Diversity Explains Inference Scaling Laws: Through a Case Study of Minimum Bayes Risk Decoding

Hidetaka Kamigaito[1], **Hiroyuki Deguchi**[1], **Yusuke Sakai**[1],
Katsuhiko Hayashi[2], **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology (NAIST), [2]The University of Tokyo
{kamigaito.h, deguchi.hiroyuki.db0, sakai.yusuke.sr9,taro}@is.naist.jp
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

Abstract

Inference methods play an important role in eliciting the performance of large language models (LLMs). Currently, LLMs use inference methods utilizing generated multiple samples, which can be derived from Minimum Bayes Risk (MBR) Decoding. Previous studies have conducted empirical analyses to clarify the improvements in generation performance achieved by MBR decoding and have reported various observations. However, the theoretical underpinnings of these findings remain uncertain. To address this, we offer a new theoretical interpretation of MBR decoding from the perspective of bias–diversity decomposition. In this interpretation, the error in the quality estimation of hypotheses by MBR decoding is decomposed into two main factors: bias, which considers the closeness between the utility function and human evaluation, and diversity, which represents the variability in the quality estimation of the utility function. The theoretical analysis reveals the difficulty of simultaneously improving bias and diversity, confirming the validity of enhancing MBR decoding performance by increasing diversity. Furthermore, we reveal that diversity can explain one aspect of inference scaling laws that describe performance improvement by increasing sample size. Moreover, experiments across multiple NLP tasks yielded results consistent with these theoretical characteristics. Our code is available at https://github.com/naist-nlp/mbr-bias-diversity.

1 Introduction

As demonstrated by the success of large language models (LLMs) (Brown et al., 2020; OpenAI et al., 2024), text generation is one of the most fundamental tasks in Natural Language Processing (NLP). In the generation, inference methods play an important role in eliciting model ability. In parallel to the advance of LLMs, inference methods utilizing multiple samples such as self-consistency (SC) (Wang et al., 2023) and Complex SC (Fu et al., 2023) are introduced. Bertsch et al. (2023) prove that these methods can be derived from Minimum Bayes Risk (MBR) decoding (Goel and Byrne, 2000).

MBR decoding can elicit models' generation performance by using a utility function, essentially an automatic evaluation metric, along with pseudo-references generated by the model. MBR decoding was initially applied to speech recognition (Goel and Byrne, 2000) and later to statistical machine translation (SMT) (Kumar and Byrne, 2002, 2004; Duan et al., 2011). Following these successes, MBR decoding has been expanded to various text generation tasks, including neural machine translation (NMT) (Stahlberg et al., 2017), text summarization (Bertsch et al., 2023), and image captioning (Borgeaud and Emerson, 2020).

Since MBR decoding has become an important inference technique in text generation, various empirical studies have explored its characteristics. Müller and Sennrich (2021); Freitag et al. (2022a); Fernandes et al. (2022); Amrhein and Sennrich (2022) highlight the importance of using high-quality evaluation metrics that is robust and correlate well with human evaluations as utility functions. Jinnai et al. (2024a); Heineman et al. (2024) emphasize the importance of high-quality pseudo-references that closely resemble human-created ones while stressing the significance of pseudo-reference diversity. Although these empirical findings cover various aspects in detail, a unified interpretation remains challenging due to the lack of theoretical frameworks explaining the relationships behind them.

To address this gap, we provide theoretical interpretations of MBR decoding through bias-diversity decomposition (Krogh and Vedelsby, 1994; Wood et al., 2024). Our theoretical interpretation focuses on errors in the estimated quality of hypotheses in MBR decoding. These errors are decomposed into two critical factors: *bias* and *diversity*. The

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29060–29094

bias term represents the closeness between the estimated quality produced by utility functions and human evaluations. The diversity term reflects the variance in the estimated quality across different utility functions. Based on this interpretation, we theoretically demonstrate the difficulty of improving both the bias and diversity terms simultaneously and highlight the effectiveness of increasing diversity in MBR decoding, verifying the correspondence with empirically induced results from previous work.

Furthermore, by focusing on information-theoretic diversity (Brown, 2009; Zhou and Li, 2010), we broaden the scope of our analysis beyond MBR decoding and reveal that diversity is also a key to explaining inference scaling laws (Wu et al., 2024; Brown et al., 2024; Chen et al., 2025; Snell et al., 2025) which describe the performance improvement by increasing sample size.

Our empirical analysis on machine translation, text summarization, and image captioning using five different sampling methods shows consistent results with our theoretical analysis.

2 Minimum Bayes Risk Decoding

MBR decoding (Eikema and Aziz, 2020, 2022) estimates the quality of a hypothesis h in the candidate set $\mathcal{H}$ by using a set $\mathcal{Y}$ of pseudo-references y sampled from the model's predicted probability $P_\pi(y|x)$ with its model weight π for the input sequence x. By treating the evaluation metric as a utility function $f_\theta(h, y)$ that measures the similarity between h and y, MBR decoding selects the best hypothesis $\hat{h}_{mbr}$ in $\mathcal{H}$ as:

$$\hat{h}_{mbr} = \underset{h\in\mathcal{H}}{\text{argmax}} \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_\theta(h, y), y \sim P_\pi(y|x). \quad (1)$$

$$\approx \underset{h\in\mathcal{H}}{\text{argmax}} \sum_{y} f_\theta(h, y) P_\pi(y|x) \quad (2)$$

Here, θ represents the parameters of the evaluation metric used in the utility function $f_\theta(h, y)$.

Alternatively, instead of using the utility function $f_\theta(h, y)$, one can assume the quality estimated by humans (Naskar et al., 2023; Suzgun et al., 2023; Jinnai et al., 2024a; Ohashi et al., 2024) as $\hat{f}_{\hat{\theta}}(h)$. Under this assumption, the ideal decoding as estimated by humans is given by:

$$\hat{h}_{human} = \underset{h\in\mathcal{H}}{\text{argmax}} \hat{f}_{\hat{\theta}}(h). \quad (3)$$

In this paper, we focus on analyzing the differences between the quality estimated by MBR decoding and that estimated by humans to better understand MBR decoding.

3 Theoretical Analysis based on Bias-Diversity Decomposition

3.1 Evaluation Discrepancy

To measure the discrepancy between the human estimated quality, $\hat{f}_{\hat{\theta}}(h)$ and the MBR decoding estimated quality, $\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_\theta(h, y)$, we define a $|\mathcal{H}|$-dimensional vector $\mathbf{u}^j$ that represents estimated quality for each hypothesis based on the j-th pseudo-reference and also define $\bar{\mathbf{u}}$, the average vector of all $\mathbf{u}^j$ as follows:

$$\mathbf{u}^j = \begin{bmatrix} u_1^j \\ \cdots \\ u_{|\mathcal{H}|}^j \end{bmatrix}, u_i^j = f_\theta(h_i, y_j), \bar{\mathbf{u}} = \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} \mathbf{u}^j. \quad (4)$$

Similarly, we can define a $|\mathcal{H}|$-dimensional vector, $\hat{\mathbf{u}}$ that represents the human estimated quality for each hypothesis as follows:

$$\hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 \\ \cdots \\ \hat{u}_{|\mathcal{H}|} \end{bmatrix}, \quad \hat{u}_i = \hat{f}_{\hat{\theta}}(h_i). \quad (5)$$

Here, by using Eqs. (4) and (5), we can reformulate MBR decoding in Eq. (1) and the ideal decoding in Eq. (3) as follows:

$$(1) \equiv \hat{h}_{mbr} = \underset{h_i}{\arg\max}\, \bar{u}_i,$$
$$(3) \equiv \hat{h}_{human} = \underset{h_i}{\arg\max}\, \hat{u}_i. \quad (6)$$

Therefore, based on Eq. (6), we can investigate the discrepancy between the estimated quality by MBR decoding and human through the comparison of $\bar{\mathbf{u}}$ and $\hat{\mathbf{u}}$. In our work, to estimate the discrepancy, we consider the prediction error of $\bar{\mathbf{u}}$ to $\hat{\mathbf{u}}$ by using Mean Squared Error (MSE) as follows:

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) = \frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}(\hat{u}_i - \bar{u}_i)^2 \quad (7)$$

$$= \mathbb{E}_{i\in\mathcal{H}}[(\hat{u}_i - \bar{u}_i)^2]. \quad (8)$$

3.2 Bias-diversity Decomposition

Our goal is to reveal the characteristics of MBR decoding through theoretical analysis. To achieve this, we focus on the bias and diversity underlying Eq. (8). Based on this approach, we can induce the following decomposition:

Theorem 1 (**Bias and Diversity Decomposition**). *The quality estimation error for MBR decoding, $MSE(\hat{\boldsymbol{u}}, \bar{\boldsymbol{u}})$, can be decomposed into bias and diversity (ambiguity) terms (Krogh and Vedelsby, 1994) as follows:*

$$\begin{aligned} &MSE(\hat{\boldsymbol{u}}, \bar{\boldsymbol{u}}) \\ =&\underbrace{\mathbb{E}_{i\in\mathcal{H}}\mathbb{E}_{j\in\mathcal{Y}}[(\hat{u}_i - f_\theta(h_i, y_j))^2]}_{Bias} \\ &- \underbrace{\mathbb{E}_{i\in\mathcal{H}}\mathbb{E}_{j\in\mathcal{Y}}[(\bar{u}_i - f_\theta(h_i, y_j))^2]}_{Diversity}. \end{aligned} \quad (9)$$

(See Appendix A.1 for the proof.)

In Eq. (9), two terms represent bias and diversity. Unlike the well-known bias-variance decomposition (Geman et al., 1992) that targets a single estimator, which is $\mathbf{u}$ in our case, the second term is negative, which is why it is referred to as diversity rather than variance (Wood et al., 2024). *Bias* indicates how closely the utility function's estimated quality for a hypothesis matches human estimation. *Diveristy* reflects how different the utility function's estimated qualities are from each other. This decomposition emphasizes the importance of increasing *Diveristy* while reducing *Bias* to improve the quality estimation error, $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$.

Even though Theorem 1 indicates the potential of *Diversity* to improve the performance in MBR decoding, there are some limitations.

Theorem 2 (**Limitation of Diversity**). *The decomposition of the quality estimation error for MBR decoding (Eq. (9)) holds the following relation:*

$$Diversity \xrightarrow[Bias\to 0]{} 0 \quad (10)$$

(See Appendix A.2 for the proof.)

According to Theorem 2, we cannot expect performance gains from *Diversity* when *Bias* is close to zero. Note that this relationship does not guarantee that *Diversity* becomes large when *Bias* is far from zero. As a broadly generalized relationship of this part, the following theorem holds.

Theorem 3 (**Bias and Diversity Trade-off**). *Bias and Diversity in the decomposition of the quality estimation error for MBR decoding (Eq. (9)) depend on each other based on the following reformulation by Brown et al. (2005):*

$$\begin{aligned} Bias &= \overline{bias}^2 + \Omega \\ Diversity &= \Omega - \left[\frac{1}{|\mathcal{Y}|}\overline{var} + \left(1 - \frac{1}{|\mathcal{Y}|}\right)\overline{cov}\right], \end{aligned} \quad (11)$$

where $\overline{bias} = \mathbb{E}_{i\in\mathcal{Y}}[\mathbb{E}_{j\in\mathcal{H}}[u^i_j] - \mathbb{E}_{j\in\mathcal{H}}[\hat{u}_j]]$, $\overline{var} = \mathbb{E}_{i\in\mathcal{Y}}[\mathbb{E}_{j\in\mathcal{H}}[(u^i_j - \mathbb{E}_{k\in\mathcal{H}}[u^i_k])^2]]$, $\overline{cov} = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)} \sum_{i=1}^{|\mathcal{Y}|}\sum_{i\neq j}^{|\mathcal{Y}|} \mathbb{E}_{k\in\mathcal{H}}[(u^i_k - \mathbb{E}_{l\in\mathcal{H}}[u^i_l])(u^i_k - \mathbb{E}_{l\in\mathcal{H}}[u^j_l])]$, *and* $\Omega = \overline{var} + \mathbb{E}_{i\in\mathcal{Y}}[(\mathbb{E}_{k\in\mathcal{H}}[u^i_k] - \mathbb{E}_{k\in\mathcal{H}}[\overline{u}_k])^2]$. *(See Appendix A.3 for the proof.)*

In Eq. (11), *Bias* and *Diversity* share Ω. Thus, Theorem 3 demonstrates the general trade-off relationship between *Bias* and *Diversity*, underscoring the difficulty of maximizing *Diversity* without affecting *Bias*.

3.3 Diversity behind Inference Scaling Laws

In some applications, MBR decoding is part of a reranking system that employs a scoring function other than $f_\theta(h, y)$. In addition, inference methods are not restricted to the one derived from MBR decoding. To broaden our scope of analysis beyond MBR decoding and analyze the inference scaling laws, we rely on the following theorem.

Theorem 4 (**Generalized Diveristy**). *Letting $\hat{\mathcal{H}}$ be a distribution for the human-selected candidate, $\mathcal{X}_{1:|\mathcal{Y}|}$ be representations corresponding to all elements in $\mathcal{Y}$, and $g(\mathcal{X}_{1:|\mathcal{Y}|})$ be a function predicting the best candidate. Its prediction error $p(\hat{\mathcal{H}} \neq g(\mathcal{X}_{1:|\mathcal{Y}|}))$ satisfies the following inequality (Brown, 2009; Zhou and Li, 2010):*

$$\begin{aligned} &\frac{H(\hat{\mathcal{H}}) - I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}}) - 1}{\log|\hat{\mathcal{H}}|} \leq p(\hat{\mathcal{H}} \neq g(\mathcal{X}_{1:|\mathcal{Y}|})), \\ &p(\hat{\mathcal{H}} \neq g(\mathcal{X}_{1:|\mathcal{Y}|})) \leq \frac{H(\hat{\mathcal{H}}) - I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})}{2}, \end{aligned} \quad (12)$$

where H is entropy and I is mutual information. To minimize the error, we should maximize $I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})$, decomposed to (Zhou and Li, 2010):

$$\underbrace{\sum_{i=1}^{|\mathcal{Y}|} I(\mathcal{X}_i; \hat{\mathcal{H}})}_{Relevancy} + \underbrace{\underbrace{\mathcal{I}(\mathcal{X}_{1:|\mathcal{Y}|}|\hat{\mathcal{H}})}_{Conditional\ Redundancy} \underbrace{-\mathcal{I}(\mathcal{X}_{1:|\mathcal{Y}|})}_{Redundancy}}_{Information-Theoretic\ Diversity}, \quad (13)$$

where $\mathcal{I}(\mathcal{X}_{1:|\mathcal{Y}|})$ and $\mathcal{I}(\mathcal{X}_{1:|\mathcal{Y}|}|\hat{\mathcal{H}})$ are total correlation and conditional total correlation, respectively. (See Appendix A.4 for the proof.)

For maximizing *Information Theoretic Diveristy* in Eq. (13), *Redundancy* must be zero. This condition is satisfied when all elements in $\mathcal{X}_{1:|\mathcal{Y}|}$ are independent of each other. That shows the importance of the diversity of generated samples in $\mathcal{Y}$. This theoretical aspect is consistent with the importance of the diversity of $\mathcal{Y}$ shown in Eq. (9).

Theorem 5 (**Monotonicity of Terms**). *When $\mathcal{X}_{1:|\mathcal{Y}|}$ increases (that means new elements are added to $\mathcal{X}_{1:|\mathcal{Y}|}$), Relevancy, Conditional Redundancy, and Redundancy monotonically increase, while Information Theoretic Diveristy and $I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})$ do not monotonically increase or decrease. (See Appendix A.5 for the proof.)*

This theorem concludes that just increasing sample size does not guarantee performance improvement in inference due to the degradation of *Information Theoretic Diveristy*. Thus, diversity is still a key to performance improvement also from the information-theoretic interpretation. The bounds further support this characteristic as follows.

Theorem 6 (**Monotonicity of Bounds**). *The lower and upper bounds of the error in Eq. (12) do not monotonically increase or decrease when $\mathcal{X}_{1:|\mathcal{Y}|}$ increases. (See Appendix A.6 for the proof.)*

However, these characteristics contradict previous work reporting performance improvement by increasing sample size. The following theorem can fill in the gap.

Theorem 7 (**Submodularity of Terms**). *When variables in $\mathcal{X}_{1:|\mathcal{Y}|}$ are independent given $\hat{\mathcal{H}}$, $I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})$ and Information-Theoretic Diversity have submodularity on $\mathcal{X}_{1:|\mathcal{Y}|}$ and $I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})$ does not decrease when $\mathcal{X}_{1:|\mathcal{Y}|}$ increases. (See Appendix A.7 for the proof.)*

This theorem shows a kind of lower bound for the performance because *Conditional Redundancy* becomes zero under the condition. Here, submodularity is a characteristic that the effect for performance improvement by increasing $\mathcal{X}_{1:|\mathcal{Y}|}$ decreases. Thus, under the conditional independence of $\mathcal{X}_{1:|\mathcal{Y}|}$ given $\hat{\mathcal{H}}$, increasing $\mathcal{X}_{1:|\mathcal{Y}|}$ makes $I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})$ increase and gradually converge based on its submodularity. This behavior is along with the inference scaling laws that indicate the logarithmic convergence of performance by increasing sample size. The following theorem supports this aspect.

Theorem 8 (**Supermodularity of Bounds**). *When variables in $\mathcal{X}_{1:|\mathcal{Y}|}$ are independent given $\hat{\mathcal{H}}$, the lower and upper bounds of the error in Eq. (12) are supermodular and non-increasing on $\mathcal{X}_{1:|\mathcal{Y}|}$. (See Appendix A.8 for the proof.)*

Therefore, prediction error $p(\hat{\mathcal{H}} \neq g(\mathcal{X}_{1:|\mathcal{Y}|}))$ decreases and converges corresponding to the increase of $I(\mathcal{X}_{1:|\mathcal{Y}|}; \hat{\mathcal{H}})$, since supermodularity is the negative of submodularity. This result supports the correspondence with inference scaling laws.

3.4 Interpretation

The decompositions and their details presented in these theorems allow us to provide theoretical interpretations for the empirically analyzed characteristics of MBR decoding and other inference methods in prior studies.

3.4.1 Correlation to Human Evaluation Results

Bias in Theorem 1 highlights the importance of considering the closeness between the human-estimated quality, $\hat{u}_i$, and the quality estimated by the utility function, $f_\theta(h_i, y_j)$, for improving the performance of MBR decoding. Specifically, since the utility function, $f_\theta(h_i, y_j)$, is influenced by the pseudo-reference y_j, *Bias* underscores the significance of considering the utility function's correlation to human evaluation and the closeness between pseudo-references and human-created references. Therefore, it emphasizes the importance of examining both utility functions and sampling strategies for generating pseudo-references.

Quality of Evaluation Metrics. Müller and Sennrich (2021); Freitag et al. (2022a); Fernandes et al. (2022); Amrhein and Sennrich (2022) support our theoretical insight, in which the quality of evaluation metrics used as utility functions is crucial for performance improvement.

Quality of Pseudo-References. Ohashi et al. (2024); Jinnai et al. (2024a) empirically show the importance of selecting appropriate pseudo-references. Our findings theoretically support these empirical insights.

Challenges in the Real World. Our theoretical findings emphasize the necessity of directly reducing the bias term. However, this requires human evaluation of the combination of pseudo-references and evaluation metrics, used as utility functions, for each hypothesis. This task is clearly challenging due to the high cost of human evaluation. As a solution, we suggest a method to approximate this in §4.1 and evaluate its correlation with task-specific performance in §5.3.

3.4.2 Diversity of Automatic Evaluation Results

Diversity in Theorem 1 shows increasing diversity can contribute to performance improvements in MBR decoding. A key insight here is that the diversity expressed by $(\bar{u}_i - f_\theta(h_i, y_j))^2$ stems from the different estimated qualities produced by each utility function $f_\theta(h_i, y_j)$. Thus, this diversity can

be influenced by the pseudo-reference y_j and/or the model parameters θ of the evaluation metric.

Diversity of Pseudo-references. This finding supports the previous studies (Freitag et al., 2023a; Jinnai et al., 2024a; Heineman et al., 2024) that conclude the diversity of sampling methods is essential for performance improvement of MBR decoding considering that the diversity of the pseudo-references can indirectly contribute to increasing the diversity of $f_\theta(h_i, y_j)$ by each y_j.

Diversity of Evaluation Metrics. Theoretically, we can anticipate performance improvements by combining multiple different evaluation metrics as utility functions to increase diversity. This is supported by empirical insights from Kovacs et al. (2024) in MBR decoding and Glushkova et al. (2023) in the quality estimation in text generation.

Unexplored Aspect. Furthermore, the effect of increasing the diversity of estimated qualities from utility functions by varying the evaluation metric's model parameters θ remains uncertain. To investigate this, we propose a method to adjust the diversity of estimated qualities by modifying θ in §4.2 and compare its behavior with that of varying pseudo-references in §5.5.

3.4.3 MBR Decoding as Ensemble Learning

Our decomposition in Theorem 1 aligns with ensemble learning, which is induced by Krogh and Vedelsby (1994). Thus, we can understand that the quality estimation by MBR decoding is a kind of ensemble learning.

Quality Estimation. Our decomposition starts from the definition $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$ in Eq. (8), the error between the estimated qualities from human evaluation and MBR decoding. We can actually observe the reduction of errors as the improvement in quality score estimation of (Naskar et al., 2023; Cheng and Vlachos, 2024) by ensembling utility functions that are similar to MBR decoding.

Weighted-voting. Furthermore, this viewpoint supports the validity of the previous work (Suzgun et al., 2023; Bertsch et al., 2023) that shows the interpretation of MBR decoding as soft-weighted voting, a variant of ensemble learning. Different from us, soft-weighted voting restricts the value range of voters (utility functions) from 0 to 1. Wood et al. (2024) shows that soft-weighted voting can be converted to the decomposition of Krogh and Vedelsby (1994), equivalent to our decomposition in Eq. (9). Therefore, weighted voting-based MBR decoding can be similarly explained in our decomposition.

Number of Pseudo-references. Generally, increasing the number of pseudo-references improves performance but demands additional computational costs. DeNero et al. (2009); Eikema and Aziz (2022); Cheng and Vlachos (2023); Deguchi et al. (2024b); Vamvas and Sennrich (2024); Trabelsi et al. (2024) prune samples to speedup inference and maintain the original quality similar to the case of pruning estimators in ensemble learning (Liu et al., 2004; Bonab and Can, 2016, 2019).

Considering an ensemble learning method, such as the Bayes optimal classifier (Mitchell, 1997), and assuming that Eq. (1) approximates the expectation by sampling y_j, we can explain the performance improvement of increased pseudo-references by the law of large numbers and the success of the pruning and weighted utility functions (Jinnai et al., 2024b) through importance sampling (Kloek and Van Dijk, 1978). (See Appendix B for more details.)

3.4.4 Bias and Diversity Trade-off

At first glance, based on the interpretation in §3.4.1 and §3.4.2, decreasing *Bias* while increasing *Diveristy* seems to be the best strategy to improve performance in MBR decoding, which was investigated by Jinnai et al. (2024a). To understand its validity, we need to focus on Theorems 2 and 3.

Limitation of MBR Decoding. Theorem 2 highlights the difficulty of increasing *Diversity* when *Bias* is close to zero. This theoretical fact indicates that even if we can prepare high-quality evaluation metrics and pseudo-references that correlate well with human behavior, there may be no performance improvement due to diminished *Diversity*. Furthermore, Theorem 3 highlights even when *Bias* is not close to zero, *Bias* influences *Diversity*.

Diversity Assists Inferior Methods. Conversely, when the evaluation metrics and pseudo-references are inferior, we can expect performance improvements through *Diversity* at the cost of increased *Bias*. This phenomenon can explain the sometimes competitive performance of BLEU (Papineni et al., 2002) against COMET (Rei et al., 2020) in Freitag et al. (2022b), and that of ancestral sampling (Robert, 1999) against other sampling methods in Freitag et al. (2023a); Ohashi et al. (2024) using MBR decoding. However, increased *Bias* does not guarantee increased *Diversity*. Therefore, we must carefully assess their diversity when using low-quality evaluation metrics and pseudo-references.

3.4.5 Diversity of Hypotheses

By replacing $\mathcal{X}_{1:|\mathcal{Y}|}$ in Theorem 4 with $\mathcal{X}_{1:|\mathcal{Y}\cup\mathcal{H}|}$, representations corresponding to all elements in $\mathcal{Y} \cup \mathcal{H}$, we can show the importance of diversity and sample size for hypotheses, through Theorems 4, 5, and 6, theoretically. That corresponds to the previous empirical studies, which focus on the sample size of hypotheses (Eikema and Aziz, 2020; Fernandes et al., 2022; Freitag et al., 2023a).

3.4.6 Diversity in Various Decoding Methods

Theorems 4, 5, and 6 demonstrate the importance of diversity for various decoding methods in addition to MBR decoding and its variants, such as universal self-consistency (Chen et al., 2024), reranking by rewards (Wu et al., 2024), and the combination with other reranking methods (Kovacs et al., 2024; Lyu et al., 2025). Basically, these theorems are applicable to other various decoding methods when they use hypotheses supported by §3.4.5.

3.4.7 Inference Scaling Laws

Theorems 7 and 8 show the theoretical background of inference scaling laws (Wu et al., 2024; Brown et al., 2024; Chen et al., 2025; Snell et al., 2025), which are applicable to various decoding methods supported by §3.4.5 and §3.4.6. This theoretical finding is based on the conditional independence of the representations of the samples given a correct output. This condition is natural in many inference approaches, where samples are independently generated to cover the correct output. Furthermore, since the characteristic shown in Theorems 7 and 8 is a kind of lower bound, we can expect further performance gains by improved *Conditional Redundancy*, interaction of samples to answer. In this situation, the performance improvement shows non-monotonic behavior. We check that in §5.4.

4 Remaining Problems & Solutions

Our theoretical analysis covers various aspects of MBR decoding. However, for a comprehensive analysis, we should investigate empirical results not addressed in previous work and bridge the gap between theory and real-world applications. To this end, we provide the following solutions.

4.1 Pseudo-Bias

As discussed in §3.4.1, the bias term suggests the importance of considering the correlation between the results of human evaluation and the evaluation metric's decisions based on pseudo-references to improve the performance of MBR decoding. However, calculating the bias term requires human evaluation, and conducting human evaluations for each setting is unrealistic and difficult. To address this issue, we introduce *pseudo-bias*, an approximation of the bias term in our decomposition. By using $|\hat{\mathcal{Y}}|$, the number of gold references $\hat{y}$, pseudo-bias is defined as follows:

$$\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\widetilde{u}_i - u_i^j)^2, \tag{14}$$

where $\widetilde{u}_i = \frac{1}{|\hat{\mathcal{Y}}|}\sum_{j=1}^{|\hat{\mathcal{Y}}|} f_\theta(h_i, \hat{y}_j)$. This formulation is based on the premise that automatic evaluation metrics correlate to human evaluation when receiving human-created references.[1] Since we can calculate the diversity term without any approximation, we compare pseudo-bias with diversity in terms of how they correlate with performance.

4.2 Metric-augmented MBR

The discussion in §3.4.2 shows the possibility of increasing the diversity of the utility function, $f_\theta(h_i, y_j)$, by changing the evaluation metric's model parameters, θ, as well as by introducing diversity through pseudo-references. To this end, we propose a new method called Metric-augmented Minimum Bayes Risk (MAMBR) decoding. In MAMBR, we employ different parameters for the evaluation metric to enhance the diversity of utility functions. Letting Θ be a set of model parameters, MAMBR is defined as follows:

$$\hat{h}_{\text{mambr}} = \arg\max_{h\in\mathcal{H}} \frac{1}{|\mathcal{Y}|\,|\Theta|}\sum_{\theta\in\Theta}\sum_{y\in\mathcal{Y}} f_\theta(h, y). \tag{15}$$

We train evaluation metrics with different initial random seeds to generate Θ, a set of diverse model parameters. Note that its concept of diversifying the internal decision of MBR decoding is similar to Daheim et al. (2025). The main difference is that we target θ, but they target π in Eq. (2).

5 Empirical Analysis

We conduct empirical analysis corresponding to our theoretical analysis through experiments to comprehensively understand MBR decoding.

[1] For the pseudo-bias, we used COMET (`Unbabel/wmt22-comet-da`) and BERTScore with `microsoft/deberta-xlarge-mnli` whose pearson correlations are 0.990 on the system-level task for English to German (Freitag et al., 2023b) and 0.7781 (`https://github.com/Tiiiger/bert_score`) on WMT16 to English (Bojar et al., 2016), respectively.

5.1 Overall Settings

We target three different text generation tasks, machine translation, text summarization, and image captioning to investigate the general performance of MBR decoding. In all tasks, we followed the settings of Jinnai et al. (2024b) for generating samples. We used epsilon sampling (Hewitt et al., 2022) to generate hypotheses.[2] For the generation of pseudo-references, we used various sampling approaches: beam decoding, nucleus sampling (Holtzman et al., 2020) with $p = 0.9$, ancestral sampling, top-k sampling (Fan et al., 2018) with $k = 10$, and epsilon sampling with $\epsilon = 0.02$. We set the sampling size for hypotheses to 64. We chose the sampling size for pseudo-references from {4, 8, 16, 32, 64}. We used the following datasets, models[3], and evaluation metrics for each task:

Machine Translation We used the WMT19 English to German (En-De) and WMT19 English to Russian (En-Ru) datasets (Barrault et al., 2019). We used `facebook/wmt19-en-de` for En-De and `facebook/wmt19-en-ru` for En-Ru, respectively. As the utility function and evaluation metric, we used COMET with the model `Unbabel/wmt22-comet-da`.

Text Summarization We used the SAMSum (Gliwa et al., 2019) and XSum (Narayan et al., 2018) datasets, and used `philschmid/bart-large-cnn-samsum` and `facebook/bart-large-xsum` for generation in SAMSum and XSum, respectively. As the utility function and evaluation metric, we used BERTScore (Zhang* et al., 2020) with the model `microsoft/deberta-xlarge-mnli`.

Image Captioning We used the MSCOCO dataset (Lin et al., 2014) with the split of Karpathy and Fei-Fei (2015) and the NoCaps dataset (Agrawal et al., 2019). We used `Salesforce/blip2-flan-t5-xl-coco` and `Salesforce/blip2-flan-t5-xl` for generation in MSCOCO and NoCaps, respectively. As the utility function and evaluation metric, we used BERTScore with the model `microsoft/deberta-xlarge-mnli`. We report the average scores on multiple references in both datasets.

Our implementation of the generation part is based on the released code of Jinnai et al. (2024b)[4] and the MBR decoding part is based on the toolkit, `mbrs` by Deguchi et al. (2024a)[5]. We generate samples on NVIDIA GeForce RTX 3090 and perform MBR decoding on an NVIDIA RTX A6000.

5.2 Correlation of Bias and Diversity to Performance

To verify our theoretical decomposition, we investigate the correlation of bias and diversity to performance on each dataset. For this purpose, we approximately compute the bias term by using our pseudo-bias in §4.1. Furthermore, we investigate the importance of considering the entire candidate or the best candidate.

Settings We compared the following measures based on our decomposition in Eq. (9): OVERALL BIAS is *Bias*; ONE BEST BIAS is *Diveristy* for the one best result by MBR decoding; OVERALL DIVERSITY is *Diversity*; ONE BEST DIVERSITY is *Diversity* for the one best result by MBR decoding; OVERALL MSE is $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$; ONE BEST MSE indicates errors for the one best result by MBR decoding in $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$. For the comparison, we calculated Spearman's rank correlation and Pearson correlation between these measures and the performance based on the results of five different sampling methods with five different sampling sizes on each dataset (See §5.1 for the details). Since lower bias and lower MSE are better for performance, we took their negative values in the correlation calculation. Moreover, we report averaged correlation across all datasets by Fisher z-transformation (Corey et al., 1998).

Results Figure 1 shows the correlation between the measures and performance for each dataset. These results show that MSE for both overall and one best results correlates well with the performance for each dataset in Spearman's rank correlation, indicating the importance of considering quality estimation in MBR decoding, as in Eq. (9). On the other hand, the decomposed bias and diversity show different tendencies. ONE BEST BIAS, which considers the one best result, is important for bias, whereas OVERALL DIVERSITY, which considers overall results, is important for diversity. This result is reasonable given the assumption that MBR decoding aims to select texts that are close to human-created ones. Based on this assumption, we can say that diversity supports the selection by

[2] Appendix F.1 includes the results with hypotheses generated by different sampling methods.

[3] We used all models from `https://huggingface.co/models` (Wolf et al., 2020).

[4] `https://github.com/CyberAgentAILab/model-based-mbr`

[5] `https://github.com/naist-nlp/mbrs`

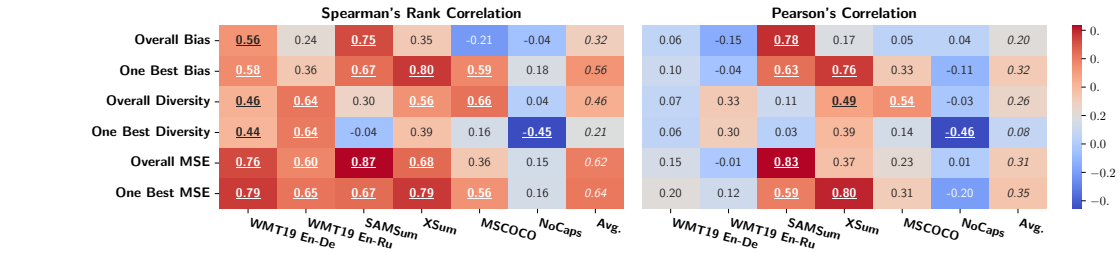

Figure 1: Correlations of each measure to the performance for each task. The underlined scores indicate statistically significant results ($p < 0.05$).[6] Note that the italic scores at *Avg.* are not the target of the significance test.

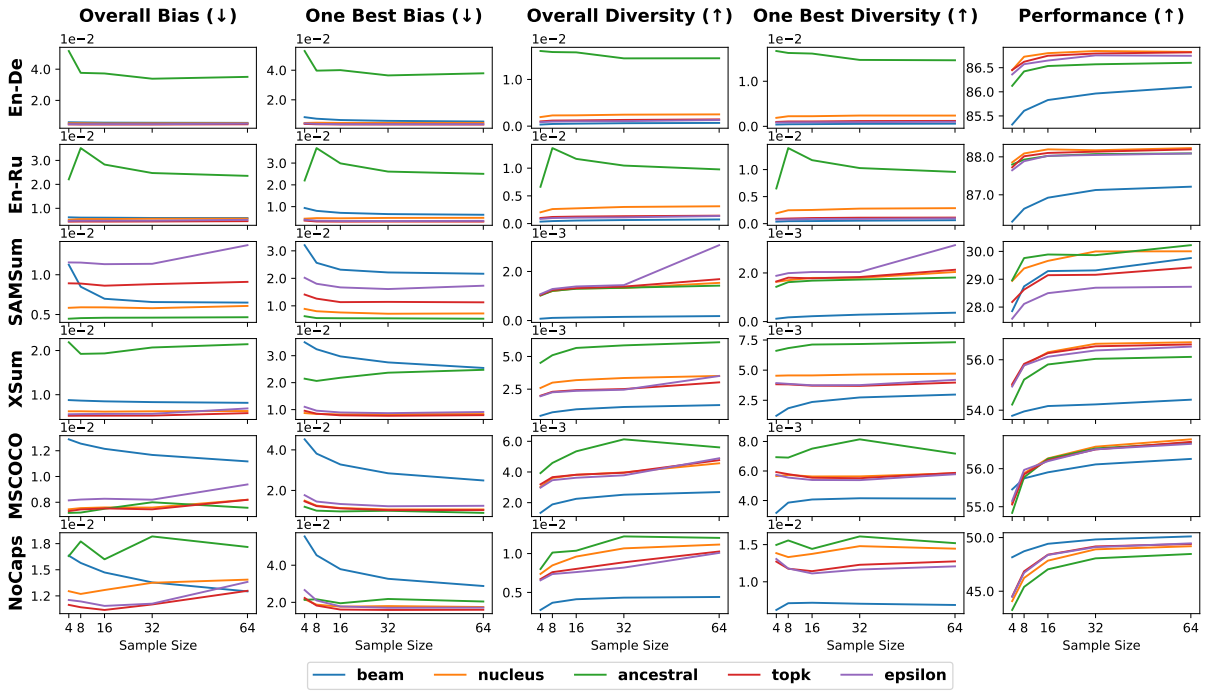

Figure 2: The relationship between bias, diversity, and performance in MBR decoding. The x-axis shows the number of used pseudo-references. (↑) indicates higher scores are better whereas (↓) indicates lower scores are better.

considering the importance of all hypotheses not covered by One Best Bias. In contrast to the results in Spearman's rank correlation, the coefficients of Pearson's correlation decrease. Based on these results, we can conclude that the measures, i.e., ONE BEST BIAS, OVERALL DIVERSITY, ONE BEST MSE and OVERALL MSE, correlate well with the rank in performance, but they are challenging to capture subtle differences of values precisely. (See Appendix C for further details.)

5.3 Bias and Diversity Trade-off

To investigate the bias-diversity trade-off in more detail, we followed the setup described in §5.2. We plotted the results for each dataset using different sampling methods in Figure 2. The results support the bias-diversity trade-off shown in Theorems 2 and 3. As a case study, while ancestral sampling exhibits the highest bias, except in the case of the SAMSum dataset, it sometimes outperforms other sampling methods owing to its greater diversity. Focusing on top-k sampling, which has the lowest bias, again excluding the SAMSum dataset, we can observe that the reduction in bias tends to limit the increase in diversity. This finding supports our previously noted bias-diversity trade-off in MBR decoding. However, as evidenced by the performance of beam decoding, which has the lowest diversity, the importance of bias and diversity varies depending on the target dataset. Therefore, while our theoretical analysis effectively explains the performance tendencies in MBR decoding, it remains essential to consider task-specific features carefully to achieve further performance improvements. (See Appendix D for further details.)

[6] We used Student's t-test (Student, 1908).

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	85.7	85.9	85.9	85.9	85.9	**87.4**	87.4	87.5	87.5	87.5
	2	85.7	**86.0**	**86.0**	85.9	85.9	**87.4**	87.4	87.5	87.5	**87.6**
	4	**85.8**	**86.0**	**86.0**	**86.0**	86.0	**87.4**	87.4	87.5	87.5	**87.6**
	8	**85.8**	**86.0**	**86.0**	**86.0**	**86.1**	**87.4**	**87.5**	**87.6**	**87.6**	**87.6**
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	28.6	29.1	29.5	29.5	29.7	**54.2**	55.2	55.7	56.0	56.1
	2	28.8	**29.6**	**29.9**	**29.9**	30.1	**54.2**	55.2	55.7	**56.1**	**56.2**
	4	**28.7**	29.5	**29.9**	29.8	**30.2**	**54.2**	55.2	**55.8**	**56.1**	**56.2**
	8	**28.7**	29.5	29.8	**29.9**	30.1	**54.3**	**55.3**	**55.8**	**56.1**	**56.2**
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	**54.9**	55.8	56.3	56.5	56.8	42.9	45.3	46.8	47.8	48.6
	2	**54.9**	55.8	56.4	56.6	56.8	43.2	45.6	47.2	48.3	48.9
	4	**54.9**	**56.0**	56.4	56.7	**56.9**	43.3	45.6	47.3	48.4	**49.0**
	8	**54.9**	**56.0**	**56.5**	**56.8**	**56.9**	**43.5**	**45.7**	**47.4**	**48.5**	**49.0**

Table 1: Results of MAMBR with ancestral sampling. Bold font indicates the best result.

5.4 Inference Scaling Laws

Figure 2 indicates improved performance by increasing the sample size, in line with Theorems 7 and 8. Since MBR decoding only considers the pairs of a pseudo-reference and hypothesis for scoring, the interaction between pseudo-references is not directly considered. This goes along with the assumption of these theorems, the conditional independence of the scores of utility functions. This is because Figure 2 shows the performance improvement our theoretical analysis can explain. Note that there is a possibility that inference methods with interaction between samples, such as universal self-consistency, do not show performance improvement like this due to the same reason.

5.5 Effectiveness of Metric-augmented MBR

We investigate the possibility of improving the performance of MAMBR in Eq. (15) by changing the automatic evaluation metric's model parameters.

Settings To prepare the set of model parameters, we trained eight models by varying their initial seeds. We trained `Unbabel/wmt22-comet-da` on the Direct Assessments (DA) task (Graham et al., 2013), using the WMT 2017 to 2020 datasets (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020) for training and the WMT 2021 dataset (Akhbardeh et al., 2021) for validation in COMET. Additionally, we trained `microsoft/deberta-large` on the MNLI dataset from GLUE (Wang et al., 2018) for BERTScore. During inference, to control model diversity, we selected the top-n models based on their proximity to the median validation scores, with n chosen from 1, 2, 4, 8. For the generation, we used ancestral sampling.

Results Tables 1 shows the MAMBR results. We observe performance improvement as the number of models increases. This suggests that MAMBR can improve performance by enhancing the diversity of evaluation metrics along with our theoretical insights. (Appendix E includes further details.)

6 Conclusion

This work provides a unified theoretical interpretation of Minimum Bayes Risk (MBR) decoding through the lens of bias-diversity decomposition. By decomposing the errors in quality estimation in MBR decoding into bias and diversity, we highlight the trade-off between improving these two factors, with an emphasis on the benefits of increasing diversity, which is behind the inference scaling laws. Our theoretical insights align with previous empirical results, and we further investigate aspects not covered by these empirical findings through the introduction of the pseudo-bias metric and MAMBR decoding. Experimental results across multiple tasks demonstrate the validity of our theoretical findings and the effectiveness of our approach in improving text generation quality. These findings bridge the gap between empirical observations and theoretical understanding of MBR decoding, offering new insights for optimizing text generation.

7 Limitation

Unlike the decomposition based on information-theoretic diversity, the bias-diversity decomposition for MBR decoding does not theoretically explain the diversity of the hypotheses and their aligneed model-side behaviors. Corresponding to this limitation, we conduct a limited empirical analysis presented in Appendix F.1, similar to previous works (Eikema and Aziz, 2020; Fernandes et al., 2022; Freitag et al., 2023a).

8 Ethical Consideration

We used GPT-4o from OpenAI in writing to check grammatical errors.

Acknowledgments

We are grateful to Graham Neubig for introducing his team's work (Bertsch et al., 2023), which reveals MBR decoding as a fundamental approach covering various decoding methods. The existence of this paper motivated us to explore MBR decoding, as we have done in this work. This work was supported by JSPS KAKENHI Grant Number JP23H03458.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, and 2 others. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. 2023. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In *Proceedings of the Big Picture Workshop*, pages 108–122, Singapore. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Hamed Bonab and Fazli Can. 2019. Less is more: A comprehensive framework for the number of components of ensemble classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2735–2745.

Hamed R. Bonab and Fazli Can. 2016. A theoretical framework on the ideal number of classifiers for online ensembles in data streams. In *Proceedings of the*

25th ACM International on Conference on Information and Knowledge Management, CIKM '16, page 2053–2056, New York, NY, USA. Association for Computing Machinery.

Sebastian Borgeaud and Guy Emerson. 2020. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 97–109, Online. Association for Computational Linguistics.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *Preprint*, arXiv:2407.21787.

Gavin Brown. 2009. An information theoretic perspective on multiple classifier systems. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, MCS '09, page 344–353, Berlin, Heidelberg. Springer-Verlag.

Gavin Brown, Jeremy L. Wyatt, and Peter Tiňo. 2005. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6(55):1621–1650.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. 2025. Simple and provable scaling laws for the test-time compute of large language models. *Preprint*, arXiv:2411.19477.

Julius Cheng and Andreas Vlachos. 2023. Faster minimum Bayes risk decoding with confidence-based pruning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12473–12480, Singapore. Association for Computational Linguistics.

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

David M Corey, William P Dunlap, and Michael J Burke. 1998. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *The Journal of general psychology*, 125(3):245–261.

Nico Daheim, Clara Meister, Thomas Möllenhoff, and Iryna Gurevych. 2025. Uncertainty-aware decoding with minimum bayes risk. In *The Thirteenth International Conference on Learning Representations*.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024a. mbrs: A library for minimum bayes risk decoding. *Preprint*, arXiv:2408.04167.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. 2024b. Centroid-based efficient minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11009–11018, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 567–575, Suntec, Singapore. Association for Computational Linguistics.

Nan Duan, Mu Li, Ming Zhou, and Lei Cui. 2011. Improving phrase extraction via MBR phrase scoring and pruning. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. 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, Melbourne, Australia. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412,

Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023a. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023b. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

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

Taisiya Glushkova, Chrysoula Zerva, and André F. T. Martins. 2023. BLEU meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 47–58, Tampere, Finland. European Association for Machine Translation.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum bayes risk decoding with multi-prompt. *Preprint*, arXiv:2407.15343.

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 degeneration. In *International Conference on Learning Representations*.

Yuu Jinnai, Ukyo Honda, Tetsuro Morimura, and Peinan Zhang. 2024a. Generating diverse and high-quality texts by minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8494–8525, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. 2024b. Model-based minimum Bayes risk decoding for text generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 22326–22347. PMLR.

Hidetaka Kamigaito, Ying Zhang, Jingun Kwon, Katsuhiko Hayashi, Manabu Okumura, and Taro Watanabe. 2025. Diversity of transformer layers: One aspect of parameter scaling laws. *Preprint*, arXiv:2505.24009.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Teun Kloek and Herman K Van Dijk. 1978. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pages 1–19.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Shankar Kumar and William Byrne. 2002. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Huan Liu, Amit Mandvikar, and Jigar Mody. 2004. An empirical study of building compact ensembles. In *Advances in Web-Age Information Management: 5th International Conference, WAIM 2004, Dalian, China, July 15-17, 2004 5*, pages 622–627. Springer.

Boxuan Lyu, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2025. Unveiling the power of source: Source-based minimum bayes risk decoding for neural machine translation. *Preprint*, arXiv:2406.11632.

Tom Mitchell. 1997. Introduction to machine learning. *Machine learning*, 7:2–5.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

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

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality estimation using minimum Bayes risk. In *Proceedings of the Eighth Conference on Machine Translation*, pages 806–811, Singapore. Association for Computational Linguistics.

Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. 2024. On the true distribution approximation of minimum Bayes-risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 459–468, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

CP Robert. 1999. Monte carlo statistical methods.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics.

Student. 1908. Probable error of a correlation coefficient. *Biometrika*, pages 302–310.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. 2024. Efficient minimum bayes risk decoding using low-rank matrix completion algorithms. *Preprint*, arXiv:2406.02832.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. 2024. A unified theory of diversity in ensemble learning. *J. Mach. Learn. Res.*, 24(1).

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

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*.

Zhi-Hua Zhou and Nan Li. 2010. Multi-information ensemble diversity. In *Proceedings of the 9th International Conference on Multiple Classifier Systems*, MCS'10, page 134–144, Berlin, Heidelberg. Springer-Verlag.

A Proofs

A.1 Proof for Theorem 1

First, we can decompose $(\hat{u}_i - \bar{u}_i)^2$ as follows:

$$
\begin{aligned}
&(\hat{u}_i - \bar{u}_i)^2 && (16)\\
=&(\hat{u}_i)^2 - 2\hat{u}_i\bar{u}_i + (\bar{u}_i)^2 && (17)\\
=&(\hat{u}_i)^2 - 2\hat{u}_i\bar{u}_i + 2(\bar{u}_i)^2 - (\bar{u}_i)^2 && (18)\\
=&(\hat{u}_i)^2 - 2\hat{u}_i\bar{u}_i + 2\bar{u}_i\bar{u}_i - (\bar{u}_i)^2 && (19)\\
=&(\hat{u}_i)^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\hat{u}_i u_i^j + \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\bar{u}_i u_i^j - (\bar{u}_i)^2 && (20)\\
=&\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i)^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\hat{u}_i u_i^j + \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|} 2\bar{u}_i u_i^j - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i)^2 && (21)\\
=&\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i)^2 - 2\hat{u}_i u_i^j + 2\bar{u}_i u_i^j - (\bar{u}_i)^2) && (22)\\
=&\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i)^2 - 2\hat{u}_i u_i^j + (u_i^j)^2 - (u_i^j)^2 + 2\bar{u}_i u_i^j - (\bar{u}_i)^2) && (23)\\
=&\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i)^2 - 2\hat{u}_i u_i^j + (u_i^j)^2 - ((u_i^j)^2 - 2\bar{u}_i u_i^j + (\bar{u}_i)^2)) && (24)\\
=&\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}((\hat{u}_i - u_i^j)^2 - (\bar{u}_i - u_i^j)^2) && (25)\\
=&\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2 && (26)
\end{aligned}
$$

By utilizing Eq. (26), we can further decompose $MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}})$ as follows:

$$
\begin{aligned}
&MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) && (27)\\
=&\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}(\hat{u}_i - \bar{u}_i)^2 && (28)\\
=&\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\Big(\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2 - \frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2\Big) && (29)\\
=&\underbrace{\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\hat{u}_i - f_\theta(h_i, y_j))^2}_{\text{Bias}} - \underbrace{\frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|}\frac{1}{|\mathcal{Y}|}\sum_{j=1}^{|\mathcal{Y}|}(\bar{u}_i - f_\theta(h_i, y_j))^2}_{\text{Diversity}} && (30)
\end{aligned}
$$

Finally, Theorem 1 is proved.

A.2 Proof for Theorem 2

In Eq. (9), when *Bias* becomes zero, $f_\theta(h_i, y_j)$ becomes $\hat{u}_i$ in any j. Because $\bar{u}_i$ is an average of $f_\theta(h_i, y_j)$ for all j, $\bar{u}_i$ becomes $\hat{u}_i$ in this condition. Since that means $(\bar{u}_i - f_\theta(h_i, y_j))^2$ becomes zero and then *Diversity* becomes zero. Therefore, Theorem 2 is proved.

A.3 Proof for Theorem 3

See Brown et al. (2005) for the proof of the decomposition. Since this decomposition is applicable when Eq. (9) holds, we can apply this decomposition to our analysis.

A.4 Proof for Theorem 4

See Zhou and Li (2010) for the decomposition based on the information-theoretic diversity. Since there is no restriction for the used variables, we can apply their decomposition to our analysis.

A.5 Proof for Theorem 5

See Kamigaito et al. (2025) for the proof. Since the monotonicity is generally applicable without limitation, we can apply their proof to our analysis.

A.6 Proof for Theorem 6

See Kamigaito et al. (2025) for the proof. Similar to Appendix A.5, we can apply their proof to our analysis.

A.7 Proof for Theorem 7

See Kamigaito et al. (2025) for the proof. Since conditional independence is an assumption, we can apply their proof to our analysis.

A.8 Proof for Theorem 8

See Kamigaito et al. (2025) for the proof. Similar to Appendix A.7, we can apply their proof to our analysis.

B Interpretation as Ensemble Learning

When $|\mathcal{Y}|$ is large enough to satisfy the law of large numbers, we can induce the following expectation in MBR decoding by using a model's prediction, $P(y|x)$:

$$\arg\max_{h \in \mathcal{H}} \sum_{y \in \Omega} f_\theta(h, y) P(y|x) \tag{31}$$

$$= \arg\max_{h \in \mathcal{H}} \mathbb{E}_{P(y|x)}[f_\theta(h, y)] \tag{32}$$

$$\approx \arg\max_{h \in \mathcal{H}} \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} f_\theta(h, y), \quad y_1, \cdots, y_{|\mathcal{Y}|} \sim P(y|x) \tag{33}$$

Since this expectation is based on $P(y|x)$, we can understand the importance of increasing the number of pseudo-references to induce a reliable $P(y|x)$.

Theorem 9. *When $f_\theta(h, y)$ is normalized as a probability $P_\theta(h|y)$, Eq. (31) is equivalent to the Bayes Optimal Classifier (BOC) in Mitchell (1997).*

Proof. Self-evident by the following reformulation:

$$\arg\max_{h \in \mathcal{H}} \sum_{y \in \Omega} f_\theta(h, y) P(y|x) = \arg\max_{h \in \mathcal{H}} \sum_{y \in \Omega} P_\theta(h|y) P(y|x) \tag{34}$$

□

Theorem 10. *When $f_\theta(h, y)$ is normalized as a probability $P_\theta(h|y)$, Eq. (33) is equivalent to the Gibbs algorithm in Mitchell (1997) that approximates BOC by sampling.*

Proof. Self-evident by the following reformulation:

$$\arg\max_{h\in\mathcal{H}} \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_\theta(h,y), \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim P(y|x) \tag{35}$$

$$=\arg\max_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} P_\theta(h|y), \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim P(y|x) \tag{36}$$

□

Hence, we can understand that MBR decoding represented as Eqs. (31) and (33) approximates the ensemble learning method, BOC. In this interpretation, since $P(y|x)$ is a prior of BOC, we can also understand that MBR approximately uses the model-predicted probability as its prior.

When pruning unnecessary y in the BOC formulation of Eq. (34), because the sum of $P_\theta(h|y)$ for all h is always 1, we can determine the importance of y based solely on $P(y|x)$. Since we can arbitrarily choose $P(y|x)$ during sampling, we understand that pruning methods select the importance of each y as a prior in BOC. Note that utility functions are not always normalized; therefore, there is a gap between this interpretation and the actual MBR decoding. Addressing this gap remains an open problem.

In practice, directly drawing samples from $P(y|x)$ is intractable. Therefore, we must use approximate search methods, which are commonly influenced by left-to-right decoding and threshold values. These factors can lead to unreachable states and biases, as seen in greedy or beam decoding and other sampling approaches. Letting $P'(y|x)$ denote the model's prediction with the approximate search, we can similarly induce the following expectation:

$$\arg\max_{h\in\mathcal{H}} \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_\theta(h,y), \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim P'(y|x) \tag{37}$$

$$\approx\arg\max_{h\in\mathcal{H}} \mathbb{E}_{P'(y|x)}[f_\theta(h,y)] \tag{38}$$

Unfortunately, due to $P'(y|x)$, Eq. (38) deviates from Eq. (32). To precisely predict Eq. (31) using samples from $P'(y|x)$, we can consider the following theorem:

Theorem 11. *When $|\mathcal{Y}|$ is large enough to satisfy the law of large numbers, by using importance sampling, we can induce Eq. (32) from $P'(y|x)$.*

Proof.

$$\arg\max_{h\in\mathcal{H}} \mathbb{E}_{P(y|x)}[f_\theta(h,y)] \tag{39}$$

$$=\arg\max_{h\in\mathcal{H}} \sum_{y} P(y|x) f_\theta(h,y) \tag{40}$$

$$=\arg\max_{h\in\mathcal{H}} \sum_{y} P(y|x) f_\theta(h,y)\frac{P'(y|x)}{P'(y|x)} \tag{41}$$

$$=\arg\max_{h\in\mathcal{H}} \sum_{y} P'(y|x) f_\theta(h,y)\frac{P(y|x)}{P'(y|x)} \tag{42}$$

$$\approx\arg\max_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} f_\theta(h,y)\frac{P(y|x)}{P'(y|x)}, \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim P'(y|x) \tag{43}$$

□

Apart from the fact that even precisely calculating $P'(y|x)$ is also difficult, we can induce the following theorem:

Theorem 12. *When $|\mathcal{Y}|$ is large enough to satisfy the law of large numbers and $P'(y|x)$ equals a discrete uniform distribution $\mathcal{U}(0,|\mathcal{Y}|)$, Eq. (43) is equivalent to Model-based MBR (MBMBR) of Jinnai et al. (2024b).*

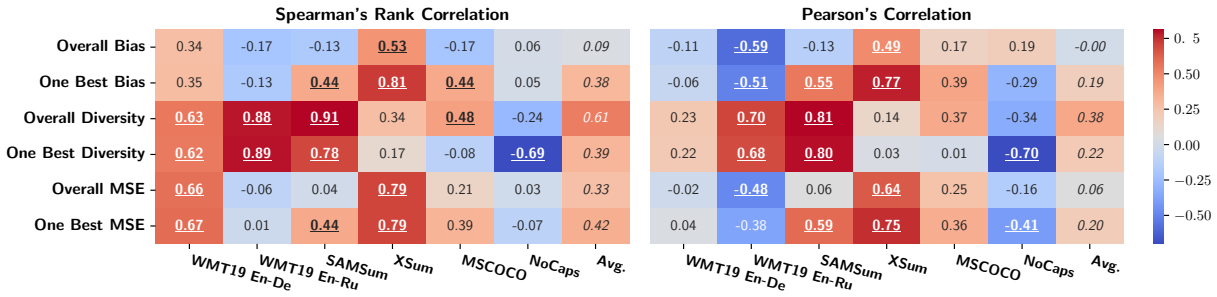

Figure 3: Correlation between measures in our decomposition and performance for each dataset when using different metrics in decoding and performance evaluation. The notations are the same as Figure 1.

Proof.

$$\arg\max_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} f_\theta(h,y)\frac{P(y|x)}{P'(y|x)}, \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim P'(y|x) \tag{44}$$

$$=\arg\max_{h\in\mathcal{H}} \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} f_\theta(h,y)P(y|x), \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim \mathcal{U}(0,|\mathcal{Y}|) \tag{45}$$

$$=\arg\max_{h\in\mathcal{H}} \sum_{y\in\mathcal{Y}} f_\theta(h,y)P(y|x), \quad y_1,\cdots,y_{|\mathcal{Y}|}\sim \mathcal{U}(0,|\mathcal{Y}|) \tag{46}$$

□

From Theorem 12, we can understand that MBMBR is an effective approach when sampling methods are unreliable. Based on the interpretation from the viewpoint of BOC, Eq. (46) estimates the importance for each y through prior $P(y|x)$, which can be used for pruning y.

Even though our interpretation can explain the pruning of pseudo-references based on priors in BOC, pruning hypotheses are out-of-scope of this interpretation.

C Correlation of Bias and Diversity to Performance

We further investigate whether our analysis in §5.2 is consistent when metrics used in MBR decoding and performance evaluation are different.

Settings Based on the inherited settings from §5.2, we changed the performance evaluation metrics, COMET and BERTScore to BLEURT (Sellam et al., 2020) and chrF++ (Popović, 2015, 2017). We used BLEURT on single sentence generation tasks, WMT19 En-De and En-Ru, XSum, MSCOCO, and NoCaps. Since SAMSum is a multiple-sentence generation task and BLEURT cannot handle it, we used chrF++ instead.

Results Figure 3 shows the correlation. Similar to the results in §5.2, the measures, i.e., One Best Bias, Overall Diversity, and One Best MSE in Spearman's rank correlation correlate well with the rank in performance, even though these correlation values are degraded by different evaluation metrics from decoding time. The lower correlation values in Pearson's correlation than Spearman's rank correlation also show similar tendencies in §5.2 and indicate the difficulty of precisely estimating the performance values from these measures. From these results, we can confirm that correlation tendencies are consistent when changing the performance evaluation metrics.

D Bias and Diversity Trade-off

Similar to Appendix C, we further investigate whether our analysis in §5.3 is consistent when metrics used in MBR decoding and performance evaluation are different.

Settings We inherited the setting of Appendix C. Thus, COMET and BERTScore used in MBR decoding are replaced with BLEURT and chrF++ in performance evaluation.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	238	211	209	221	274	214	212	216	236	282
	2	341	466	330	359	472	324	331	341	367	457
	4	592	554	682	715	754	562	608	578	621	866
	8	1,129	1,067	1,059	1,119	1,549	1,205	1,018	1,014	1,158	1,531
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	390	424	544	748	1,093	3,612	4,249	4,451	5,947	10,936
	2	592	717	902	1,300	1,945	5,267	5,978	7,201	10,478	20,404
	4	1,149	1,122	1,388	1,994	4,005	8,805	11,584	15,265	19,438	38,137
	8	1,838	2,100	2,626	3,764	7,821	16,397	19,392	28,844	37,442	60,800
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	1,209	1,143	1,582	1,906	2,849	1,235	1,179	1,406	1,831	3,453
	2	1,739	1,772	2,580	3,261	5,144	1,574	1,647	2,172	2,935	5,636
	4	2,439	3,035	4,541	5,986	9,758	2,644	3,330	4,170	5,201	8,916
	8	4,316	6,099	7,557	11,397	19,108	3,784	5,432	6,666	10,210	17,052

Table 2: Time usages (seconds) by MAMBR in each setting with ancestral sampling.

Results Figure 4 shows the results. We can see the changed performances in the subfigures of the rightmost column. The entire tendencies of beam decoding are almost the same as Figure 2, excluding the case of the performance drop in SAMSum, whose evaluation metric is changed from BERTScore to chrF++. However, this behavior is reasonable considering the highest `One Best Bias` and lowest `Overall Diversity` of beam decoding in SAMSum. This result shows the possibility of adopting bias and diversity in a metric to estimate performance in other evaluation metrics. On the other hand, these relationships are not always consistent, as represented by the uncorrelated values on NoCaps that permit diversified generation, as shown by its 10 gold references.

E Detailed Results of MAMBR

E.1 Computational Costs

The computational costs of MAMBR are proportional to the number of used models. Tables 2 and 3 show the computational cost of running the settings corresponding to Table 1. We used one NVIDIA RTX A6000 for the measurement.

E.2 Results on other Sampling Strategies

Tables 4 and 5 show the results of MAMBR with epsilon sampling and beam decoding, respectively. In the following discussion, we also consider Table 1. In ancestral and epsilon sampling, the best and moderately diversified sampling strategies (as shown in Figure 2), we observe performance improvement as the number of models increases. On the other hand, in the lowest diversity method, beam decoding, performance improvement is limited. These results suggest that MAMBR can improve performance by enhancing the diversity of evaluation metrics, although the diversity of the sampling strategy itself remains important.

E.3 Bias and Diversity of MAMBR

Figures 5, 6, and 7 show the bias and diversity corresponding to the results in Tables 1, 4, and 5, respectively. The results show that MAMBR actually increases the diversities in WMT19 En-De and En-Ru and SAMSum but not in the other dataset. Thus, this improvement depends on the datasets. On the other hand, we can see the improvement of bias in some cases. This is reasonable because using multiple metric models itself is an ensembling approach and can contribute to performance improvement.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	1,501	1,501	1,501	1,501	1,528	1,533	1,533	1,533	1,533	1,566
	2	2,610	2,610	2,610	2,610	2,636	2,642	2,642	2,642	2,642	2,674
	4	4,828	4828	4,828	4,828	4,855	4,859	4,859	4,859	4,859	4,891
	8	9,261	9,261	9,261	9,261	9,289	9,293	9,293	9,293	9,293	9,326
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	2,525	2,529	2,525	2,526	2,524	2,104	2,102	2,106	2,102	2,106
	2	3,734	3,738	3,734	3,736	3,734	3,313	3,311	3,315	3,311	3,315
	4	6,152	6,156	6,153	6,154	6,152	5,732	5,730	5,733	5,730	5,733
	8	10,990	10,994	10,990	10,992	10,990	10,569	10,567	10,570	10,567	10,571
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	1,621	1,621	1,621	1,621	1,621	1,654	1,654	1,655	1,654	1,655
	2	2,831	2,831	2,831	2,831	2,831	2,863	2,864	2,863	2,864	2,863
	4	5,249	5,249	5,249	5,249	5,249	5,282	5,282	5,282	5,283	5,282
	8	10,086	10,086	10,086	10,086	10,086	10,119	10,119	10,119	10,120	10,119

Table 3: GPU memory usages (MB) by MAMBR in each setting with ancestral sampling.

F Experimental Results on the First 1000 Examples

To consider more detailed configurations and reveal the possibility of a more efficient investigation, we conducted an additional evaluation using only the first 1000 examples for each dataset based on the setting of Jinnai et al. (2024b).

F.1 Hypotheses generated by different sampling strategies

Figures 8 to 12 present the bias and diversity decomposition plots for different hypothesis generation strategies. The results indicate that differences in the generated hypotheses influence performance in some cases, whereas the overall tendencies of the sampling strategy used for generating pseudo-references remain similar despite these variations.

F.2 MAMBR

Tables 6 to 8 show the MAMBR results for the first 1000 lines. From these results, we observe a similar trend to those obtained when the dataset is fully used, as described in §5.5. Similarly, Figures 13 to 15 demonstrate that the results are nearly identical to those obtained when the dataset is fully utilized.

G Reproducibility Statement

We performed our experiments by running publicly available models, facebook/wmt19-en-de (Apache license 2.0), facebook/wmt19-en-ru (Apache license 2.0), philschmid/bart-large-cnn-samsum (MIT License), facebook/bart-large-xsum (MIT License), Salesforce/blip2-flan-t5-xl-coco (MIT License), and Salesforce/blip2-flan-t5-xl (MIT License) in HuggingFace Transformers (Wolf et al., 2020) on the publicly available datasets, WMT19 English to German (Barrault et al., 2019), WMT19 English to Russian (Barrault et al., 2019), SAMSum (Gliwa et al., 2019), XSum (Narayan et al., 2018), MSCOCO (Lin et al., 2014; Karpathy and Fei-Fei, 2015), and NoCaps (Agrawal et al., 2019), respectively with utilizing the publicly available MBR decoding toolkit, `mbrs` (Deguchi et al., 2024a) as described in §5.1.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	85.9	86.1	86.2	86.2	86.2	87.3	87.6	**87.7**	**87.7**	87.7
	2	**86.0**	86.1	86.1	86.2	86.3	87.3	87.6	**87.7**	**87.7**	87.7
	4	**86.0**	86.1	86.2	86.2	86.3	**87.4**	87.6	**87.7**	**87.7**	87.7
	8	**86.0**	**86.2**	**86.3**	**86.3**	**86.4**	**87.4**	**87.7**	**87.7**	**87.7**	**87.8**
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	27.5	27.9	28.3	28.4	28.5	**54.9**	55.7	56.1	56.3	56.4
	2	**27.7**	28.1	28.5	**28.6**	**28.7**	**54.9**	55.7	56.1	56.3	**56.5**
	4	**27.7**	**28.2**	28.5	**28.6**	28.6	**54.9**	55.7	56.1	**56.4**	**56.5**
	8	27.6	**28.2**	**28.6**	**28.6**	**28.7**	**54.9**	**55.8**	**56.2**	**56.4**	**56.5**
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	55.2	55.9	**56.3**	56.5	56.7	44.4	46.7	48.5	49.1	49.5
	2	55.2	55.9	**56.3**	56.5	56.7	44.4	46.8	48.6	49.2	49.6
	4	55.2	56.0	**56.3**	**56.6**	**56.8**	44.5	46.9	**48.7**	49.3	**49.7**
	8	**55.3**	**56.1**	**56.3**	**56.6**	**56.8**	**44.6**	**47.0**	**48.7**	**49.4**	**49.7**

Table 4: Results of MAMBR with epsilon sampling. Notations are the same as Table 1.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	85.2	85.4	85.6	85.7	85.8	86.5	86.8	**87.0**	**87.1**	87.1
	2	**85.3**	**85.5**	**85.7**	**85.8**	85.8	86.5	86.8	86.9	**87.1**	87.1
	4	**85.3**	**85.5**	**85.7**	**85.8**	85.8	**86.6**	**86.9**	**87.0**	**87.1**	**87.2**
	8	**85.3**	**85.5**	**85.7**	**85.8**	**85.9**	86.5	86.8	**87.0**	**87.1**	**87.2**
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	27.6	28.7	**29.2**	29.3	**29.7**	**53.8**	**54.0**	**54.2**	**54.2**	**54.4**
	2	**27.8**	**28.9**	**29.2**	**29.4**	**29.7**	**53.8**	53.9	54.1	**54.2**	**54.4**
	4	**27.8**	**28.9**	**29.2**	**29.4**	**29.7**	**53.8**	**54.0**	**54.2**	**54.2**	**54.4**
	8	**27.8**	**28.9**	**29.2**	**29.4**	**29.7**	**53.8**	**54.0**	**54.2**	**54.2**	**54.4**
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	55.4	55.7	**55.9**	**56.1**	**56.3**	**48.2**	**48.8**	49.4	**49.9**	50.2
	2	55.4	**55.8**	**55.9**	**56.1**	**56.3**	**48.2**	**48.8**	49.4	**49.9**	**50.3**
	4	**55.5**	55.7	**55.9**	**56.1**	**56.3**	**48.2**	**48.8**	**49.5**	**49.9**	50.2
	8	**55.5**	55.7	**55.9**	**56.1**	**56.3**	**48.2**	**48.8**	**49.5**	**49.9**	50.2

Table 5: Results of MAMBR with beam decoding. Notations are the same as Table 1.

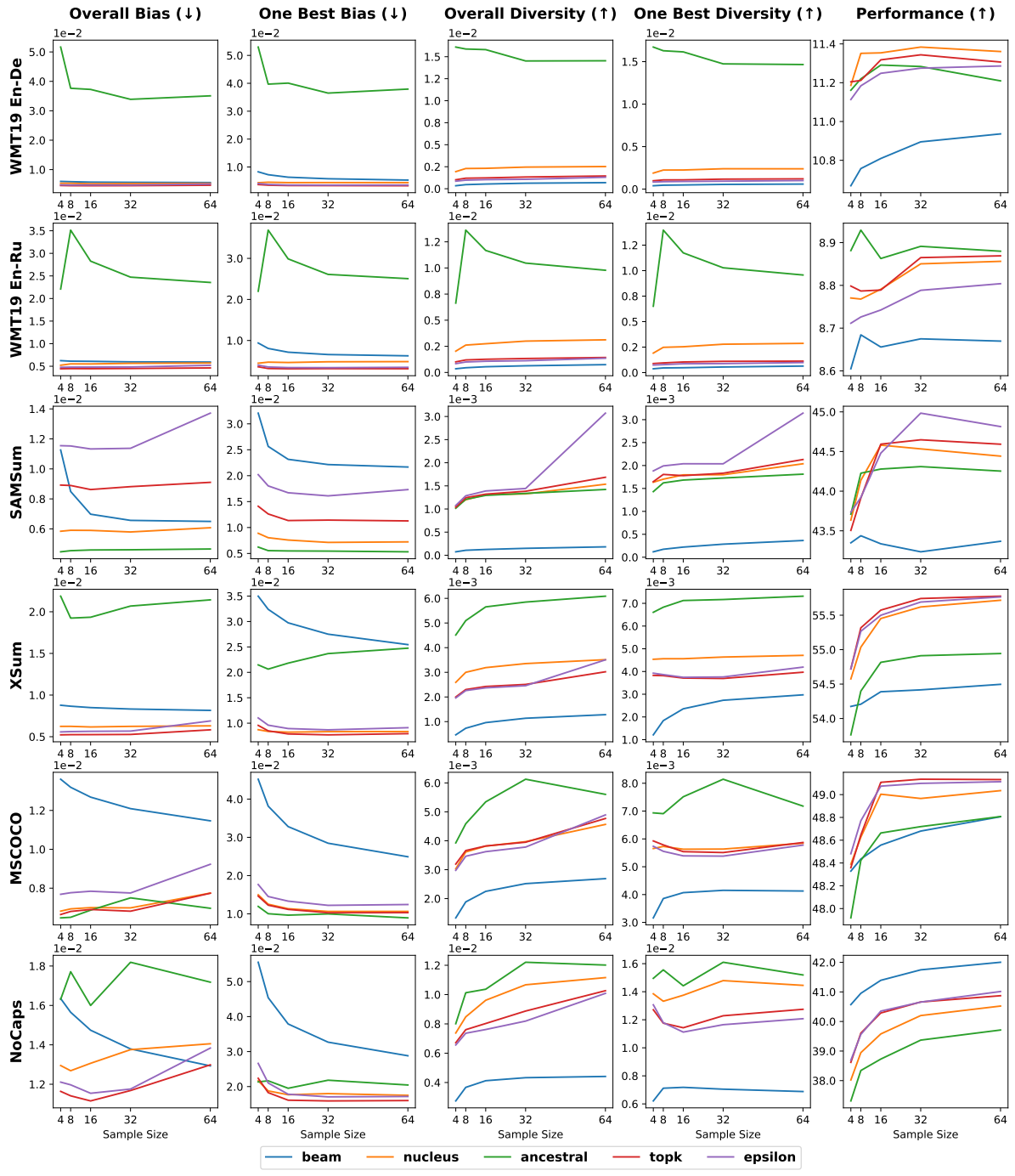

Figure 4: The relationship between bias, diversity, and performance in MBR decoding when using different metrics in decoding and performance evaluation. The notations are the same as Figure 2.

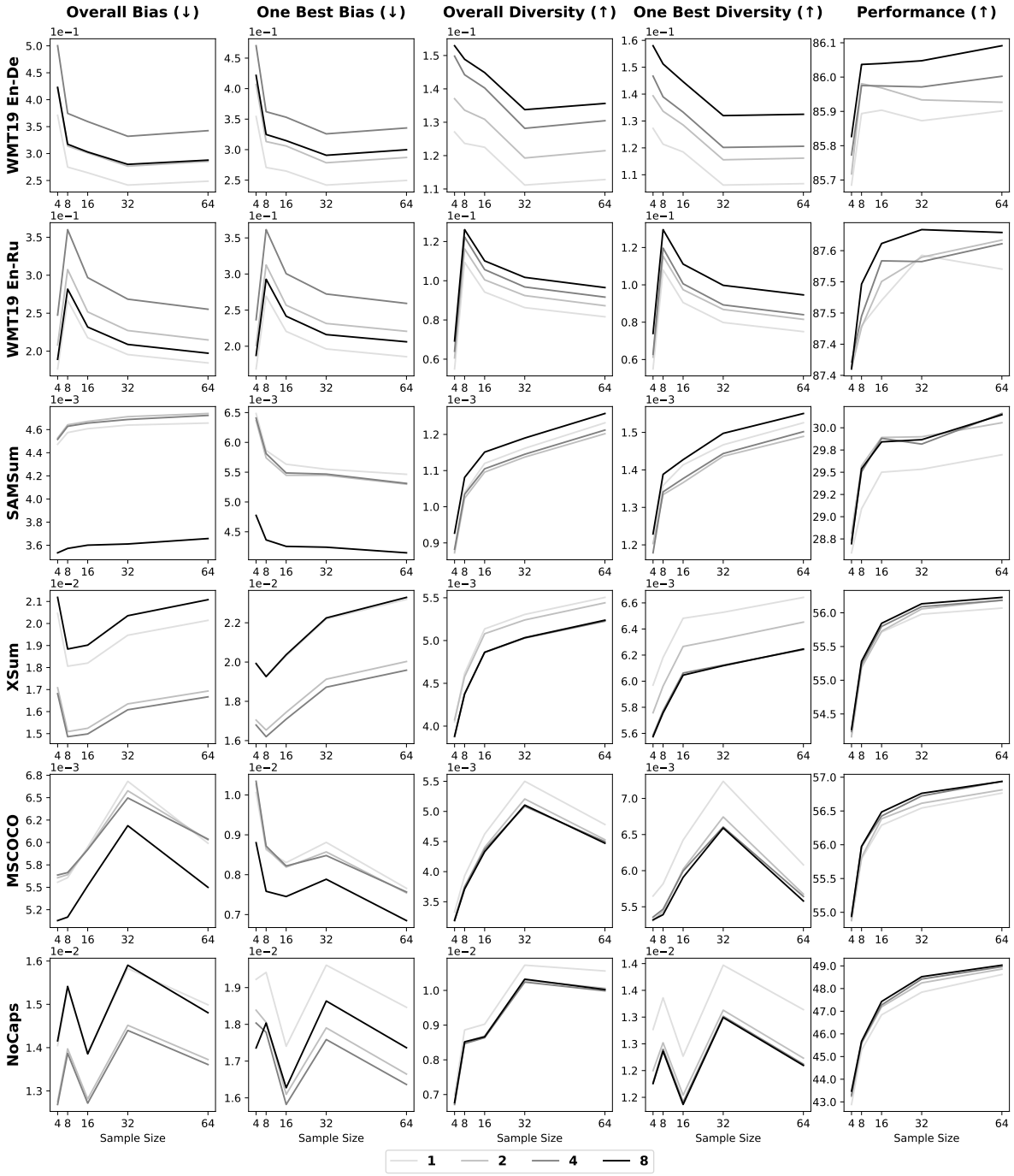

Figure 5: The relationship between bias, diversity, and performance in MAMBR decoding with pseudo-references generated by ancestral sampling. The lines indicate the score for each number of used metric models. Other notations are the same as Figure 2.

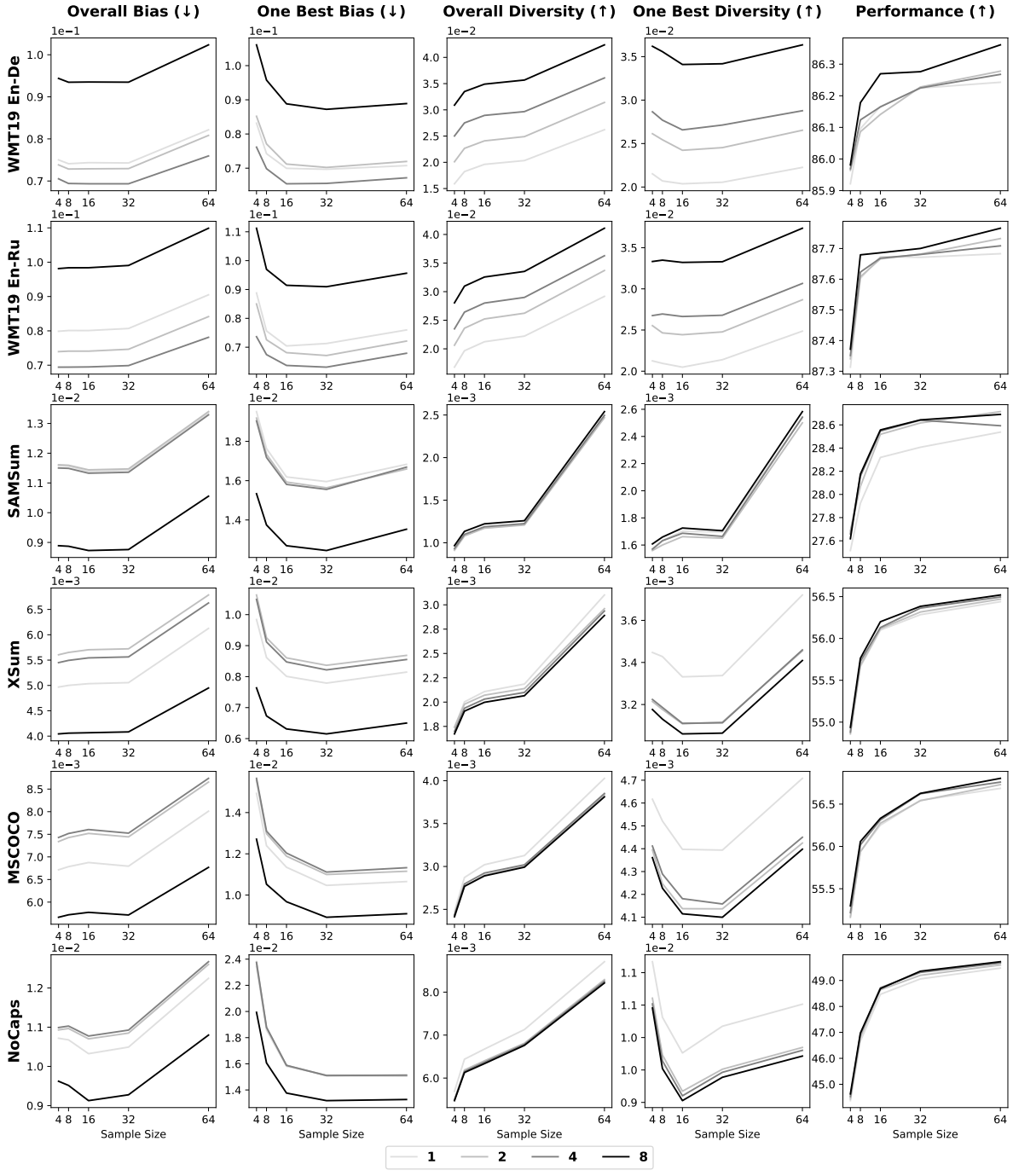

Figure 6: The relationship between bias, diversity, and performance in MAMBR decoding with pseudo-references generated by epsilon sampling. The notations are the same as Figure 5.

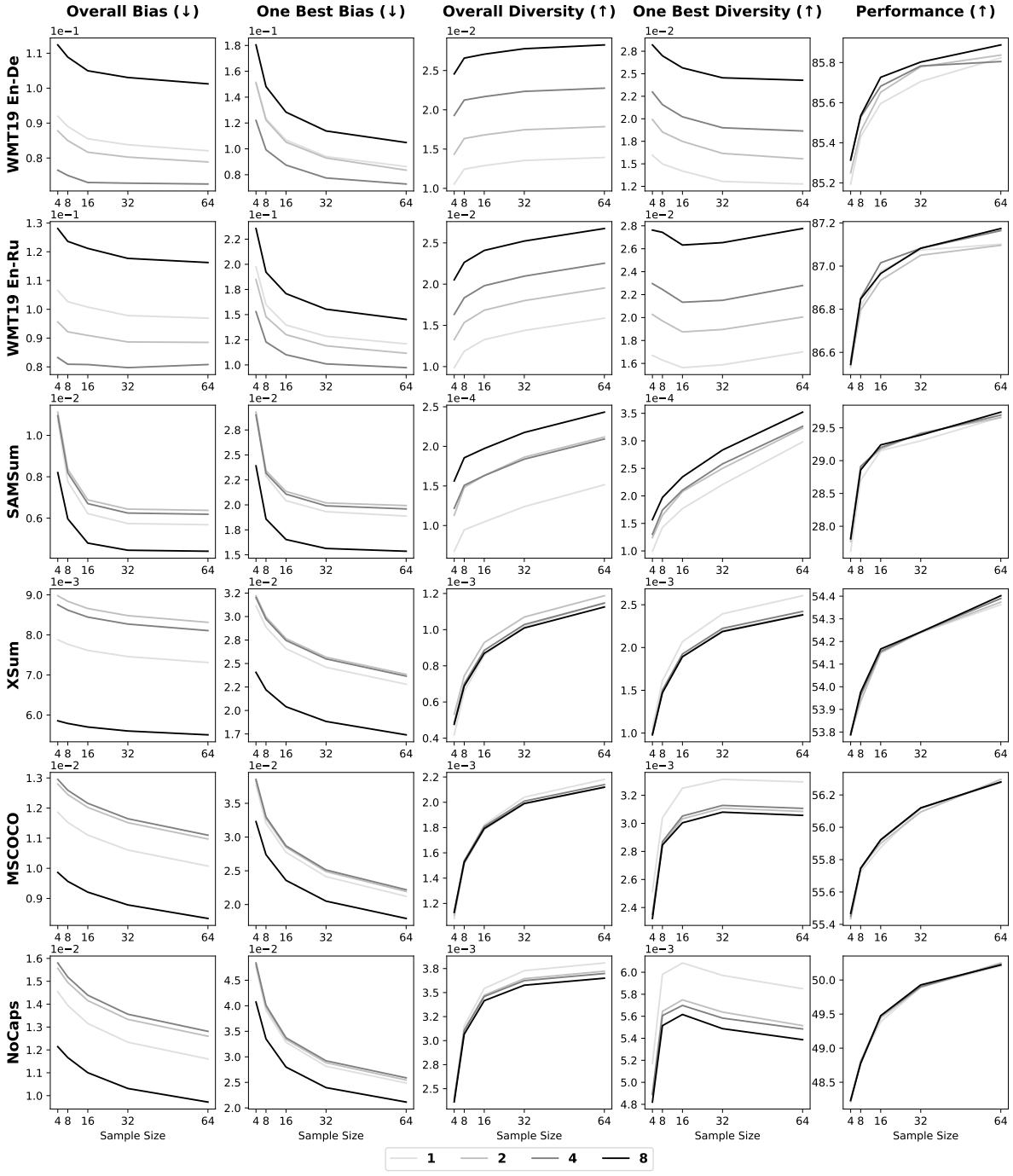

Figure 7: The relationship between bias, diversity, and performance in MAMBR decoding with pseudo-references generated by beam decoding. The notations are the same as Figure 5.

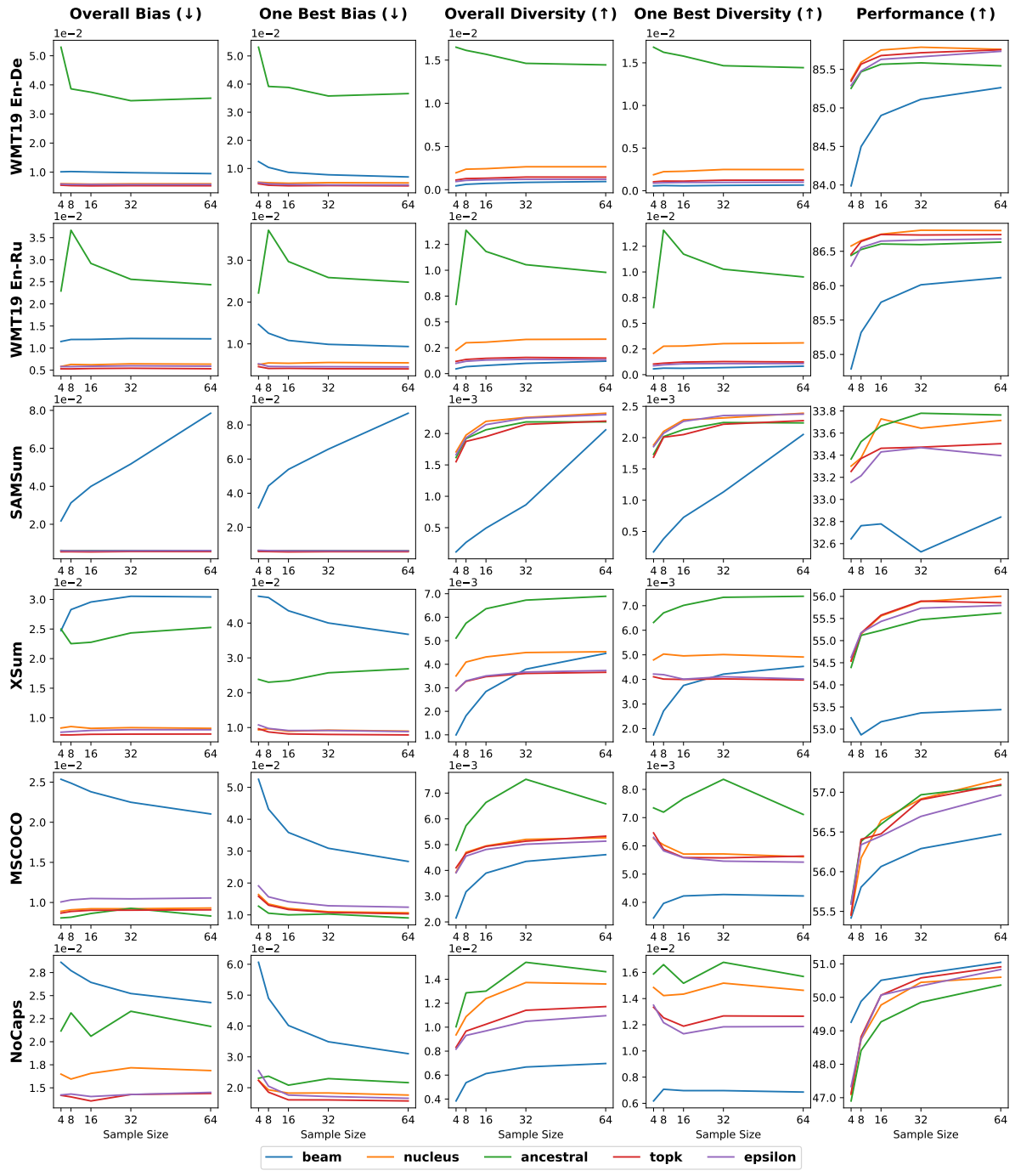

Figure 8: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with hypotheses generated by beam decoding. The notations are the same as Figure 2.

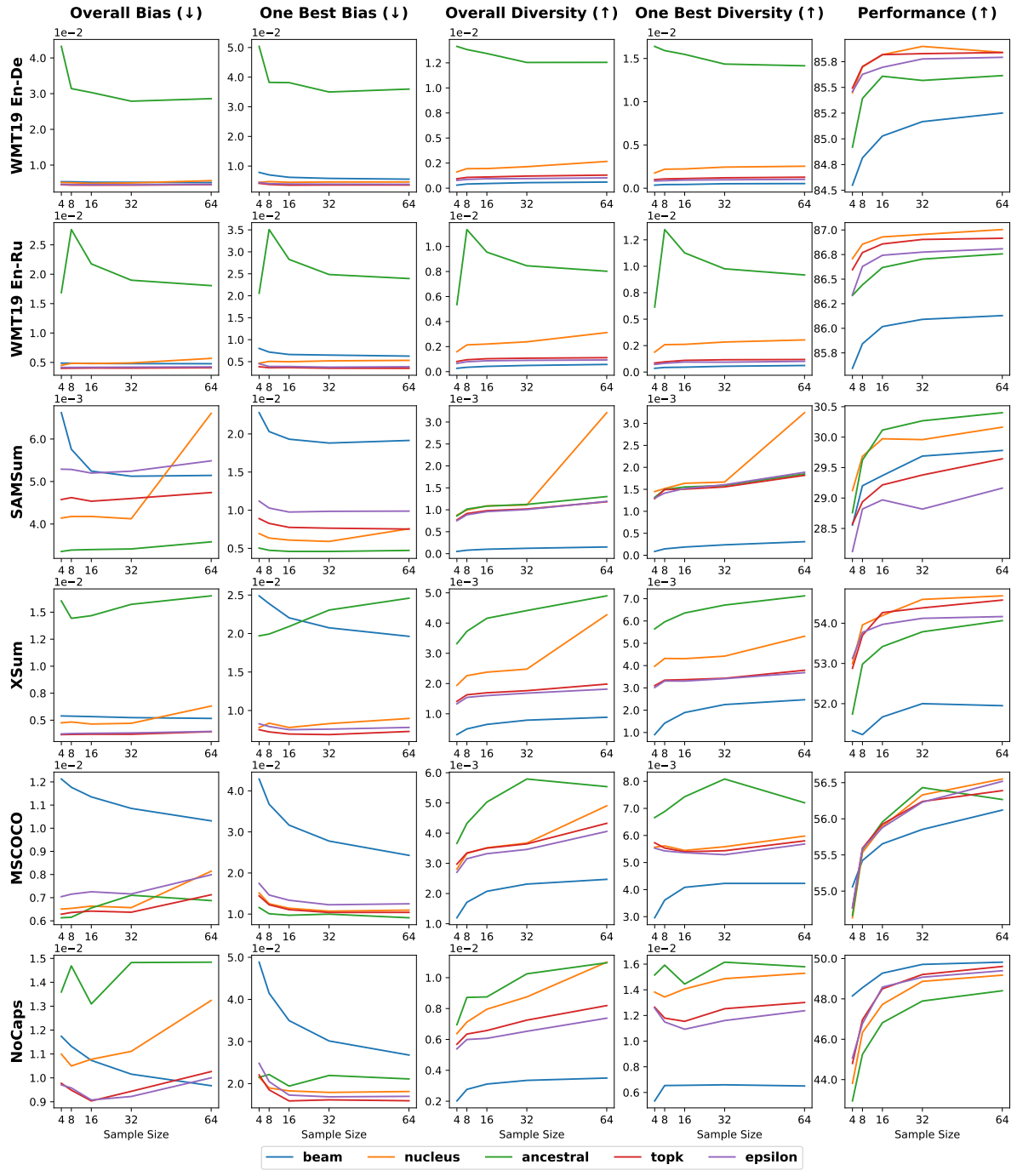

Figure 9: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with hypotheses generated by nucleus sampling. The notations are the same as Figure 2.

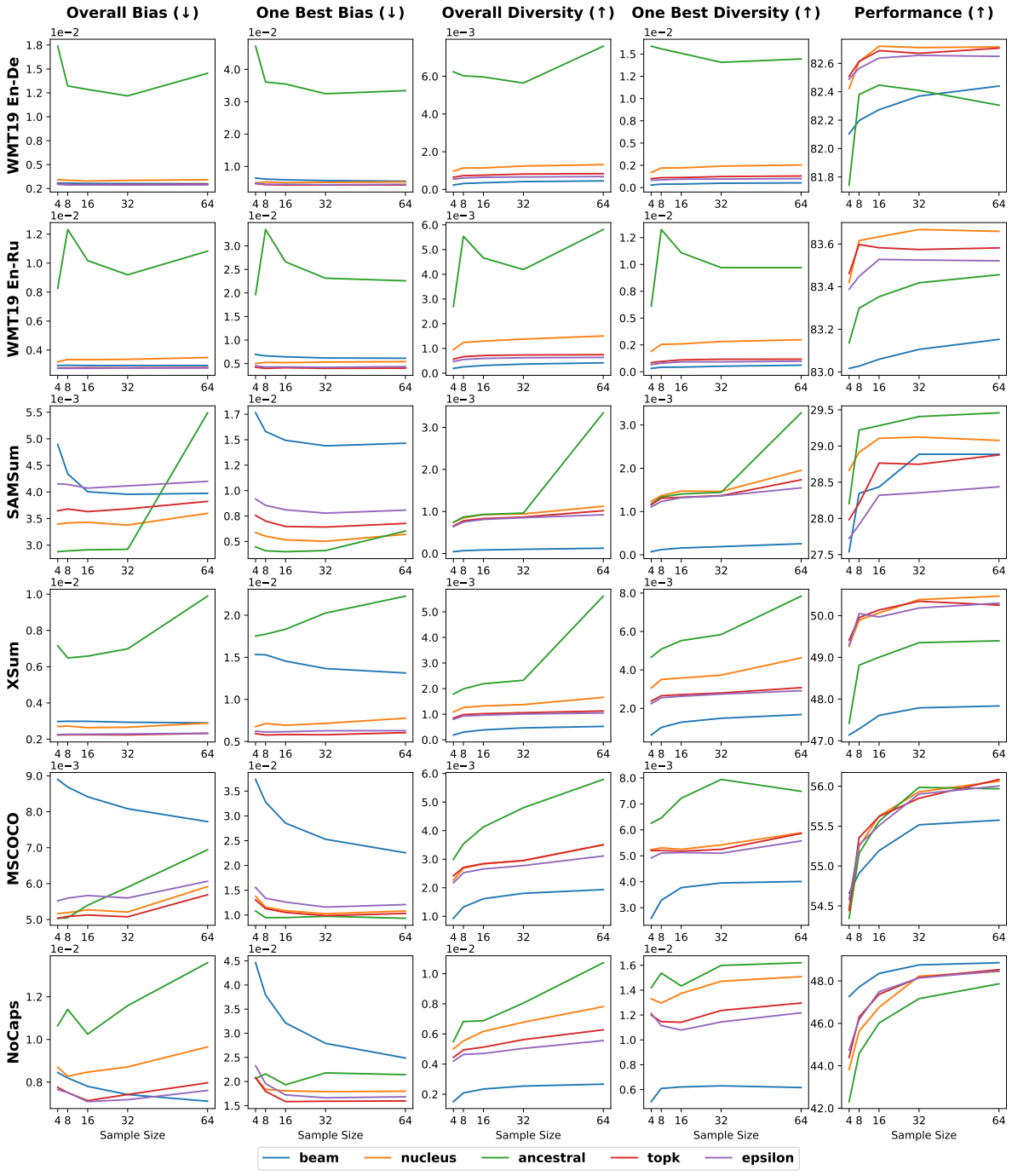

Figure 10: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with hypotheses generated by ancestral sampling. The notations are the same as Figure 2.

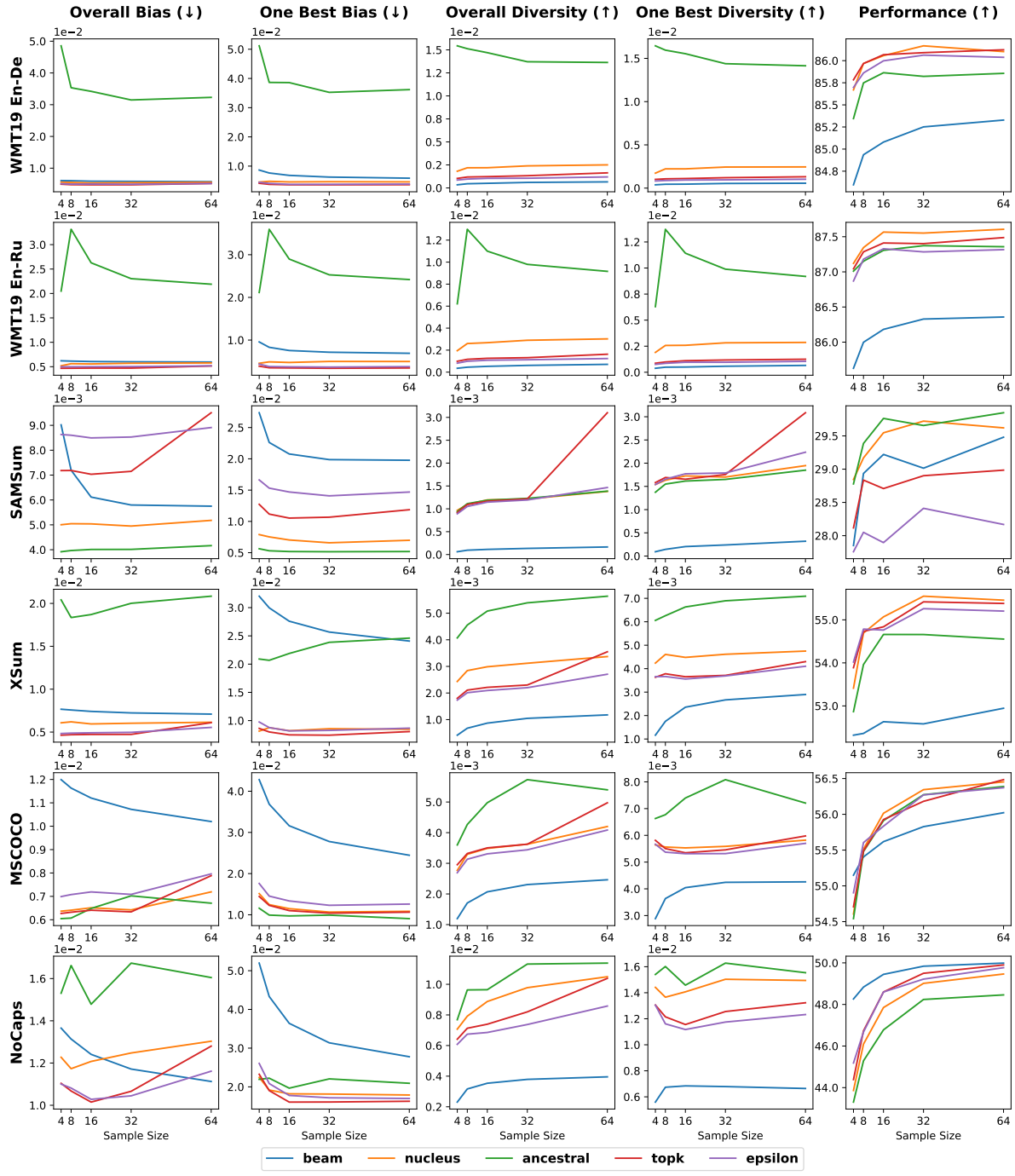

Figure 11: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with hypotheses generated by top-k sampling. The notations are the same as Figure 2.

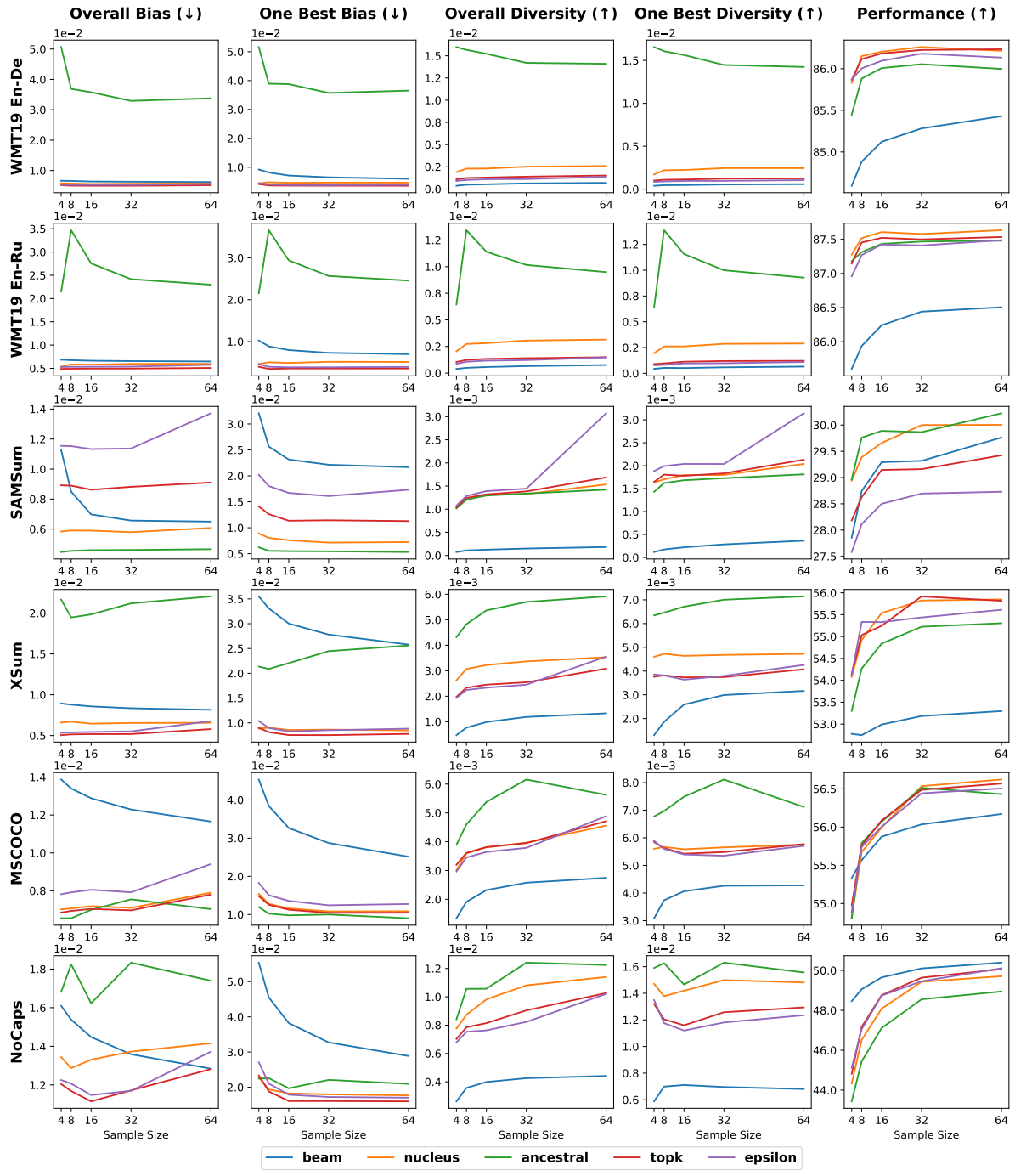

Figure 12: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with hypotheses generated by epsilon sampling. The notations are the same as Figure 2.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	84.5	84.7	84.8	85.0	85.1	**85.8**	**86.1**	86.3	**86.5**	86.5
	2	84.5	**84.8**	**85.0**	**85.1**	85.2	**85.8**	**86.1**	86.3	86.4	86.5
	4	84.6	**84.8**	**85.0**	**85.1**	85.2	**85.8**	**86.1**	**86.4**	**86.5**	**86.6**
	8	**84.7**	**84.8**	**85.0**	**85.1**	**85.3**	**85.8**	**86.1**	86.3	**86.5**	**86.6**
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	28.6	29.1	29.5	29.5	29.7	53.3	54.1	**55.0**	55.1	55.2
	2	**28.8**	**29.6**	**29.9**	**29.9**	30.1	53.2	54.2	**55.0**	**55.3**	**55.3**
	4	28.7	29.5	**29.9**	29.8	**30.2**	53.3	54.2	**55.0**	**55.3**	**55.3**
	8	28.7	29.5	29.8	**29.9**	30.1	**53.4**	**54.3**	**55.0**	55.2	**55.3**
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	54.8	55.8	56.1	56.4	56.6	43.2	45.3	46.9	48.2	49.0
	2	54.8	55.8	56.2	56.4	56.6	43.4	45.7	47.2	48.7	49.1
	4	54.8	**55.9**	**56.4**	56.5	**56.8**	43.8	45.8	**47.5**	48.8	**49.5**
	8	**54.9**	**55.9**	56.3	**56.6**	**56.8**	**43.9**	**45.9**	47.4	**49.0**	**49.5**

Table 6: Results of MAMBR with samples generated by ancestral sampling. The notations are the same as Table 1.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	85.2	85.4	85.6	85.6	85.6	**86.7**	87.0	**87.1**	87.0	87.1
	2	**85.3**	85.5	85.6	**85.7**	85.7	**86.7**	87.0	**87.1**	**87.1**	87.1
	4	**85.3**	85.5	85.6	**85.7**	85.7	**86.7**	87.0	**87.1**	87.0	87.1
	8	**85.3**	**85.6**	**85.7**	**85.7**	**85.8**	**86.7**	**87.1**	**87.1**	87.0	**87.2**
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	27.5	27.9	28.3	28.4	28.5	54.1	55.1	55.1	55.4	55.5
	2	**27.7**	28.1	28.5	**28.6**	**28.7**	54.1	**55.2**	55.1	55.4	55.4
	4	**27.7**	**28.2**	28.5	**28.6**	**28.6**	54.1	**55.2**	55.2	**55.5**	**55.6**
	8	27.6	**28.2**	**28.6**	**28.6**	**28.7**	**54.2**	**55.2**	**55.3**	55.4	**55.6**
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	55.0	55.8	56.0	56.4	56.6	45.0	46.9	48.6	49.5	49.8
	2	55.0	55.7	56.1	56.4	56.6	45.0	47.1	48.9	49.5	50.1
	4	55.0	**55.9**	**56.2**	**56.6**	56.7	**45.2**	**47.2**	48.9	**49.8**	50.2
	8	**55.1**	55.8	**56.2**	56.5	**56.8**	**45.2**	**47.2**	**49.0**	**49.8**	**50.3**

Table 7: Results of MAMBR with samples generated by epsilon sampling. The notations are the same as Table 1.

		WMT19 En-De					WMT19 En-Ru				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	85.1	85.3	85.4	85.4	85.3	**86.7**	**86.9**	**86.9**	**86.9**	86.9
	2	85.1	85.4	**85.5**	85.4	85.4	**86.7**	86.8	**86.9**	**86.9**	**87.0**
	4	**85.2**	85.4	**85.5**	85.4	85.4	**86.7**	86.8	**86.9**	**86.9**	86.9
	8	**85.2**	**85.5**	**85.5**	**85.5**	**85.5**	86.6	**86.9**	**86.9**	**86.9**	86.9
		SAMSum					XSum				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	27.6	28.7	**29.2**	29.3	**29.7**	**52.8**	52.8	**53.1**	**53.2**	53.3
	2	**27.8**	**28.9**	**29.2**	**29.4**	**29.7**	52.7	**52.9**	53.0	**53.2**	**53.4**
	4	**27.8**	**28.9**	**29.2**	**29.4**	**29.7**	**52.8**	52.8	53.0	**53.2**	53.3
	8	**27.8**	**28.9**	**29.2**	**29.4**	**29.7**	52.7	**52.9**	**53.1**	**53.2**	53.3
		MSCOCO					NoCaps				
Num. of Samples		4	8	16	32	64	4	8	16	32	64
Num. of Models	1	**55.3**	**55.5**	**55.9**	56.0	56.2	**48.5**	49.0	49.6	50.1	**50.5**
	2	**55.3**	**55.5**	55.8	**56.1**	56.2	**48.5**	49.0	**49.8**	50.2	**50.5**
	4	**55.3**	**55.5**	55.8	56.0	**56.3**	**48.5**	49.1	49.7	50.2	**50.5**
	8	**55.3**	**55.5**	55.8	**56.1**	56.2	**48.5**	**49.2**	49.7	**50.3**	**50.5**

Table 8: Results of MAMBR with samples generated by beam decoding. The notations are the same as Table 1.

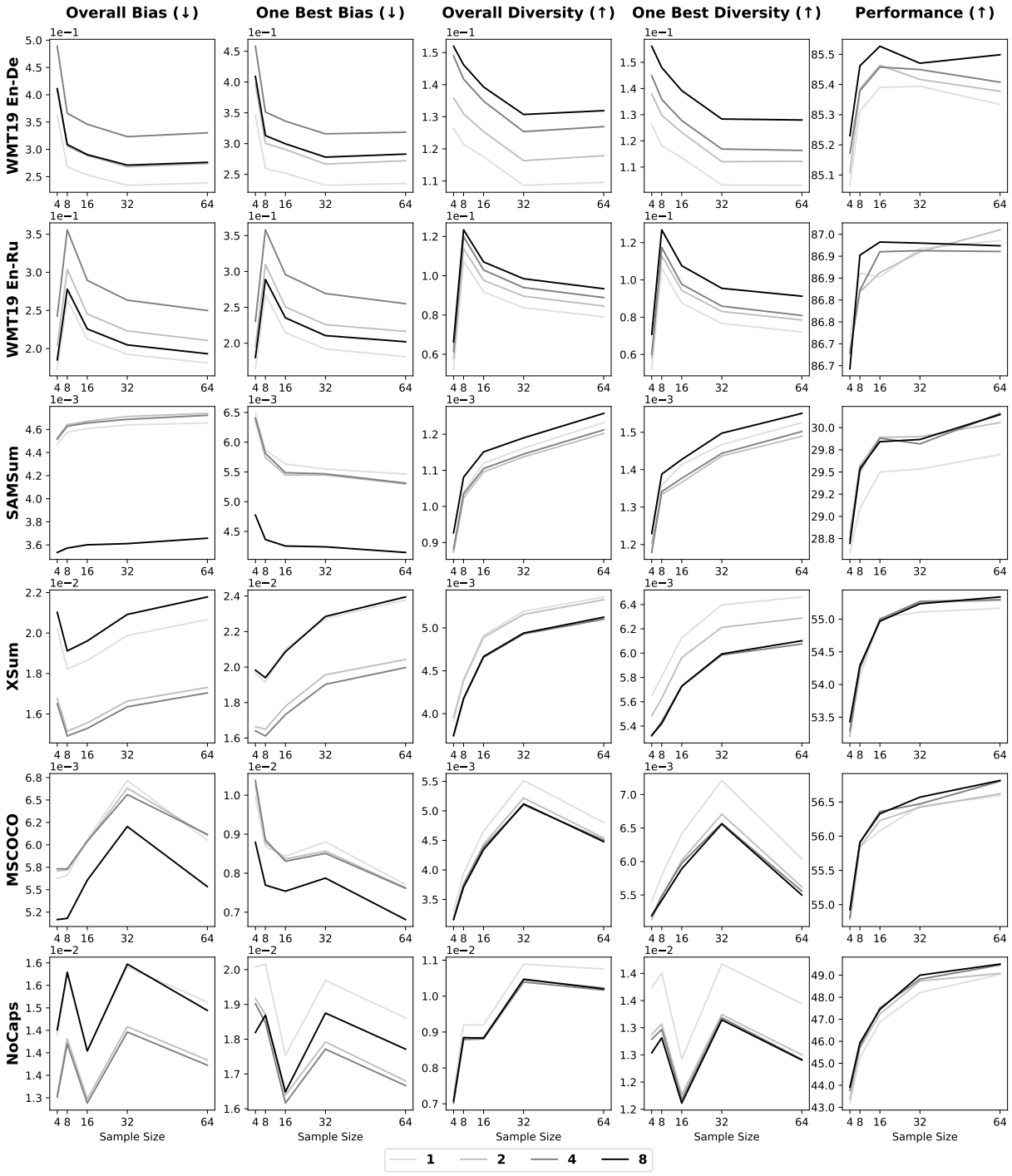

Figure 13: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with pseudo-references generated by ancestral sampling. The notations are the same as Figure 5.

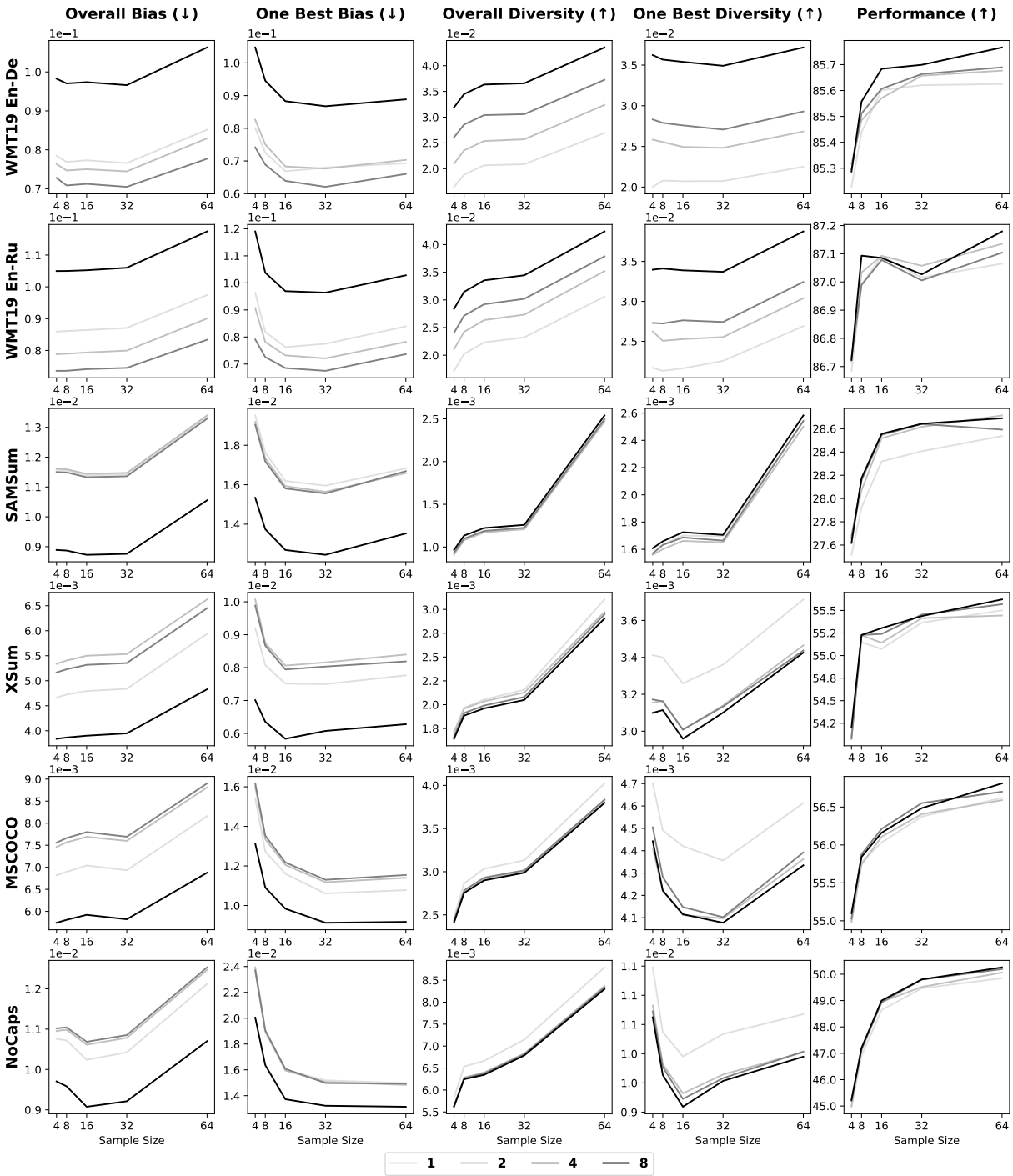

Figure 14: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with pseudo-references generated by epsilon sampling. The notations are the same as Figure 5.

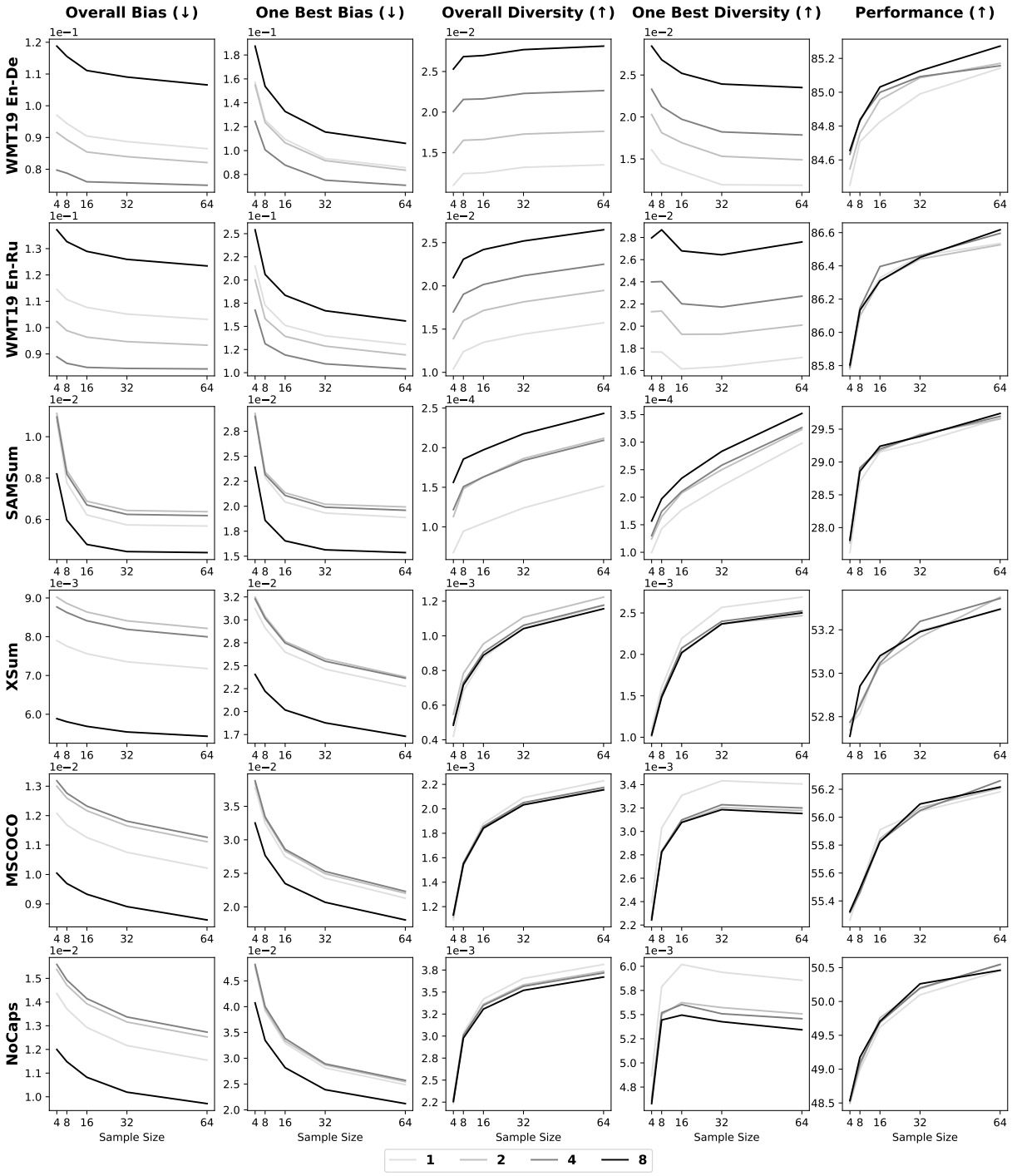

Figure 15: The relationship between bias, diversity, and performance on the first 1000 lines of each dataset in MBR decoding with pseudo-references generated by beam decoding. The notations are the same as Figure 5.