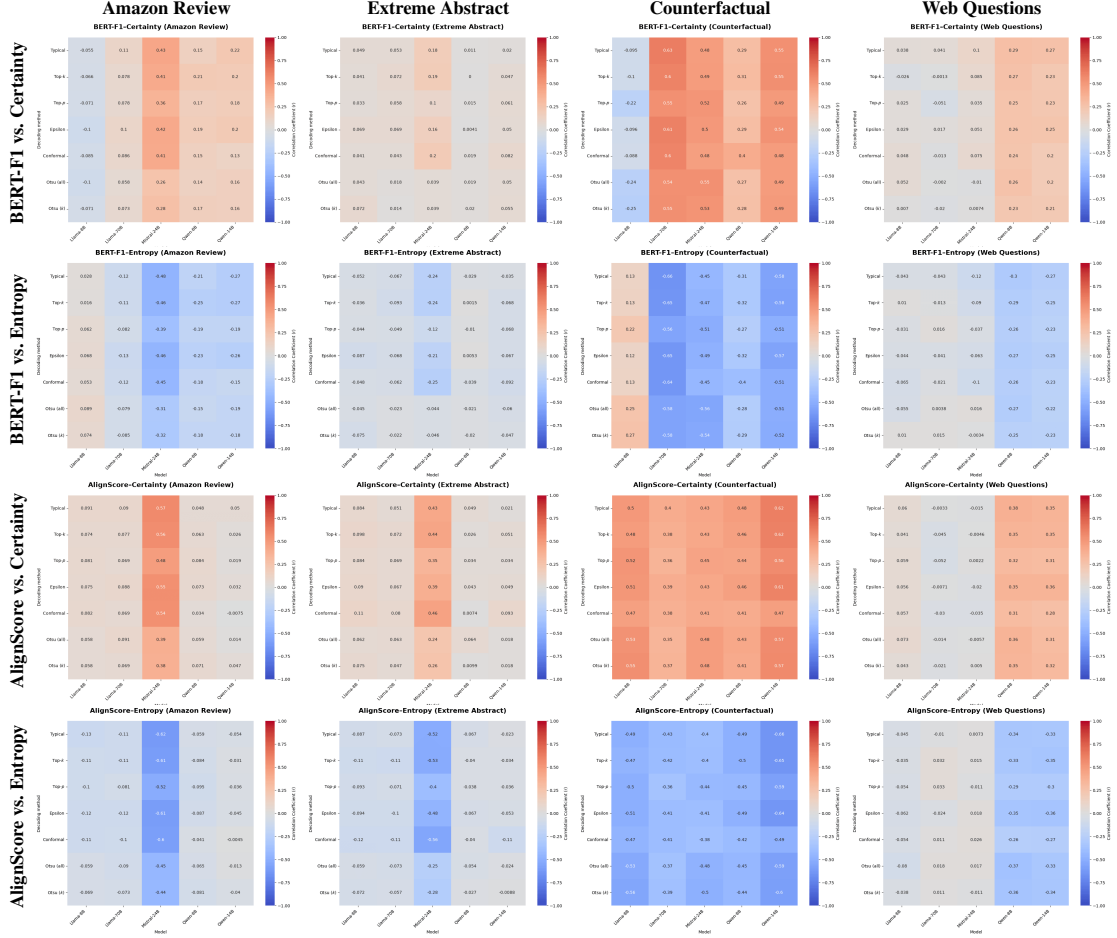


Figure 2: Correlations between generation-time uncertainty and output reliability across tasks, models, and decoding strategies. We show BERT-F1 and AlignScore as representative quality and factuality metrics, respectively, together with generated-token certainty and predictive entropy.

5.1 Analysis of Otsu Thresholding and Other Decoding Strategies

Table 2 summarizes the average rank of each decoding strategy across datasets and models. For generation quality, Otsu- k achieves the best overall mean rank (1.83), followed by top- p (2.78) and Otsu-all (3.03). For factuality, Otsu-all ranks first overall (3.00), followed by top- p (3.27).

The uncertainty-related measures provide a complementary view of generation behaviour. Top- p produces the most confident overall profile, ranking first in generated-token certainty, predictive entropy, and candidate-set size. Both Otsu variants follow closely on token-level certainty and predictive entropy, indicating similarly concentrated predictive distributions, while showing somewhat greater variation in semantic entropy.

Overall, Otsu-based decoding achieves a favorable balance between output performance and generation confidence: Otsu- k is strongest on quality,

while Otsu-all ranks best on factuality, with both exhibiting confidence profiles close to top- p . Importantly, higher confidence does not necessarily imply higher quality or factuality; we examine this relationship directly in Section 5.2.

5.2 Correlations Between Different Types of Metrics

We next examine whether uncertainty observed during decoding provides a reliable indication of the quality and factuality of the resulting text. We thus computed the Pearson correlations between the quality-certainty and factuality-certainty metrics. Rather than aggregating correlations across models and decoding strategies, Figure 2 presents the correlations separately for each model-strategy combination and task. For readability, we show BERT-F1- and AlignScore as representative quality and factuality metrics, together with generated-token certainty and predictive entropy. Results for the remaining metrics are provided in Appendix E.

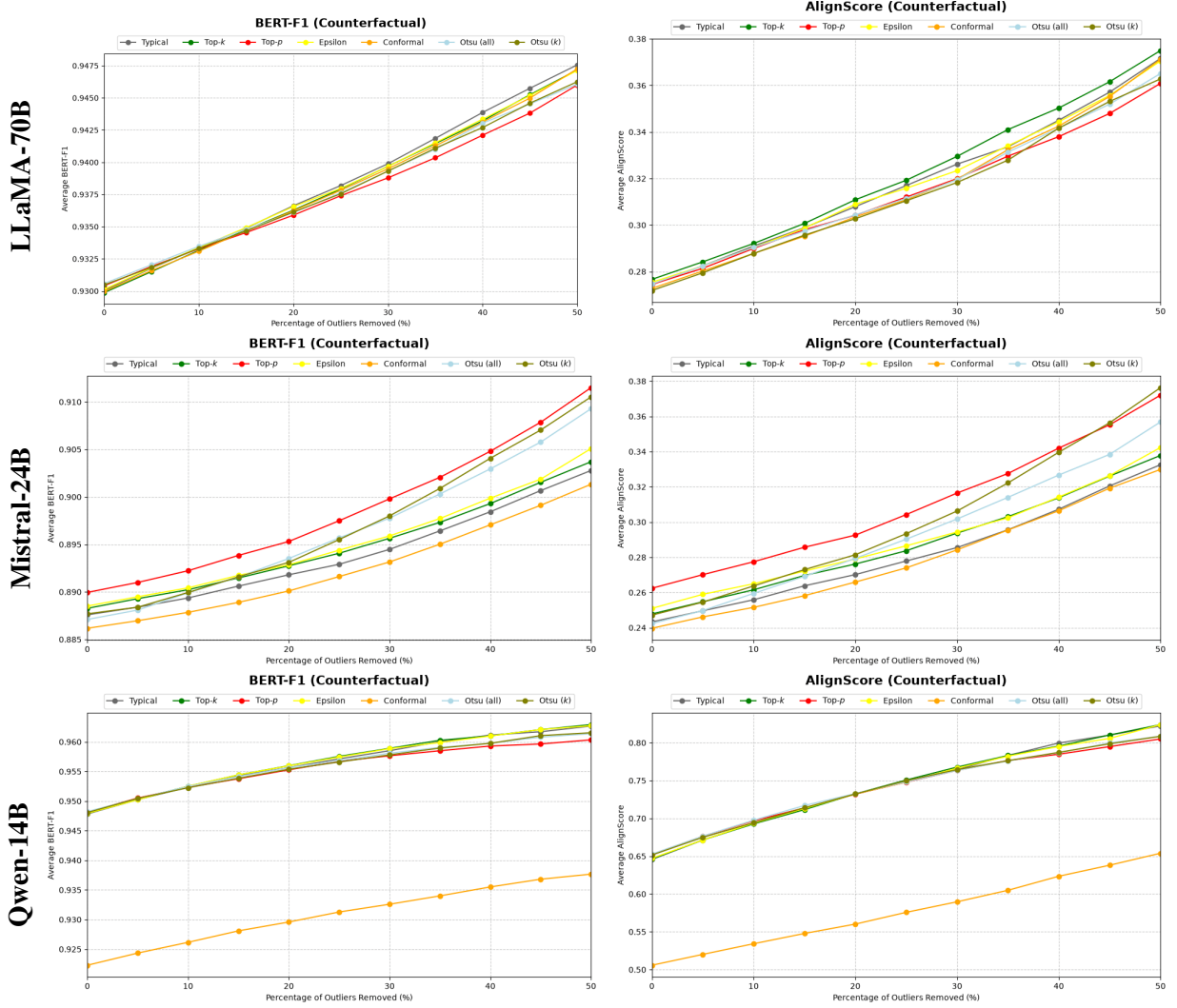


Figure 3: Effect of predictive-entropy-based selective filtering on the Counterfactual task for larger models from three model families. Increasing fractions of the generations with highest average predictive entropy are removed, and BERT-F1 (quality) and AlignScore (factuality) are recomputed over the retained outputs.

Overall, the expected directional relationship is visible across many settings: higher generated-token certainty tends to be associated with higher quality and factuality, while higher predictive entropy tends to show the opposite relationship. However, the strength of this association varies substantially across tasks, models, and decoding strategies. In particular, the Counterfactual task exhibits considerably stronger relationships between uncertainty and output performance, whereas the correlations for Extreme Abstract are generally weaker. Considering the specifics of each task, it seems that one of the factors impacting the strength of the correlations is the average length of the expected output and potentially the degree of creative freedom implied in the expected output, although the latter would merit further analysis to be confirmed.

The heatmaps further show that this variation is not explained by the task alone: different models and decoding strategies can exhibit noticeably different uncertainty–performance relationships within the same task. These results suggest that generation-time uncertainty provides a useful signal of output reliability, but its relationship with correctness remains model- and decoding-dependent.

5.3 Selective Filtering

Furthermore, we examined whether generation-time uncertainty can be used to identify less reliable outputs. Following Section 3.1, we progressively remove generations with the highest average predictive entropy and recompute the quality and factuality metrics over the retained set.

Figure 3 shows representative results on the *Counterfactual* task for larger models from three model families. Overall, uncertainty provides a useful reliability signal: factuality generally improves as increasingly uncertain generations are removed, while the effect on reference-based quality is smaller and more model-dependent. This suggests that uncertainty is useful for selective filtering, although it should not be interpreted as a direct proxy for correctness, since especially for smaller models it seems less consistent. Additional results across all tasks and models are reported in Appendix F.

6 Conclusions

We investigated the relationship between generation-time uncertainty and output reliability across multiple tasks, models, and sampling-based decoding strategies, while proposing *Otsu Thresholding* as a parameter-free adaptive alternative, which are mutual ideas for more reliable text generation. Our results show that Otsu-based decoding achieves competitive quality and factuality, with confidence profiles comparable to strong baselines such as top- p sampling. At the same time, the relationship between uncertainty and downstream performance is clearly task-, model-, and metric-dependent, suggesting that uncertainty should not be interpreted as a direct proxy for correctness. Nevertheless, the selective-filtering experiments show that simple uncertainty signals available during generation can be practically useful for identifying less reliable outputs and improving the quality or factuality of the retained generations.

Limitations

Our evaluation is limited to four generation tasks and five instruction-tuned LLMs. Although these settings cover summarization, story generation, and question answering, broader experiments across additional datasets, model families, and parameter scales would be needed to establish the generalisability of the observed uncertainty–quality relationships. Along the same lines, it would be important to also expand to other languages, to assess whether observed patterns hold in non English-centric settings.

Our analysis also focuses on a relatively small and simple set of uncertainty signals, including generated-token probability, predictive entropy, se-

mantic entropy, and candidate-set size. Future work could consider richer sequence-level, semantic, claim-level, or learned uncertainty estimators.

Finally, our conclusions rely primarily on automatic quality and factuality metrics. While multiple complementary metrics are used, human evaluation would provide stronger evidence that the observed improvements correspond to genuinely better and more reliable outputs. Moreover, the selective filtering results are not uniform across all settings, reinforcing that generation-time uncertainty should be treated as a useful reliability signal rather than a universal proxy for correctness.

Acknowledgments

Mrs Partalidou acknowledges the financial support of the Archimedes Research Unit (Athena RC) during her PhD. Also, GCP resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility. Furthermore, this research used resources of the Oak Ridge Leadership Computing Facility (OLCF) and Argonne Leadership Computing Facility (ALCF), which are a DOE Office of Science User Facility. This work was supported by an award from the ASCR Leadership Computing Challenge (ALCC) under project ERCAP0034861. Moreover, this allocation was secured through a research collaboration with Oumi, who provided compute access in support of this work. In addition, this research was partially funded by an unrestricted gift from Google, by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), by Fundação para a Ciência e Tecnologia (FCT) through 'OptiGov' (DOI 10.54499/2024.07385.IACDC), funded by the PRR under the measure RE-C05-i08.m04, by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and also by FCT/MECI through national funds and, when applicable, cofunded EU initiatives under contract UID/50008 for Instituto de Telecomunicações. Lastly, Sophia Ananiadou acknowledges support from the British Heart Foundation Centre of Research Excellence, University of Manchester (Award Code: RC000797).

References

Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it's lying](#). In *Find-*

- ings 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*.
- Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. *HHEM-2.1-Open*.
- Taehun Cha and Donghun Lee. 2024. [Pre-trained language models return distinguishable probability distributions to unfaithfully hallucinated texts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12630–12639, Miami, Florida, USA. Association for Computational Linguistics.
- Nellia Dzhubaeva, Katharina Trinley, and Laura Pissani. 2025. Unstructured minds, predictable machines: A comparative study of narrative cohesion in human and llm stream-of-consciousness writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1079–1096.
- Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2587–2601.
- 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.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- Alina Fastowski, Bardh Prenkaj, and Gjergji Kasneci. 2025. From confidence to collapse in llm factual robustness. In *EMNLP (Findings)*, pages 8650–8667.
- 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, Suzhou, China. Association for Computational Linguistics.
- John Hewitt, Christopher Manning, and Percy Liang. 2022. [Truncation sampling as language model desmoothing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation](#). In *International Conference on Learning Representations*.
- 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*.
- Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru Wang, Sheng Bi, Yongrui Chen, Tongtong Wu, and Jeff Z Pan. 2025. Can llms evaluate complex attribution in qa? automatic benchmarking using knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17096–17118.
- Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. [Uncertainty in language models: Assessment through rank-calibration](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, Miami, Florida, USA. 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*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Matéo Mahaut, Laura Aina, Paula Czarrowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4570.
- Hyangsuk Min, Yuho Lee, Minjeong Ban, Jiaqi Deng, Nicole Hee-Yeon Kim, Taewon Yun, Hang Su, Jason Cai, and Hwanjun Song. 2025. Towards multi-dimensional evaluation of llm summarization across domains and languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14417–14450.
- Nobuyuki Otsu. 1979. [A threshold selection method from gray-level histograms](#). *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

- Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. In *Findings of the association for computational linguistics: ACL 2023*, pages 27–34.
- Tan Min Sen, Zachary Choy Kit Chun, Swagat Bikash Saikia, Syed Ali Redha Alsagoff, Banerjee Mohor, Nadya Yuki Wangsajaya, and Alvin Chan. 2025. Automating evaluation of creativity in llms with semantic entropy and efficient multi-agent judge. In *Workshop on Reasoning and Planning for Large Language Models*.
- 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](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.
- 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*.
- Yingjia Wan, Haochen Tan, Xiao Zhu, Xinyu Zhou, Zhiwei Li, Qingsong Lv, Changxuan Sun, Jiaqi Zeng, Yi Xu, Jianqiao Lu, and 1 others. 2025. Fastfact: Faster, stronger long-form factuality evaluations in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23814–23854.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Gianis 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281.
- 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.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTscore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.
- Rongxin Zhu, Jianzhong Qi, and Jey Han Lau. 2023. Annotating and detecting fine-grained factual errors for dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6825–6845.

A Large Language Models

LLaMA-v3 uses a dense, decoder-only transformer architecture, designed to balance reasoning capabilities with inference speed. The models support a context length of 128,000 tokens and utilize Grouped-Query Attention (GQA) and Rotary Positional Embeddings (RoPE) to remain efficient. Moreover, they are trained on over 15 trillion tokens and refined through extensive Supervised Fine-Tuning and Reinforcement Learning with Human Feedback to ensure they are safe, helpful, and highly capable in multilingual environments.

Mistral-2501 is built as a high-efficiency model, packing 24 billion parameters into an architecture, optimized for low latency and high throughput. It supports a 128,000 context window and is specifically designed for enterprise applications, such as tool calling and extracting structured data. By prioritizing a dense, knowledge-heavy design, it achieves high reasoning performance in a compact footprint, making it a versatile middle-ground for tasks that require more power than smaller models but more speed than massive 70B systems.

Qwen-v3 offers a range of parameter counts (from 0.6B to 235B) that are known for exceptional performance in coding and mathematics. These models support a native context length of 128,000 tokens with specific versions, capable of handling up to 1 million tokens. The architecture relies on an advanced transformer design with SwiGLU activations and GQA. A key training insight is their use of a progressive context-scaling technique, which allows them to maintain high precision in short tasks, while excelling at long-context information retrieval.

B Instructions of the datasets

The instructions that the datasets 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.
- **Counterfactual:** Given a premise, an initial context, an original ending, and a counterfactual context, the task is to generate a new story ending aligned with the counterfactual context and as close to the original ending as possible. Each instance consists of a five-sentence story. The premise is the first sentence of a story, and the second sentence, which is the initial context, provides more information about the story’s context and the story’s general plot. The original ending is the last three sentences of the story. Also, a counterfactual context is a slight modification to the initial context. You should write a new story ending that edits the original story ending as little as possible to regain coherence with the counterfactual context. To sum up, you should write the last three sentences of a story based on the premise(first sentence) and the counterfactual context(second sentence) of the story.
- **Web Questions:** A question is presented to you in this task, and your job is to write a potentially correct answer.

C Hyperparameter Settings

Our experiments were conducted on two computing environments: NERSC Perlmutter cluster and Google Cloud Platform (GCP). The Perlmutter nodes were configured with 250 GB of system memory and four NVIDIA A100 (80 GB) GPUs. The GCP VM instance featured 208 vCPUs, 1,872 GB of RAM, a 200 GB boot disk, and eight NVIDIA H100 (80 GB) GPUs. For the text generation configuration, we set the number of generated sequences to 5, the temperature to 0.7, the epsilon cutoff to $9e^{-4}$, top- k to 50, and top- p to 0.80, where applicable. These hyperparameters

were selected to balance computational efficiency with reliable estimation and to match standard values in recent work.

D Analysis of Otsu Thresholding and Other Decoding Strategies

This section provides the complete results underlying the aggregate decoding-strategy comparison in Section 5.1. Tables 3–5 report the quality, factuality, uncertainty, and candidate-set statistics for each combination of task, model, and decoding strategy. Specifically, we report results for the two summarization tasks, Amazon Review and Extreme Abstract, the Counterfactual story-generation task, and the Web Questions question-answering task.

These detailed results complement the average ranks reported in Table 2 by showing how the relative behaviour of the decoding methods varies across individual models and tasks. The $\uparrow$ and $\downarrow$ symbols indicate the preferred direction for each metric.

Across the individual settings, Otsu-based decoding remains competitive, although the relative advantage of Otsu-all and Otsu- k varies across tasks and evaluation metrics. Otsu- k frequently achieves among the strongest quality scores, while Otsu-all is particularly competitive on factuality measures. Both variants also consistently produce concentrated predictive distributions, with candidate-set sizes and uncertainty profiles close to top- p , despite determining the filtering threshold adaptively rather than through a fixed probability cutoff. These results support the aggregate trends reported in Table 2, while highlighting the model- and task-dependent differences hidden by average rankings.

E Correlations Between Different Types of Metrics

This section provides the complete correlation results underlying the representative analysis reported in Section 5.2. Figures 4–6 show Pearson correlations between the quality metrics and the uncertainty-related measures defined in Section 3.1, while Figures 7–9 show the respective correlations for the factuality metrics. We report ROUGE-L, BERT-F1, and F1-Q/A for generation quality, and HHEM, AlignScore, and MiniCheck for factuality.

For each evaluation metric, correlations are shown with generated-token certainty, predictive entropy, semantic entropy, and candidate-set size across all tasks. Each heatmap further separates the

Dataset	Model	Method	ROUGE-L $\uparrow$	BERT-F1 $\uparrow$	HHEM $\uparrow$	AlignScore $\uparrow$	MiniCheck $\uparrow$	Certainty $\uparrow$	Entropy $\downarrow$	Semantic Entropy $\downarrow$	Candidate-set Size $\downarrow$
Amazon Review	LLaMA-8B	Typical	0.1499	0.8482	0.2405	0.7673	0.3891	0.7387	0.5367	0.749	128256.8036
		Top- k	0.1484	0.8482	0.2371	0.7629	0.3859	0.7381	0.5344	0.7522	50.6921
		Top- p	0.1516	0.8492	0.2355	0.782	0.4012	0.823	0.2863	0.73	1.5961
		Epsilon	0.1498	0.8487	0.2411	0.766	0.3888	0.7399	0.522	0.7367	5.3993
		Conformal	0.1509	0.8484	0.2343	0.7656	0.3819	0.7373	0.5371	0.7523	128256.8064
		Otsu (all)	0.1526	0.8485	0.2388	0.7817	0.4012	0.8441	0.2513	0.7327	1.6083
		Otsu (k)	0.1542	0.8491	0.2405	0.783	0.3995	0.8444	0.2507	0.7316	1.6073
	LLaMA-70B	Typical	0.1721	0.8694	0.2302	0.86	0.4736	0.9294	0.1254	0.2986	128255.737
		Top- k	0.171	0.8694	0.2255	0.8659	0.4653	0.9297	0.1244	0.2894	50.489
		Top- p	0.1702	0.8691	0.2247	0.863	0.4716	0.9577	0.0628	-	1.1064
		Epsilon	0.1726	0.8695	0.2288	0.8662	0.4699	0.9293	0.1232	0.2969	1.6924
		Conformal	0.172	0.8694	0.2296	0.8636	0.4711	0.929	0.1253	0.2974	128255.7271
		Otsu (all)	0.1708	0.8692	0.225	0.8657	0.4736	0.9539	0.071	0.3032	1.2046
		Otsu (k)	0.1721	0.8694	0.2289	0.861	0.4732	0.9543	0.0715	0.3159	1.2045
	Mistral-24B	Typical	0.1057	0.8413	0.1561	0.3934	0.4055	0.7306	0.5928	2.1165	131071.8738
		Top- k	0.1068	0.8419	0.1617	0.4067	0.4291	0.7375	0.5646	2.0943	50.8049
		Top- p	0.0988	0.8392	0.1381	0.3319	0.4426	0.8349	0.2704	2.0691	1.6044
		Epsilon	0.1058	0.8416	0.1571	0.4006	0.4152	0.7431	0.5379	2.0958	6.5658
		Conformal	0.1059	0.8414	0.1605	0.3903	0.4138	0.7326	0.5873	2.0959	131071.8728
		Otsu (all)	0.0994	0.8391	0.1316	0.3333	0.4386	0.8598	0.2261	2.0179	1.5764
		Otsu (k)	0.1001	0.8392	0.1368	0.3403	0.4459	0.8575	0.2289	2.0235	1.5824
	Qwen-8B	Typical	0.1864	0.8688	0.3223	0.9095	0.5562	0.9288	0.1247	0.8088	151935.7518
		Top- k	0.1861	0.8688	0.3183	0.9087	0.5505	0.9291	0.1245	0.8095	50.6135
		Top- p	0.186	0.8688	0.3203	0.9065	0.5499	0.9575	0.0624	0.7984	1.1059
		Epsilon	0.1871	0.8688	0.3214	0.9102	0.553	0.9296	0.1232	0.8178	1.6805
		Conformal	0.1776	0.8624	0.3135	0.866	0.5139	0.9186	0.1462	0.8283	151480.3652
		Otsu (all)	0.1873	0.8689	0.3198	0.9124	0.5539	0.9525	0.0738	0.8019	1.2152
		Otsu (k)	0.1864	0.8689	0.317	0.908	0.5517	0.9534	0.0728	0.7979	1.2135
	Qwen-14B	Typical	0.1789	0.8681	0.3158	0.9087	0.5496	0.9421	0.1019	0.6868	151935.7778
		Top- k	0.1795	0.8683	0.3217	0.9101	0.5552	0.9419	0.1011	0.6882	50.5762
		Top- p	0.1789	0.868	0.3177	0.9085	0.5555	0.9648	0.0512	0.662	1.0861
		Epsilon	0.1796	0.8682	0.3172	0.9081	0.5572	0.9426	0.1011	0.6867	1.5446
		Conformal	0.169	0.8573	0.3103	0.871	0.5291	0.9307	0.1226	0.7156	151591.7873
		Otsu (all)	0.1789	0.8681	0.3208	0.9077	0.5509	0.9615	0.0602	0.666	1.1793
		Otsu (k)	0.1803	0.8682	0.3177	0.9076	0.5547	0.962	0.0598	0.6678	1.1774
Extreme Abstract	LLaMA-8B	Typical	0.2043	0.8808	0.3124	0.9318	0.7918	0.7672	0.4651	0.2929	128256.9519
		Top- k	0.2026	0.8809	0.3021	0.9327	0.7946	0.7673	0.4628	0.2971	50.667
		Top- p	0.2056	0.8813	0.3073	0.9402	0.807	0.8422	0.2509	-	1.5021
		Epsilon	0.2028	0.8805	0.3036	0.9346	0.8021	0.769	0.4525	0.297	4.4215
		Conformal	0.2019	0.8806	0.3103	0.9335	0.796	0.7666	0.4651	0.3005	128256.9518
		Otsu (all)	0.2037	0.8809	0.3028	0.9376	0.8069	0.8573	0.2276	0.3509	1.546
		Otsu (k)	0.2059	0.8813	0.3098	0.9397	0.8019	0.8574	0.2276	0.3691	1.5441
	LLaMA-70B	Typical	0.2022	0.8837	0.1603	0.9275	0.9552	0.9173	0.1489	0.3253	128255.8748
		Top- k	0.2026	0.8836	0.1604	0.925	0.9538	0.9169	0.1499	0.3364	50.5471
		Top- p	0.2032	0.8839	0.1611	0.9293	0.9566	0.9491	0.0762	0.3079	1.1329
		Epsilon	0.2018	0.8835	0.16	0.9284	0.9571	0.9184	0.1467	0.3444	1.8719
		Conformal	0.2033	0.8837	0.1579	0.9256	0.9547	0.9169	0.1495	0.342	128255.8735
		Otsu (all)	0.2048	0.8839	0.1616	0.9264	0.9567	0.9471	0.0828	0.3093	1.2256
		Otsu (k)	0.2045	0.8838	0.1594	0.9303	0.9585	0.9471	0.0829	0.3067	1.2255
	Mistral-24B	Typical	0.1685	0.8693	0.4467	0.8914	0.773	0.8219	0.3614	1.6972	131071.9729
		Top- k	0.1686	0.8691	0.4495	0.8929	0.7771	0.8232	0.3549	1.6971	50.7894
		Top- p	0.172	0.87	0.4541	0.9032	0.7935	0.8987	0.1571	1.6207	1.3025
		Epsilon	0.1698	0.8698	0.4521	0.9011	0.7851	0.8272	0.339	1.6941	3.8751
		Conformal	0.1687	0.8691	0.4449	0.8944	0.778	0.8238	0.3568	1.6922	131071.973
		Otsu (all)	0.1733	0.871	0.4588	0.9183	0.8051	0.9055	0.1503	1.5962	1.403
		Otsu (k)	0.1722	0.8706	0.4637	0.9162	0.8027	0.9062	0.1499	1.5918	1.4043
	Qwen-8B	Typical	0.1971	0.8798	0.3653	0.9308	0.7894	0.9363	0.1102	0.7893	151935.9537
		Top- k	0.1979	0.88	0.366	0.9306	0.791	0.9372	0.1091	0.7762	50.5983
		Top- p	0.198	0.8801	0.3737	0.9309	0.7937	0.9619	0.0555	0.752	1.0921
		Epsilon	0.198	0.8801	0.3672	0.9314	0.7932	0.937	0.1094	0.7734	1.5537
		Conformal	0.1842	0.8658	0.3671	0.8662	0.7199	0.9219	0.1367	0.6416	150771.7144
		Otsu (all)	0.1974	0.88	0.3709	0.93	0.7924	0.9571	0.0668	0.7576	1.1962
		Otsu (k)	0.1982	0.88	0.3673	0.9302	0.7907	0.957	0.0669	0.7616	1.1963
	Qwen-14B	Typical	0.2022	0.8804	0.36	0.9475	0.7742	0.9315	0.1198	0.3599	151935.952
		Top- k	0.2027	0.8805	0.3606	0.9488	0.7774	0.9319	0.1203	0.3543	50.6398
		Top- p	0.2028	0.8806	0.3631	0.949	0.7708	0.9603	0.0588	0.4119	1.0989
		Epsilon	0.2013	0.8804	0.3626	0.9474	0.7713	0.9317	0.1183	0.3576	1.6316
		Conformal	0.1767	0.8542	0.336	0.8435	0.6658	0.9204	0.1419	0.363	151171.6299
		Otsu (all)	0.2025	0.8806	0.364	0.9471	0.7754	0.9546	0.0711	0.4023	1.2112
		Otsu (k)	0.2034	0.8807	0.3637	0.951	0.7715	0.9539	0.0716	0.3945	1.2125

Table 3: Analysis of the decoding strategies for the text summarization task.

results by model and decoding strategy, allowing the variation hidden by aggregate statistics to be examined directly. These results provide the full context for the BERT-F1 and AlignScore correlations selected for presentation in Figure 2.

The complete heatmaps reinforce the directional trends observed in Section 5.2. Across many model-strategy combinations, higher generated-token certainty is positively associated with quality and factuality, while higher predictive entropy shows the corresponding negative relationship. The strength of these associations is nevertheless highly task-dependent: the Counterfactual task exhibits

particularly clear relationships across several evaluation metrics, whereas Extreme Abstract generally produces weaker correlations.

The extended results also show that the relationship depends on the uncertainty signal itself. Generated-token certainty and predictive entropy exhibit comparatively consistent directional patterns, whereas semantic entropy and candidate-set size show greater variation across models, decoding strategies, and tasks. Thus, the full analysis supports the use of generation-time uncertainty as an informative reliability signal, while also demonstrating that no individual uncertainty measure pro-

Dataset	Model	Method	ROUGE-L $\uparrow$	BERT-F1 $\uparrow$	HHEM $\uparrow$	AlignScore $\uparrow$	MiniCheck $\uparrow$	Certainty $\uparrow$	Entropy $\downarrow$	Semantic Entropy $\downarrow$	Candidate-set Size $\downarrow$
Counterfactual	LLaMA-8B	Typical	0.2892	0.8749	0.0969	0.357	0.1561	0.787	0.4653	2.3493	128256.7483
		Top- k	0.2891	0.8753	0.091	0.3562	0.1506	0.7856	0.4682	2.3519	50.7552
		Top- p	0.3113	0.8774	0.1082	0.3836	0.1571	0.8703	0.2173	2.3437	1.5007
		Epsilon	0.2921	0.8754	0.0981	0.3625	0.1508	0.7874	0.4528	2.3493	5.6383
		Conformal	0.2922	0.875	0.0973	0.3473	0.1475	0.7828	0.4747	2.3486	128256.7481
		Otsu (all)	0.3047	0.8762	0.1054	0.3973	0.1613	0.8908	0.1786	2.3141	1.4556
		Otsu (k)	0.3062	0.8763	0.1062	0.3951	0.164	0.8907	0.178	2.3005	1.4547
	LLaMA-70B	Typical	0.56	0.93	0.2036	0.2745	0.1239	0.9432	0.1052	1.7745	128255.6366
		Top- k	0.5583	0.9299	0.2031	0.2767	0.1262	0.9426	0.1056	1.7908	50.5353
		Top- p	0.5623	0.9304	0.2068	0.2745	0.1225	0.9649	0.0531	1.6596	1.0944
		Epsilon	0.5612	0.9301	0.2044	0.2754	0.1209	0.9433	0.1044	1.7845	1.6784
		Conformal	0.5591	0.93	0.2039	0.2728	0.125	0.9422	0.1064	1.7822	128255.2686
		Otsu (all)	0.5629	0.9306	0.211	0.2747	0.1289	0.9642	0.0563	1.6605	1.1583
		Otsu (k)	0.5631	0.9305	0.2046	0.2719	0.1255	0.9634	0.0575	1.67	1.1607
	Mistral-24B	Typical	0.3248	0.8876	0.0992	0.2433	0.0997	0.7116	0.6967	2.5735	131071.684
		Top- k	0.3285	0.8883	0.0897	0.2478	0.1013	0.7139	0.675	2.5687	50.8764
		Top- p	0.3435	0.8899	0.1008	0.2624	0.1176	0.8356	0.2815	2.6324	1.6946
		Epsilon	0.3283	0.8885	0.0993	0.251	0.1059	0.7239	0.6287	2.5771	9.3298
		Conformal	0.3152	0.8862	0.0824	0.2397	0.1022	0.7088	0.7008	2.5681	131071.6883
		Otsu (all)	0.3328	0.8871	0.0951	0.2423	0.1199	0.8663	0.2198	2.6158	1.5905
		Otsu (k)	0.3363	0.8877	0.101	0.2471	0.1187	0.864	0.2242	2.6126	1.6028
	Qwen-8B	Typical	0.7046	0.9443	0.2619	0.8217	0.4613	0.9867	0.0235	1.3722	151935.5737
		Top- k	0.7075	0.9449	0.2634	0.822	0.463	0.9872	0.023	1.3695	50.6955
		Top- p	0.7078	0.9449	0.262	0.8239	0.4669	0.9925	0.0112	1.3269	1.0192
		Epsilon	0.7078	0.9449	0.2645	0.8213	0.4641	0.987	0.0231	1.3704	1.1344
		Conformal	0.6282	0.9208	0.2639	0.6586	0.3534	0.9575	0.0796	1.5053	151426.0019
		Otsu (all)	0.7086	0.9449	0.2632	0.8241	0.4674	0.9917	0.0129	1.3355	1.0385
		Otsu (k)	0.7072	0.9449	0.2628	0.8238	0.4617	0.9917	0.0132	1.3351	1.0393
	Qwen-14B	Typical	0.6806	0.9478	0.2132	0.6456	0.324	0.9735	0.0483	1.4324	151935.5383
		Top- k	0.6809	0.9479	0.2152	0.6463	0.3258	0.9736	0.0483	1.4264	50.6424
		Top- p	0.6837	0.9481	0.2159	0.6523	0.33	0.9846	0.0232	1.3429	1.0407
		Epsilon	0.6812	0.9479	0.2162	0.6472	0.3255	0.9739	0.0476	1.4294	1.3033
		Conformal	0.6053	0.9223	0.2224	0.5058	0.2635	0.9472	0.0978	1.5442	151724.3278
		Otsu (all)	0.6838	0.9482	0.2181	0.6527	0.3263	0.9835	0.0261	1.353	1.0756
		Otsu (k)	0.6828	0.9481	0.2156	0.6515	0.3284	0.9837	0.0255	1.3494	1.0746

Table 4: Analysis of the decoding strategies for the story generation task.

Dataset	Model	Method	ROUGE-L $\uparrow$	BERT-F1 $\uparrow$	F1-Q/A $\uparrow$	HHEM $\uparrow$	AlignScore $\uparrow$	MiniCheck $\uparrow$	Certainty $\uparrow$	Entropy $\downarrow$	Semantic Entropy $\downarrow$	Candidate-set Size $\downarrow$
Web Questions	LLaMA-8B	Typical	0.0804	0.8078	0.1052	0.2693	0.3777	0.0874	0.8444	0.3033	1.1429	128256.9638
		Top- k	0.083	0.8083	0.1072	0.2669	0.3805	0.0949	0.844	0.3017	1.157	50.6867
		Top- p	0.0851	0.8088	0.1101	0.2731	0.3786	0.0925	0.902	0.1523	1.162	1.2941
		Epsilon	0.0816	0.8078	0.1068	0.2644	0.3776	0.0895	0.8466	0.2927	1.1428	3.1994
		Conformal	0.0812	0.8077	0.1051	0.2636	0.3784	0.0902	0.8446	0.3016	1.1545	128256.9641
		Otsu (all)	0.0828	0.8083	0.109	0.2676	0.3781	0.0939	0.9077	0.1477	1.147	1.3827
		Otsu (k)	0.0848	0.8087	0.1101	0.27	0.3786	0.0929	0.9074	0.1478	1.1481	1.3817
	LLaMA-70B	Typical	0.0709	0.8067	0.0969	0.302	0.3434	0.0721	0.952	0.0832	1.2206	128255.9724
		Top- k	0.0718	0.8071	0.0976	0.3083	0.343	0.0725	0.9518	0.0833	1.2402	50.5194
		Top- p	0.072	0.8068	0.0977	0.3078	0.3437	0.0734	0.9713	0.0417	1.1486	1.0692
		Epsilon	0.071	0.8067	0.0976	0.3066	0.3431	0.0727	0.9519	0.0824	1.2293	1.4334
		Conformal	0.073	0.8068	0.0988	0.3081	0.3437	0.0706	0.9513	0.0843	1.2388	128254.3571
		Otsu (all)	0.071	0.8068	0.0977	0.3052	0.3438	0.0757	0.9679	0.0497	1.1695	1.1428
		Otsu (k)	0.0718	0.807	0.0981	0.3065	0.3432	0.0752	0.9682	0.0497	1.1641	1.1431
	Mistral-24B	Typical	0.0557	0.8073	0.0761	0.2581	0.3287	0.0619	0.8466	0.3125	1.5687	131071.9748
		Top- k	0.0558	0.8074	0.0749	0.262	0.3302	0.0589	0.848	0.3071	1.564	50.7641
		Top- p	0.0584	0.8093	0.0792	0.2744	0.3276	0.0572	0.905	0.1488	1.4796	1.2926
		Epsilon	0.0575	0.8078	0.0786	0.2654	0.3296	0.0573	0.852	0.2913	1.5706	3.647
		Conformal	0.0559	0.8075	0.0754	0.2543	0.3281	0.0611	0.847	0.3109	1.5737	131071.9746
		Otsu (all)	0.0588	0.8091	0.0794	0.2747	0.3295	0.0588	0.9115	0.1413	1.4833	1.3823
		Otsu (k)	0.0593	0.8095	0.0802	0.2758	0.3289	0.0553	0.9117	0.1418	1.4631	1.3848
	Qwen-8B	Typical	0.0802	0.8052	0.1109	0.2774	0.4101	0.0757	0.9278	0.1383	1.0161	151935.9647
		Top- k	0.0789	0.805	0.1093	0.2753	0.4097	0.0787	0.9276	0.1376	1.0127	50.6288
		Top- p	0.08	0.8056	0.1113	0.2814	0.4095	0.0702	0.9547	0.0706	-	1.14
		Epsilon	0.0786	0.8054	0.1093	0.2766	0.4092	0.0737	0.9273	0.1362	-	2.0263
		Conformal	0.077	0.8026	0.1054	0.274	0.4108	0.0682	0.9207	0.1525	1.0146	151828.5611
		Otsu (all)	0.0797	0.8053	0.1109	0.2797	0.4089	0.0741	0.9538	0.0734	-	1.2074
		Otsu (k)	0.0799	0.8055	0.1114	0.2794	0.409	0.0774	0.9537	0.0738	-	1.2084
	Qwen-14B	Typical	0.0843	0.8068	0.1163	0.3007	0.4111	0.0819	0.9358	0.121	0.9527	151935.9649
		Top- k	0.0849	0.8069	0.1163	0.3017	0.4119	0.08	0.9359	0.1191	0.9747	50.5804
		Top- p	0.084	0.8068	0.1161	0.3027	0.4103	0.0776	0.9596	0.0619	0.9262	1.1176
		Epsilon	0.0847	0.8065	0.1152	0.3008	0.4102	0.0787	0.9367	0.1165	0.9725	1.8001
		Conformal	0.0809	0.8014	0.1059	0.3026	0.4043	0.0703	0.9276	0.1362	0.971	151780.6482
		Otsu (all)	0.0844	0.8068	0.1162	0.3029	0.4107	0.0813	0.9584	0.0652	0.9151	1.1819
		Otsu (k)	0.0843	0.8069	0.1167	0.3036	0.4098	0.0775	0.9581	0.0658	0.9244	1.1833

Table 5: Analysis of the decoding strategies for the question-answering task.

vides a model- and task-independent proxy for output correctness.

F Selective Filtering Results

This section extends the selective-filtering analysis presented in Section 5.3. Following the same procedure, we rank generations by their average predictive entropy $H(y)$, progressively remove those with the highest entropy, and recompute the evaluation metrics over the retained outputs. To facilitate comparison with the main-text analysis, we

report BERT-F1 and AlignScore as representative measures of generation quality and factuality, respectively.

Figures 10–14 present results across all four tasks and all five models. The extended results illustrate the task- and model-dependent nature of the filtering effect and provide the broader context for the representative Counterfactual results reported in Figure 3.

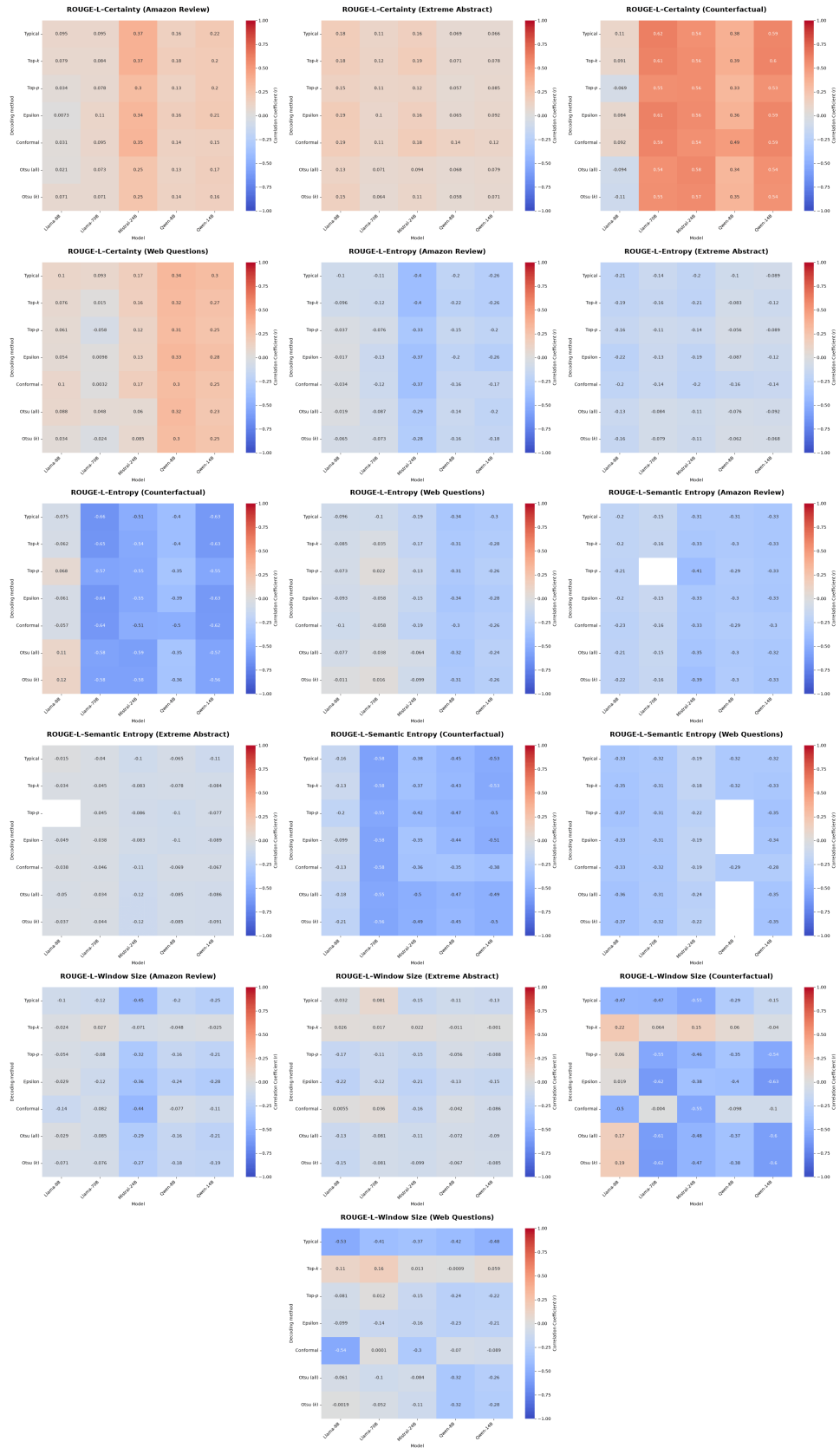


Figure 4: Correlation heatmaps between the ROUGE-L and uncertainty-related measures.

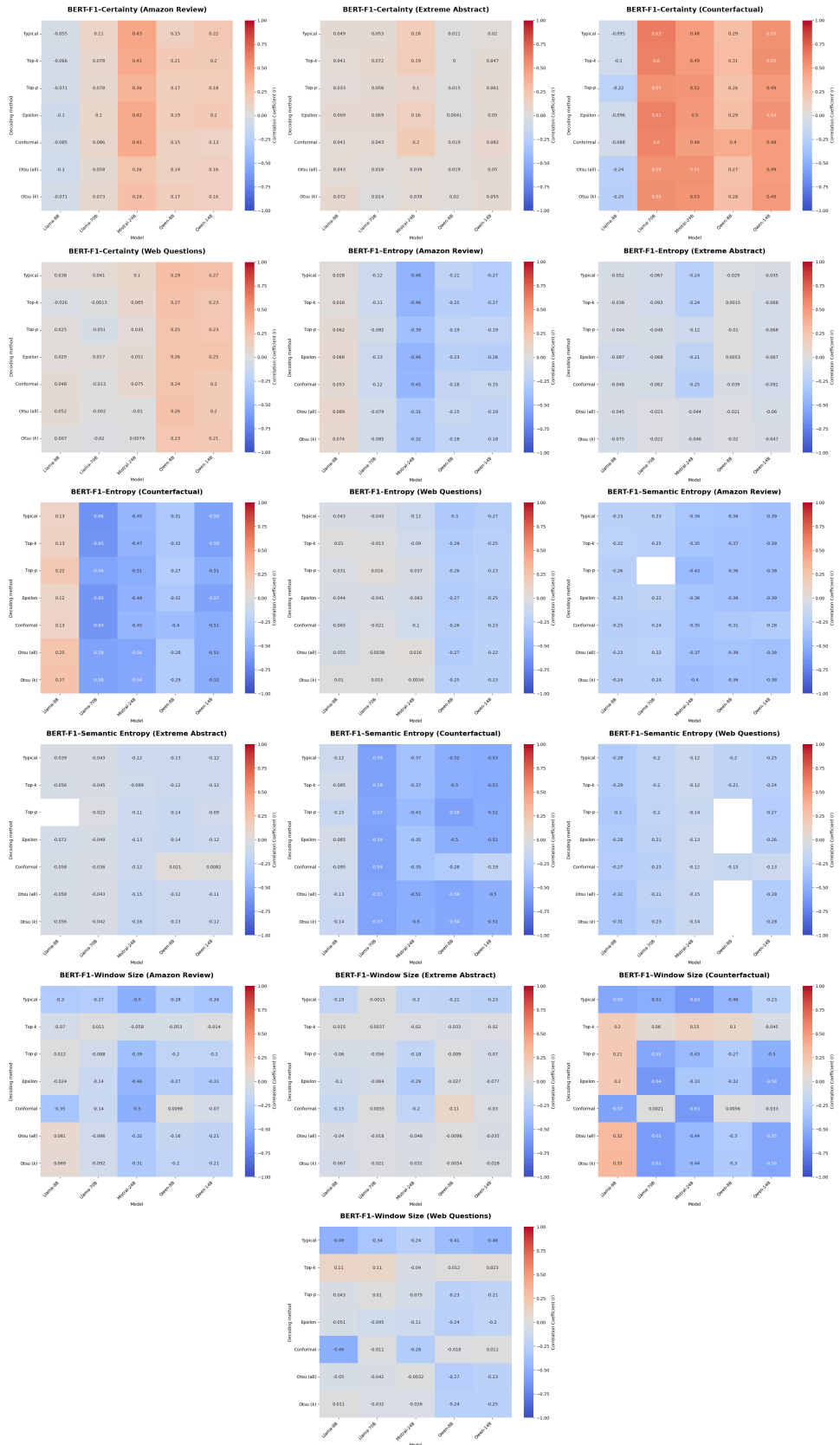


Figure 5: Correlation heatmaps between the BERT-F1 and uncertainty-related measures.

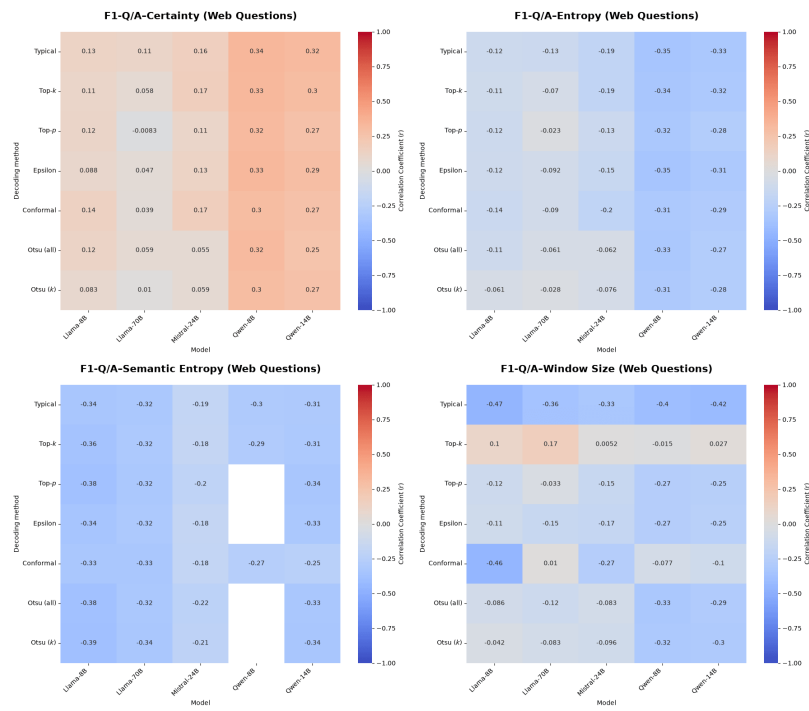


Figure 6: Correlation heatmaps between the F1-Q/A and uncertainty-related measures.

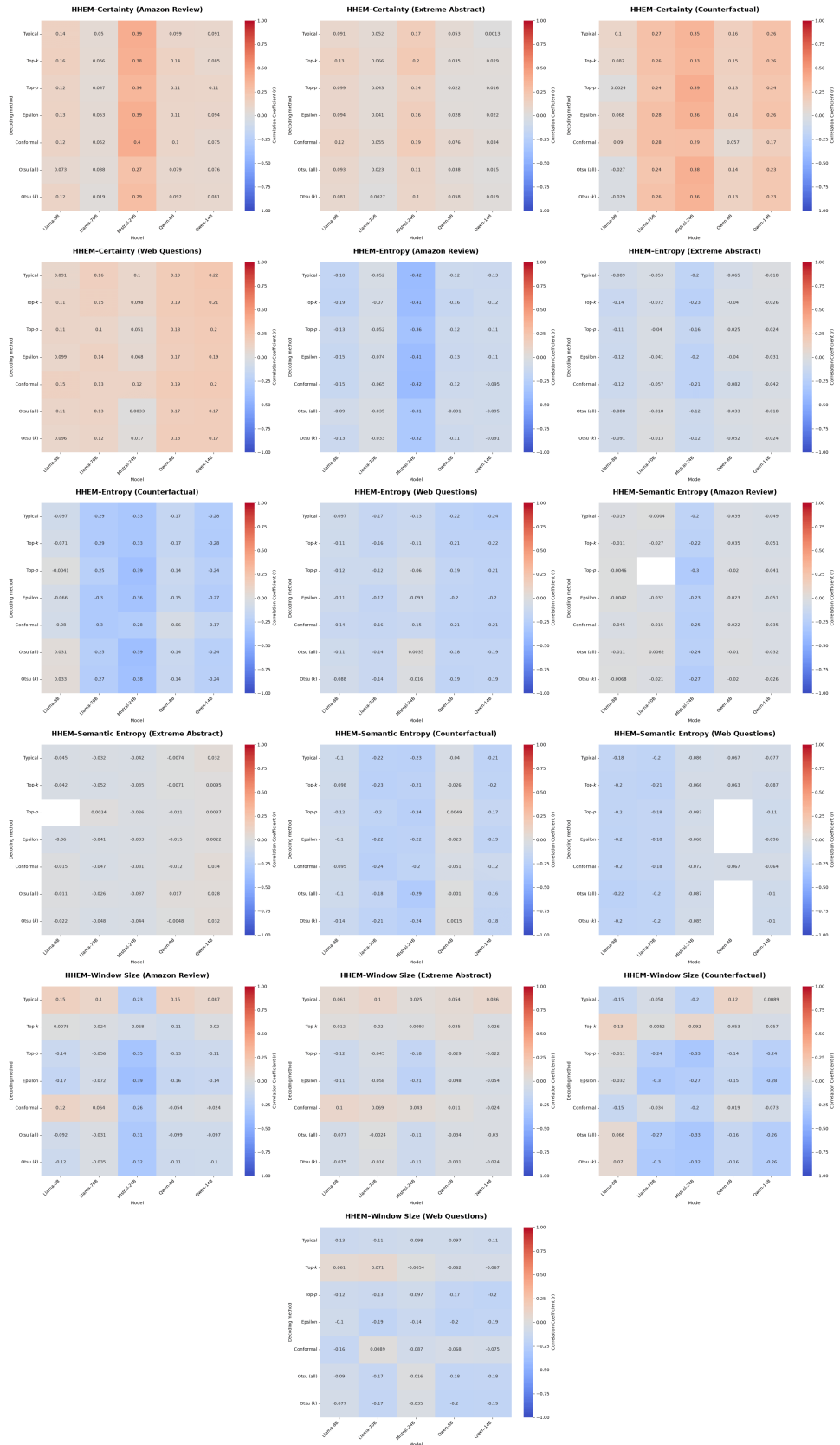


Figure 7: Correlation heatmaps between the HHEM and uncertainty-related measures.

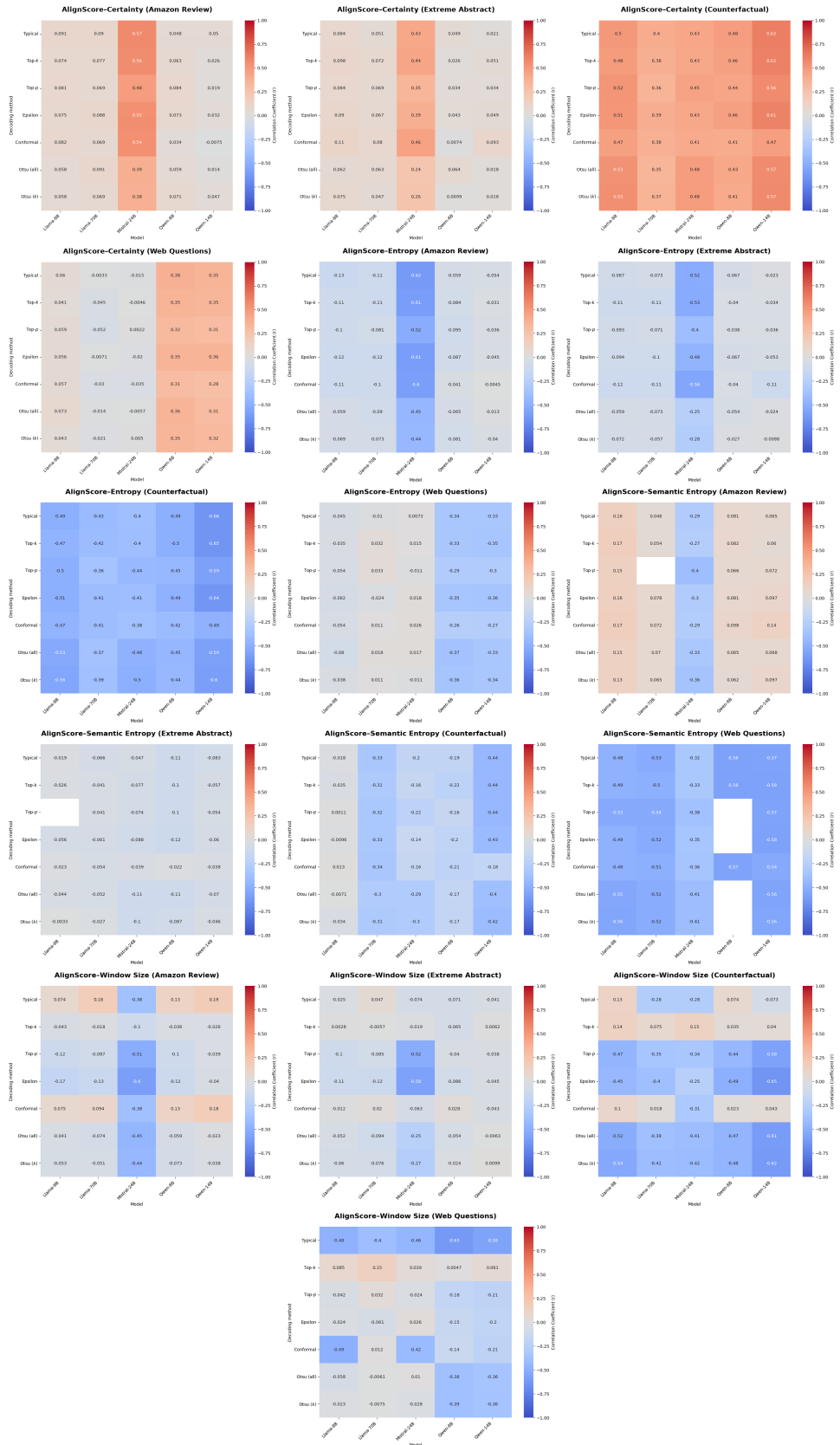


Figure 8: Correlation heatmaps between the AlignScore and uncertainty-related measures.

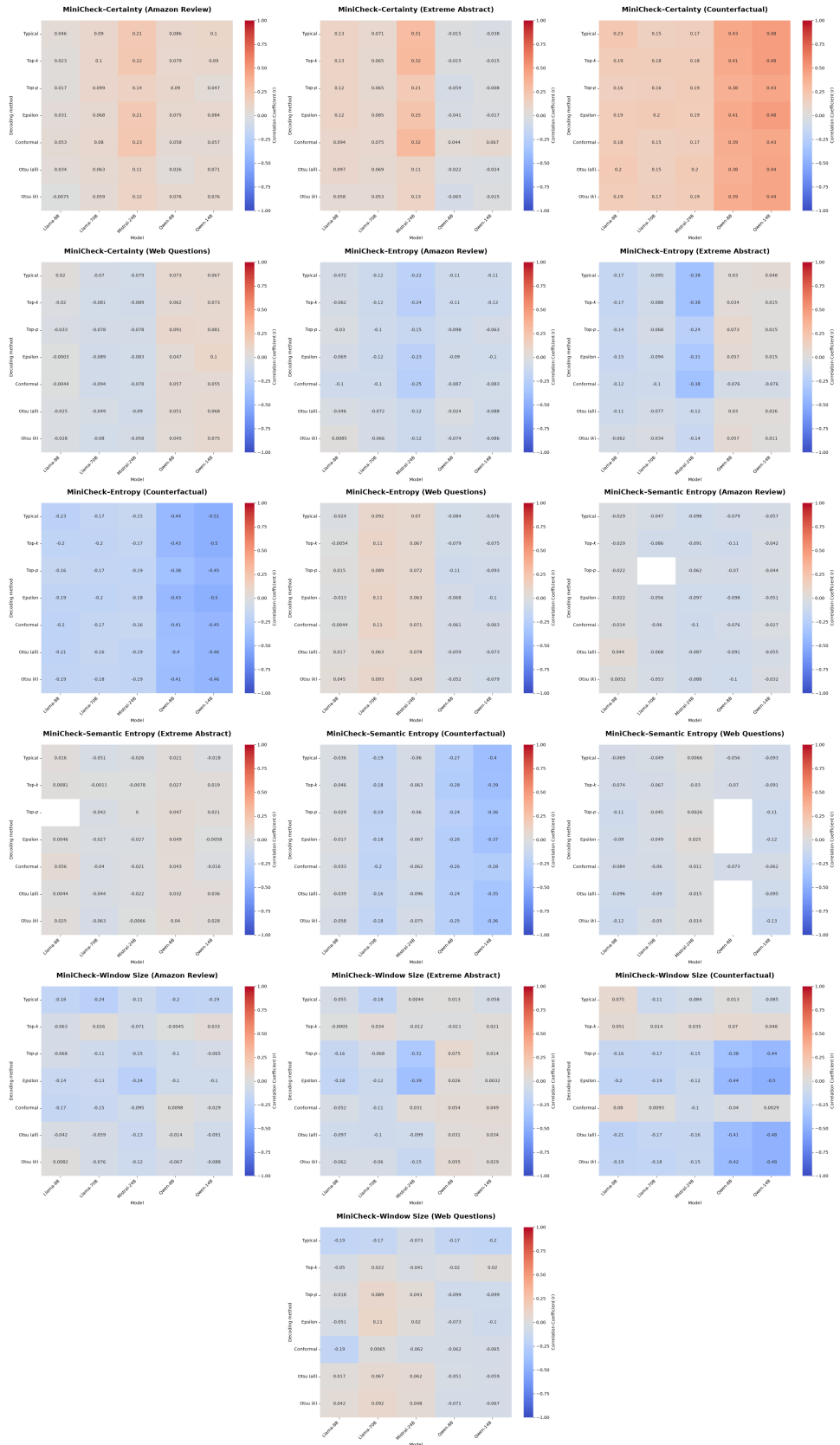


Figure 9: Correlation heatmaps between the MiniCheck and uncertainty-related measures.

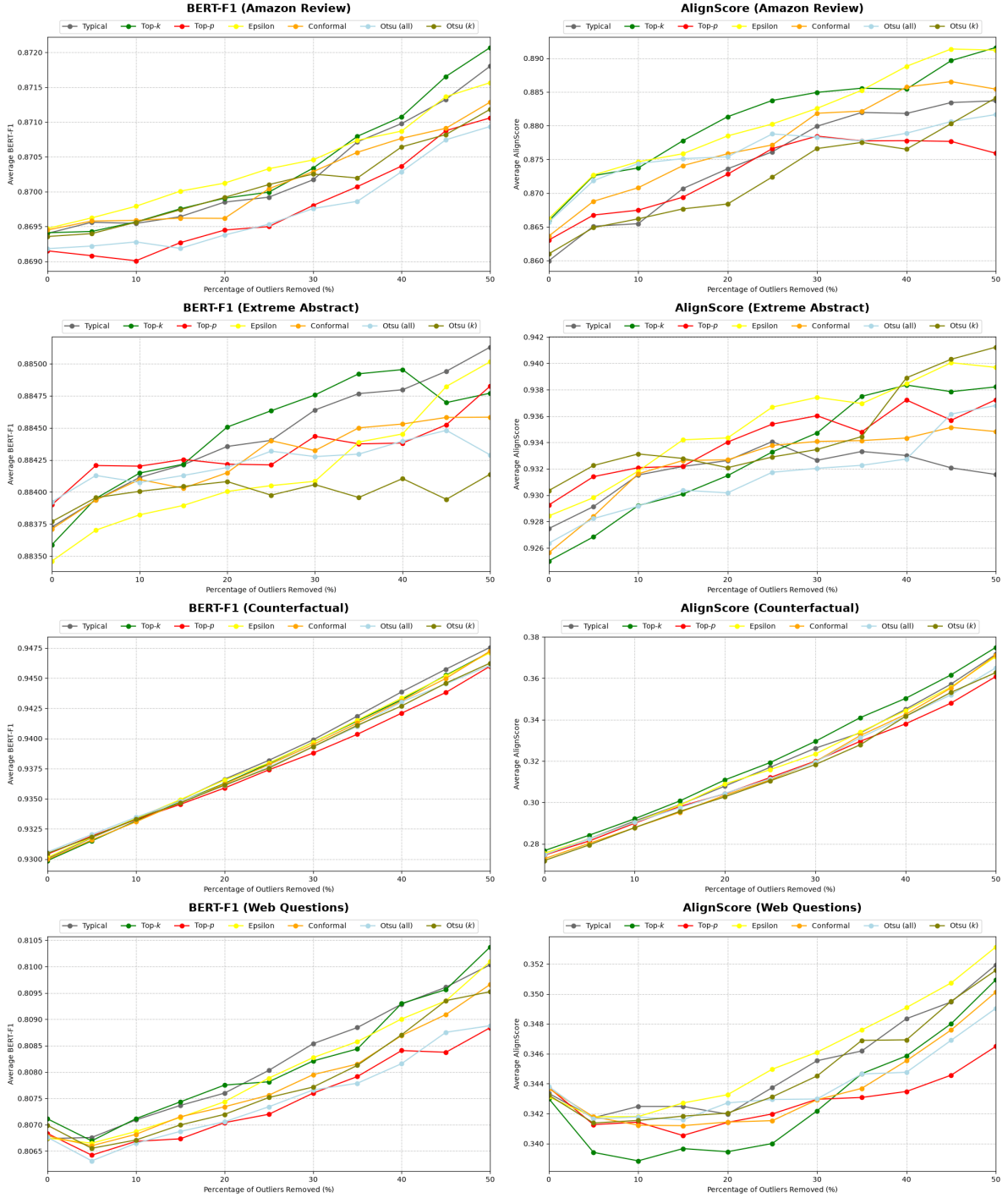


Figure 10: Selective-filtering results for LLaMA-70B. The left column reports BERT-F1 and the right column AlignScore; rows correspond, from top to bottom, to Amazon Review, Extreme Abstract, Counterfactual, and Web Questions. Increasing fractions of the most uncertain generations are removed and the metrics are recomputed over the retained outputs.

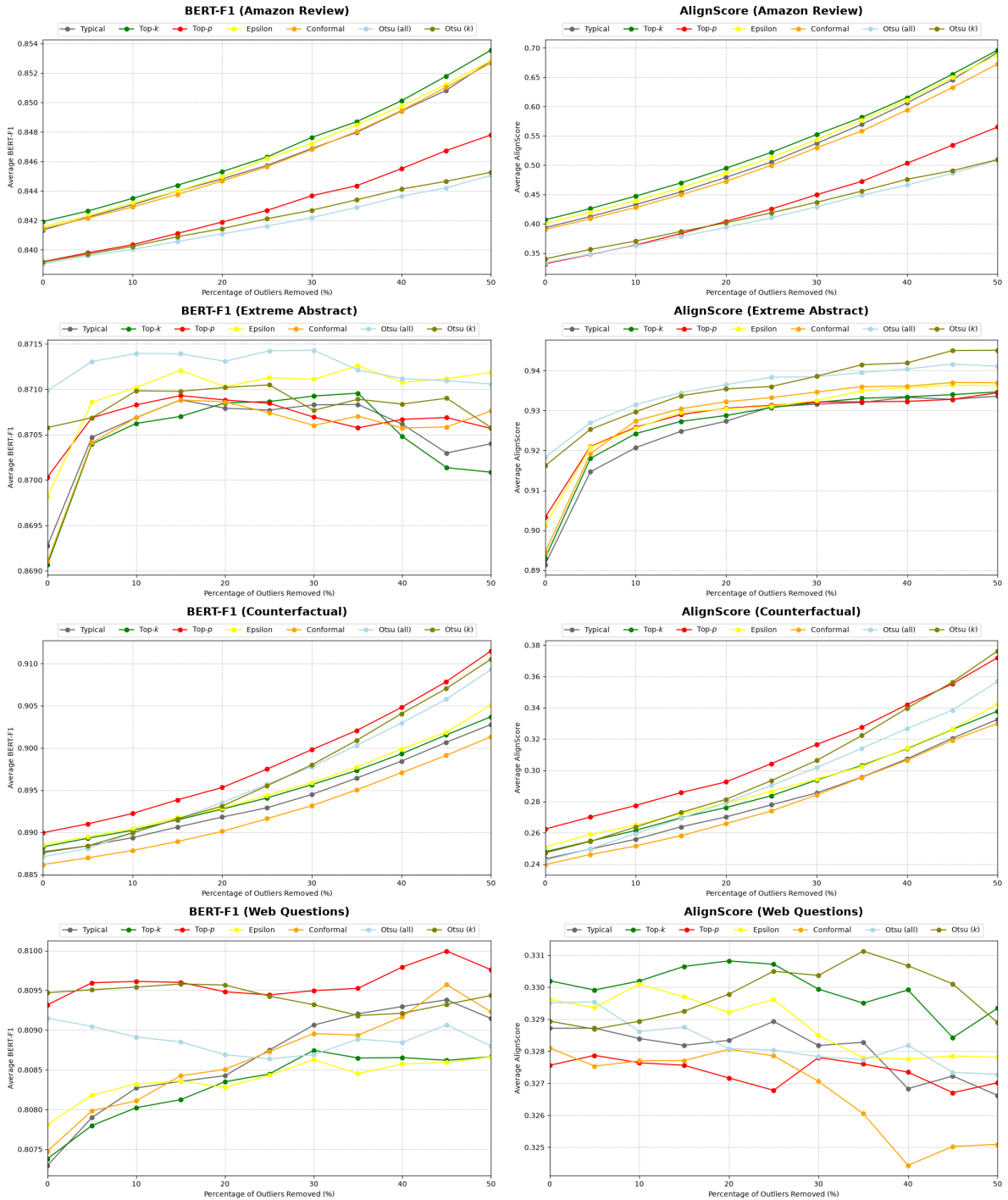


Figure 11: Selective-filtering results for Mistral-24B. The left column reports BERT-F1 and the right column AlignScore; rows correspond, from top to bottom, to Amazon Review, Extreme Abstract, Counterfactual, and Web Questions. Increasing fractions of the most uncertain generations are removed and the metrics are recomputed over the retained outputs.

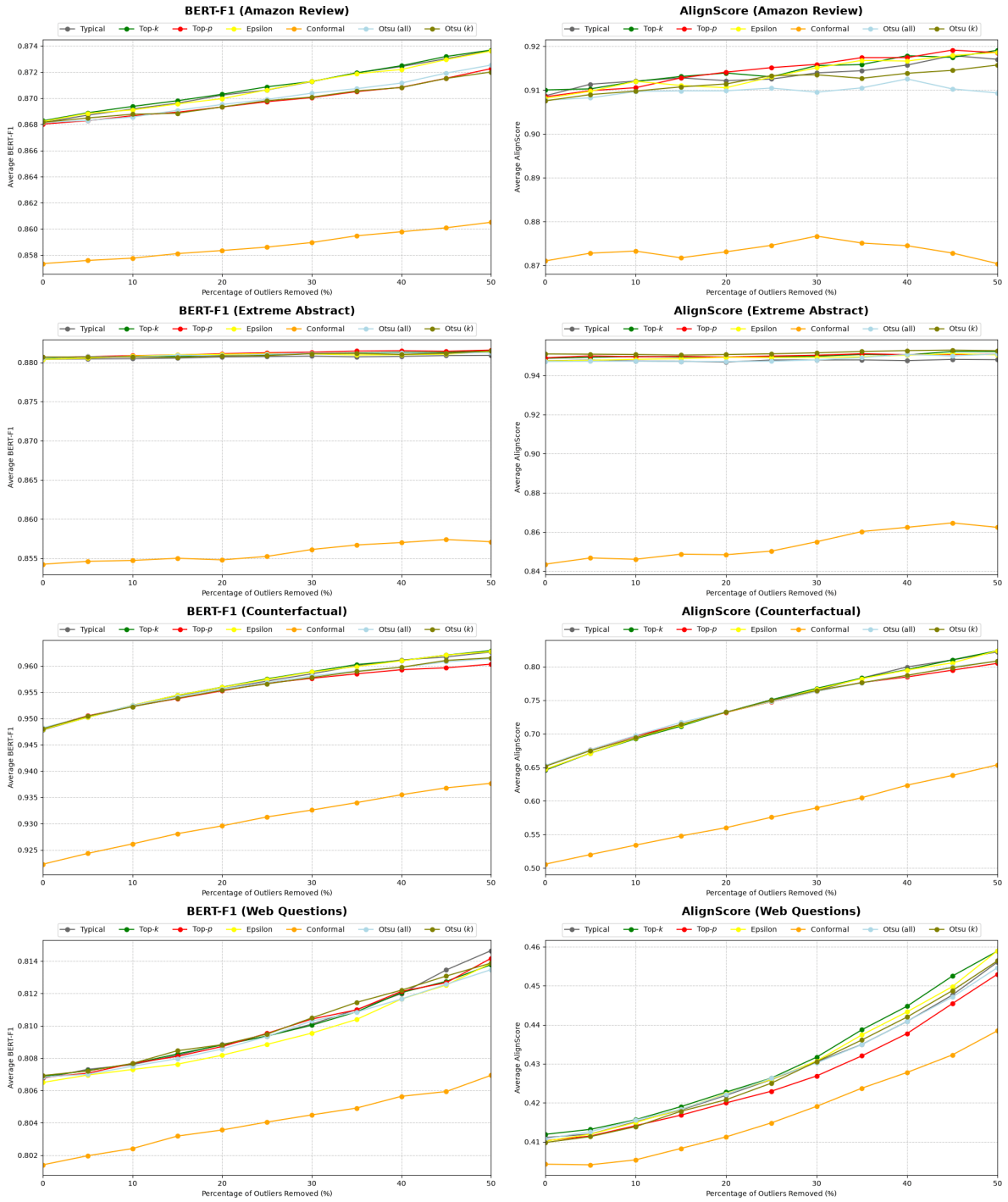


Figure 12: Selective-filtering results for Qwen-14B. The left column reports BERT-F1 and the right column AlignScore; rows correspond, from top to bottom, to Amazon Review, Extreme Abstract, Counterfactual, and Web Questions. Increasing fractions of the most uncertain generations are removed and the metrics are recomputed over the retained outputs.

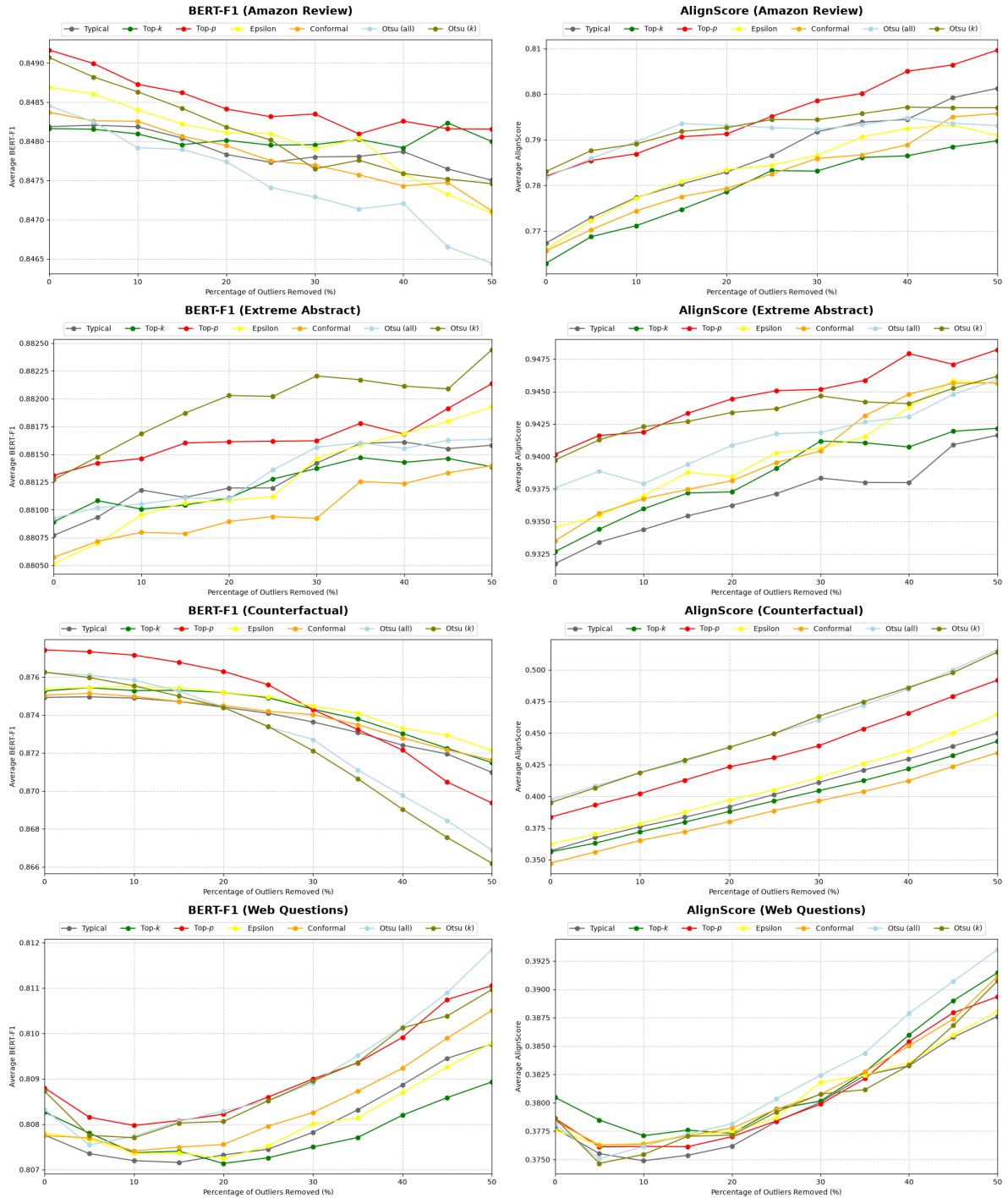


Figure 13: Selective-filtering results for LLaMA-8B. The left column reports BERT-F1 and the right column AlignScore; rows correspond, from top to bottom, to Amazon Review, Extreme Abstract, Counterfactual, and Web Questions. Increasing fractions of the most uncertain generations are removed and the metrics are recomputed over the retained outputs.

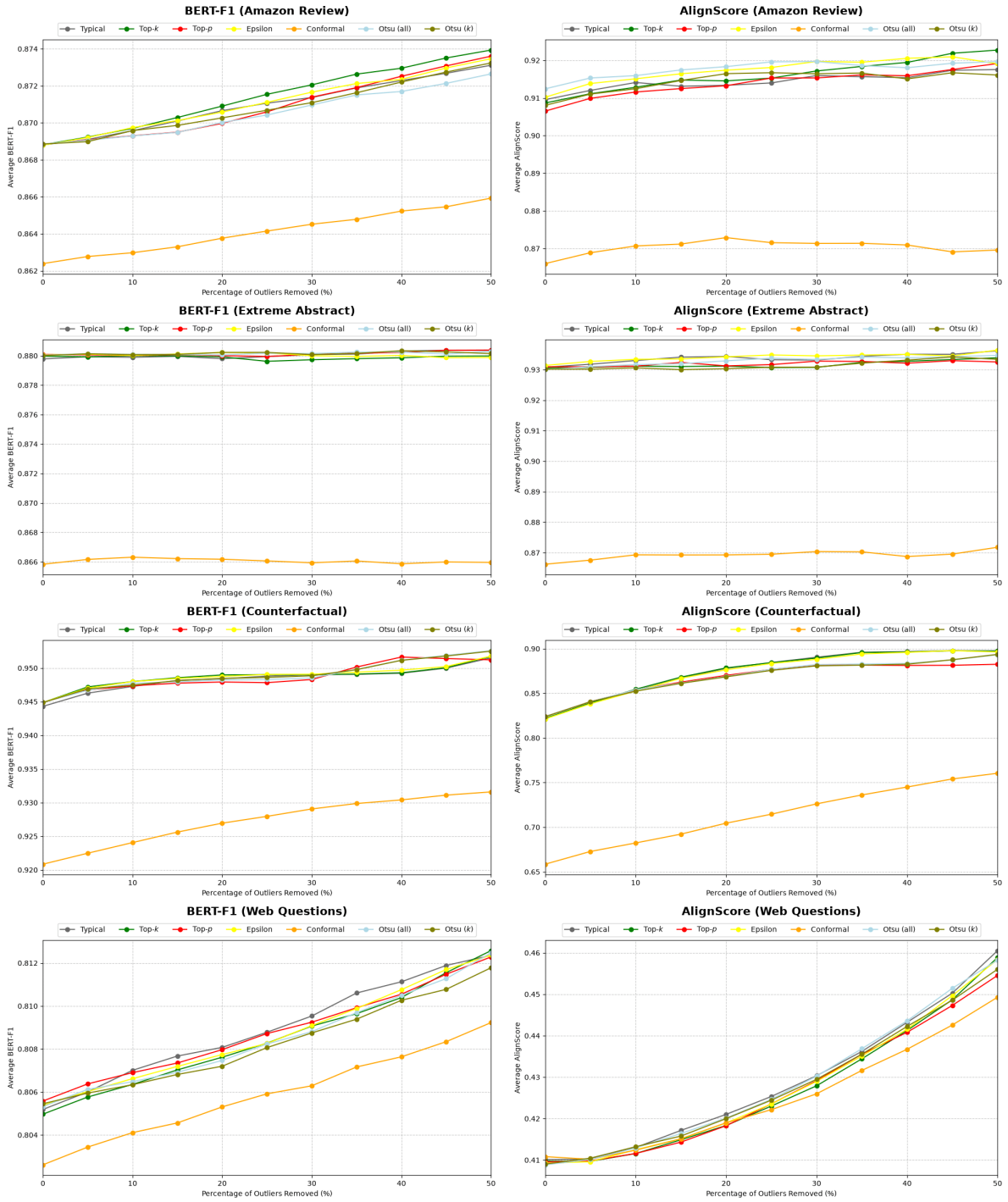


Figure 14: Selective-filtering results for Qwen-8B. The left column reports BERT-F1 and the right column AlignScore; rows correspond, from top to bottom, to Amazon Review, Extreme Abstract, Counterfactual, and Web Questions. Increasing fractions of the most uncertain generations are removed and the metrics are recomputed over the retained outputs.

G Examples of Otsu-based generation

For a better understanding of the results, we present some examples of text generations from the experiments. In detail, we note the summaries with the lowest entropy for all the dataset categories, the Qwen-14B model, and the Otsu’s variants.

Amazon Review (Otsu all)

ID: task618-4fa348bcac24454d8a97f504ebd519eb

Output: The customer is still waiting for the product protection plan.

Amazon Review (Otsu k)

ID: task618-4fa348bcac24454d8a97f504ebd519eb

Output: The customer is still waiting for the product protection plan.

Extreme Abstract (Otsu all)

ID: task668-c2deb513b74645bd8e20136b25b925ac

Output: This study proposes a data augmentation method using bilingual dictionaries to improve semi-supervised neural machine translation, enhancing vocabulary and performance in low-resource settings.

Extreme Abstract (Otsu k)

ID: task668-c2deb513b74645bd8e20136b25b925ac

Output: This study proposes a data augmentation method using bilingual dictionaries to improve semi-supervised neural machine translation, enhancing vocabulary and performance in low-resource settings.

Counterfactual (Otsu all)

ID: task269-ab6dfdd399f1428199e05fa493168c45

Output: When he logged into his account he saw he had 20 free pizzas due. Ralph threw a party and got all 20 pizzas free. The next day, Ralph ate at Dominos instead.

Counterfactual (Otsu k)

ID: task269-ab6dfdd399f1428199e05fa493168c45

Output: When he logged into his account he saw he had 20 free pizzas due. Ralph threw a party and got all 20 pizzas free. The next day, Ralph ate at Dominos instead.

Web Questions (Otsu all)

ID: task1412-c1be6342f18646fcb247d8697f004372

Output: The first Ford Motor Company was located in Detroit, Michigan.

Web Questions (Otsu k)

ID: task1412-c1be6342f18646fcb247d8697f004372

Output: The first Ford Motor Company was located in Detroit, Michigan.