On Temperature-Constrained Non-Deterministic Machine Translation: Potential and Evaluation

Anonymous ACL submission

Abstract

In recent years, the non-deterministic properties of language models have garnered considerable attention and have shown a significant influence on real-world applications. However, such properties remain under-explored in machine translation (MT), a complex, non-deterministic NLP task. In this study, we systematically evaluate modern MT systems and identify temperature-constrained **N**on-**D**eterministic **MT** (**ND-MT**) as a distinct phenomenon. Additionally, we demonstrate that ND-MT exhibits significant potential in addressing the multi-modality issue that has long challenged MT research and provides higher-quality candidates than **D**eterministic MT (D-MT) under temperature constraints. However, ND-MT introduces new challenges in evaluating system performance. Specifically, the evaluation framework designed for D-MT fails to yield consistent evaluation results when applied to ND-MT. We further investigate this emerging challenge by evaluating five state-of-the-art ND-MT systems across three open datasets using both lexical-based and semantic-based metrics at varying sampling sizes. The results reveal a Buckets effect across these systems: the lowest-quality candidate generated by ND-MT consistently determines the overall system ranking across different sampling sizes for all reasonable metrics. Furthermore, we propose the ExpectoSample strategy to automatically assess the reliability of evaluation metrics for selecting robust ND-MT.

1 Introduction

The revolutionary development of large language models and their emergent capabilities (Wei et al., 2022) has demonstrated significant influence across various fields, including complex downstream NLP tasks (Wang et al., 2019; Hendrycks et al., 2021; Li et al., 2024), science (D'Souza et al., 2025), and mathematical reasoning (Ahn et al., 2024). In recent years, researchers have increasingly recognized the non-deterministic properties (Atil et al., 2025; Song et al., 2025) of LLMs and revealed their potential in enabling chat-box applications (DeepSeek-AI, 2025; OpenAI et al., 2024; Yang et al., 2025). Recent studies have also conducted fine-grained exploration and analysis of this property, primarily focusing on deterministic tasks (Song et al., 2025; Kuhn et al., 2023) such as question answering. However, the impact of such properties on machine translation—a complex, non-deterministic NLP task—remains under-explored. In this paper, we examine modern **N**on-**D**eterministic **Machine Translation** (**ND-MT**) systems to investigate the potential and challenges of non-determinism in this context.

We first address one of the most prominent challenges in MT: multi-modality (Papineni et al., 2002; Bao et al., 2023), which refers to the phenomenon where a single source sentence can have multiple candidates. This challenge becomes particularly problematic in automatic evaluation due to the scarcity of comprehensive reference sets (Papineni et al., 2002; Popović, 2015; Rei et al., 2022b) in most cases, as well as the need to evaluate one candidate from D-MT. Previous research has employed human assessment (Kocmi et al., 2024, 2023, 2025) to mitigate this issue; however, this approach faces the challenge of endless assessment requirements due to domain shifts in source texts (Kocmi et al., 2025). We reformulate this challenge as a dual requirement for MT: the candidates for a source sentence should demonstrate lexical diversity (Ploeger et al., 2024) while maintaining semantic equivalence (Kuhn et al., 2023) with the original source sentence. Notably, candidates generated from ND-MT may potentially satisfy both principles, as observed through direct examination of the generated outputs. In this work, we systematically investigate the potential of ND-MT in addressing multi-modality, particularly its ability to provide both lexical diversity and semantic equivalence across

22 modern MT systems in six language directions under the same temperature setting (0.5). Additionally, we design a reference-free lexical metric, the Group Lexical Variance Score (GLVS), to address the scarcity of references. We employ both lexical-based and semantic-based metrics to measure the effect of non-determinism on lexical diversity and semantic equivalence, respectively. The results demonstrate significant lexical variance with nearly identical semantic meanings compared to D-MT systems using the same underlying models across all ND-MT systems. Furthermore, we investigate the impact of temperature, a crucial parameter in ND-MT, on system performance. The results indicate that all temperature settings can generate candidates with lexical diversity, while only low temperatures preserve semantic equivalence; we therefore characterize modern MT systems as temperature-constrained ND-MT systems.

However, ND-MT presents challenges to the current evaluation scheme (Kocmi et al., 2024, 2023, 2025) (automatic evaluation followed by human assessment) due to the large number of generated candidates that satisfy both lexical diversity and semantic equivalence criteria. To address these emerging challenges in ND-MT, we first apply an intuitive approach: utilizing the ranking of the corresponding D-MT version. The results reveal inconsistent relationships across five group-based measurements: *min*,*max*, *mean*, *random*, and *std (standard deviation)*, demonstrating the unreliability of the current D-MT evaluation scheme when applied to ND-MT. Furthermore, we examine ranking consistency across these five measurements with varying sampling sizes ($\{10, 20, 50\}$) on five state-of-the-art ND-MT systems at a fixed temperature (0.5). The results uncover a strong Buckets effect, where the lowest-quality candidate for each source consistently determines the ranking across different sample sizes. For practical application, we propose the *ExpectoSample* strategy, which considers the average performance of candidate groups to identify reliable metrics and select robust ND-MT systems.

Our contributions are threefold: (1) We demonstrate that ND-MT systems address the multi-modality challenge through lexical diversity while maintaining semantic equivalence under temperature constraints. (2) We uncover the Buckets effect in ND-MT evaluation, where the lowest-quality candidate determines system ranking, and propose the ExpectoSample strategy to identify reliable metrics for robust system selection. (3) We systematically investigate 22 ND-MT systems across six language directions with 11,947 source cases, and release all code, data, and evaluation results to support future research.

2 Related Works

2.1 Modern MT Systems

Modern machine translation follows the sequence-to-sequence paradigm (Sutskever et al., 2014) with the Transformer (Vaswani et al., 2017) as the backbone and is divided into two main types: encoder-decoder models pre-trained on multilingual text then fine-tuned on bilingual text, and decoder-only architectures pre-trained on multilingual text without specific fine-tuning requirements. From an inference perspective, encoder-decoder models (Liu et al., 2020; Team et al., 2022) require explicit language signals as input during both training and inference, while decoder-only models (Touvron et al., 2023; Grattafiori et al., 2024; Qwen et al., 2025; Yang et al., 2025; DeepSeek-AI, 2025) leverage the inherent multilingual semantic alignment of LLMs and activate MT capabilities through various prompts. Different LLM-based MT approaches exhibit distinct characteristics: pre-training-only MT systems typically use few-shot methods (Brown et al., 2020; Vilar et al., 2023) (commonly five-shot) but inevitably introduce repetition and language mismatch issues (Wang et al., 2024); instruction-tuned MT systems use direct MT prompts but sometimes produce noise without strict constraints (Touvron et al., 2023; Grattafiori et al., 2024) (e.g., Chinese translations including Pinyin in Llama series models); RL-based reasoning MT systems use direct MT prompts and can provide detailed translation steps but require substantial computational resources for both post-editing and inference (DeepSeek-AI, 2025; Yang et al., 2025). Generally, modern MT systems use a generate-once approach (Kocmi et al., 2025) to produce deterministic results, while their potential to generate multiple candidate translations through non-deterministic sampling remains underexplored.

2.2 Non-determinism of LLMs

Previously, substantial effort focused on deterministic tasks such as sentiment classification (Zhang et al., 2024) and parsing (Ginn and Palmer, 2025), with most attention directed toward extracting deterministic capabilities from LLMs. In recent

years, the non-deterministic properties of LLMs have emerged and been leveraged to satisfy customized user requirements (Tseng et al., 2024). Some models now implement non-determinism as a default property (DeepSeek-AI, 2025; Yang et al., 2025), enabling LLMs to provide various reasonable outputs under the same prompt to increase user satisfaction. Previous studies have found that this property can benefit certain deterministic NLP tasks (Song et al., 2025), such as question answering, by generating semantically equivalent responses (Kuhn et al., 2023). However, systematic research on complex non-deterministic tasks such as MT remains limited. In this work, we systematically investigate the effects of non-determinism in LLM-based MT systems across various architectures, revealing both the potential and challenges introduced by this property.

2.3 Automatic Evaluation on MT

Automatic evaluation methods play a key role in evaluating MT systems by avoiding the substantial costs of human assessment. In this work, we investigate the potential of ND-MT to provide lexical diversity and semantic equivalence. To achieve this goal, we categorize current metrics into two main categories: lexical-based methods and semantic-based methods, to measure the capabilities of ND-MT. For lexical-based methods, BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004), which focus on lexical overlap. ChrF++ (Popović, 2015) focuses on character overlap and TER (Snover et al., 2006) focuses on error edit distance. Specifically, these methods rely on references, suffer from the multi-modality issue, and fail without references. For semantic-based methods, BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020) utilize the token information to model the semantic score. COMET20-DA (Rei et al., 2020) and COMET22-KIWI (Rei et al., 2022a) include a training stage to learn the semantic equivalence between source and candidates. XCOMET (Guerreiro et al., 2024) further evaluates on the error spans. Other methods measure semantic alignment through semantic similarity between the source and candidates in a unified semantic embedding space. including SentTrans (Reimers and Gurevych, 2019) with direct LMs, LASER (Heffernan et al., 2022), and XNLI (Conneau et al., 2020) using bilingual pairs. In this work, we test the reliability of these metrics on evaluating ND-MT systems.

3 Modern MT Systems Are Temperature-Constraint ND-MT

In this section, we systematically investigate the non-deterministic properties of modern MT systems. We begin with experimental preparation by selecting state-of-the-art modern MT systems across encoder-decoder and decoder-only architectures with varying model sizes. We then generate the candidates and evaluate them based on group measurements. The results demonstrate how ND-MT addresses the multi-modality challenge, as evidenced by our findings. Finally, we examine the role of the temperature parameter in ND-MT, reveal its influence on translation quality, and characterize modern MT systems as temperature-constrained ND-MT.

3.1 Experimental Preparation

3.1.1 ND-MT Systems

We explore the mainstream autoregressive generation method for MT across two highly successful architectures: encoder-decoder and decoder-only. We further categorize LLM-based MT (decoder-only) into three types based on training approaches: pre-trained models using multilingual texts, instruction-tuned models with post-training alignment, and reasoning models that generate thinking steps through reinforcement learning. Notably, model series are differentiated into distinct MT systems based on their training type. For encoder-decoder architectures, we select mBART (Liu et al., 2020) trained on 50 multilingual texts (0.68B parameters) and NLLB-200 (Team et al., 2022) with three model scales (0.6B, 3.3B, and 54.6B parameters). For LLM-based MT, we include the Llama-2 series (Touvron et al., 2023), Llama-3 series (Grattafiori et al., 2024), Qwen-2.5 series (Qwen et al., 2025), Qwen-3 (Yang et al., 2025) series, and DeepSeek series (DeepSeek-AI, 2025), examining both small-scale (7-8B parameters) and large-scale (70-72B and 671B parameters) variants across pre-trained, instruction-tuned, and reasoning types when available. We provide comprehensive information about these models in Appendix A.

3.1.2 Datasets

3.2 dataset statistics

In this work, we adopt sentence-level MT as our starting point and leverage existing, well-established open-source datasets to study both ND-MT and their corresponding D-MT. Specifically,

Table 1: Dataset Statistics Information

Source	Language Pair	Size
WMT23	En→Zh	2,074
WMT23	Zh→En	1,976
WMT23	En→De	557
WMT23	De→En	549
WMT23	En→Ru	2,074
WMT23	Ru→En	1,723
WMT24	En→Zh	998
WMT24	En→De	998
WMT24	En→Ru	998

we use the latest WMT data from 2023–2024 [1] across six translation directions (ZHEN, ENDE, ENRU), covering three language pairs: ⟨English, Chinese⟩, ⟨English, German⟩, and ⟨English, Russian⟩. We identify ⟨English, Chinese⟩ translation as particularly valuable for investigation due to substantial differences in language families and structural characteristics, making it our primary experimental setting to explore the potential of ND-MT. We also evaluate ⟨English, German⟩ and ⟨English, Russian⟩ to demonstrate the generalizability of ND-MT across diverse language pairs. We present detailed statistics in Table 1.

3.2.1 Evaluation Methods

Lexical-based Methods We include BLEU (Papineni et al., 2002), an n-gram-based metric evaluating lexical overlap; ChrF++ (Popović, 2015), an n-gram-based metric capturing both lexical and character-level information; METEOR (Banerjee and Lavie, 2005), a token-level alignment metric; ROUGE(-1, -2, -L)(Lin, 2004), a recall-oriented n-gram overlap metric; and TER (Snover et al., 2006), a token-level edit distance metric.

Semantic-based Methods We include COMET22KIWI(Rei et al., 2022a) and COMET20DA (Rei et al., 2020) to measure with the neural network; LASER (Heffernan et al., 2022), LaBSE (Heffernan et al., 2022), SentTrans (Reimers and Gurevych, 2019), and XNLI (Conneau et al., 2020) to test the semantic equivalence on a unified semantic space; BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2019) to measure the semantic equivalence with token information.

Group Lexical Variance Score (GLVS) To address reference scarcity and enhance sensitivity to variance among group candidates, we propose the Group Lexical Variance Score (GLVS), which evaluates a group of candidates from an ND-MT system for a given source. The algorithm is straightforward and consists of three steps:

Step 1: Tokenize each candidate c_i into words $\mathcal{W}_i = w_1, w_2, \ldots, w_l$

Step 2: Construct a frequency vocabulary $\mathcal{V}_t$ from the combined word sets

Step 3: Compute the GLVS for c_i:

$$V(c_i) = \sum_{w \in \mathcal{W}i^U} f\mathcal{V}_t(w), \tag{2}$$

where $\mathcal{W}i^U$ represents the unique word set in c_i, and $f\mathcal{V}_t(w)$ denotes the frequency of word w.

3.2.2 Experimental Set-up

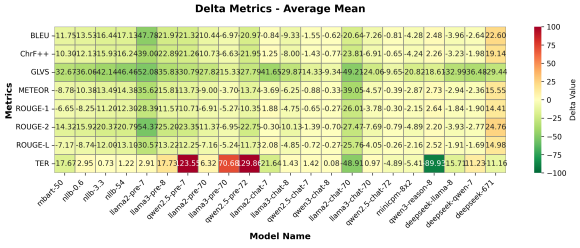
Decoding Strategy For decoding strategies, we employ greedy decoding as the deterministic baseline and sampling-based decoding for the non-deterministic setting with adjustable temperature, generating K candidates for each source. We use an initial setting of temperature 0.5 and sampling size 10 to investigate the potential of ND-MT, following established practices from prior semantic equivalence research.

Group-based Measurements For each evaluation metric, we design group-based measurements—*min*, *max*, *mean*, *random*, and *std* (standard deviation), to capture different aspects of ND-MT performance: lower bound, upper bound, average performance, single-response simulation (representing real-world usage), and performance variability, respectively, at the group level for generated candidates of each source. We then aggregate results across the entire dataset to obtain the average values, yielding overall ND-MT system performance metrics.

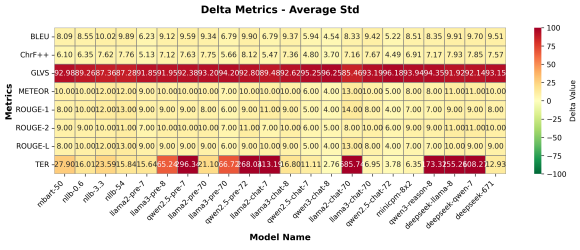
3.3 The Potential of ND-MT to Solve Multi-Modality

To explore the potential of ND-MT in addressing the multi-modality challenge across two dimensions: lexical variance and semantic equivalence. We conduct experiments on 22 ND-MT systems for the ⟨English, Chinese⟩ pair, with a temperature of 0.5 and a sampling size of 10 (Kuhn et al., 2023), without additional non-deterministic settings. We also include the corresponding D-MT systems as baselines to enable direct comparison, where each system generates only one candidate. Notably,

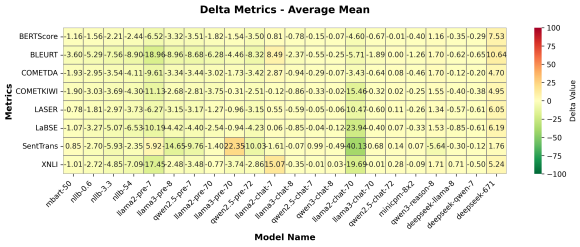
[1] https://github.com/wmt-conference/wmtX-news-systems, x = {23, 24}



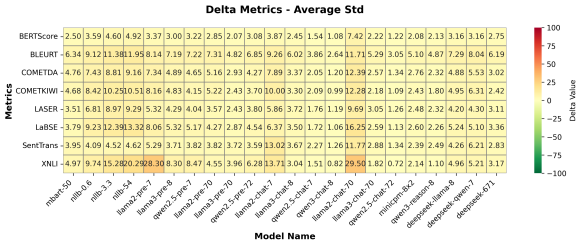
(a) Lexical Metrics–Delta Mean Values



(b) Lexical Metrics–Delta Std Values



(c) Semantic Metrics–Delta Mean Values



(d) Semantic Metrics–Delta Std Values

Figure 1: Delta Mean and Std (Standard deviation) values on WMT23 EN-ZH measured by lexical and semantic metrics under temperature of 0.5 with 10 candidates, respectively. The delta value is computed with deterministic results on the same dataset under greedy decoding.

we access DeepSeek-671B through an API that does not allow modification of the default sampling method; consequently, DeepSeek-671B produces non-deterministic outputs for both D-MT and ND-MT configurations, differing only in sampling size. This scenario provides valuable insights into evaluating emerging closed-source ND-MT systems. So temporarily keep the results of DeekSeek-671 while not analyzing it. We compute the *delta* results based on the D-MT for easy understanding. The results are shown in Figure 1.

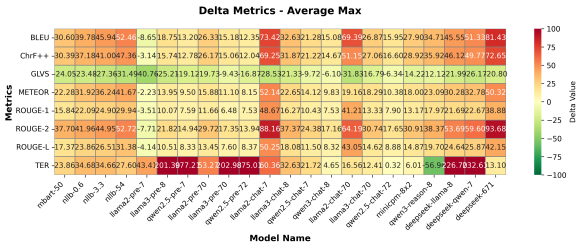
ND-MT can provide lexical diversity Reference-based metrics (Figure 1a) reveal modest performance variations between D-MT and ND-MT systems across most lexical metrics, with differences typically within 10 percent, except for TER values. These results demonstrate the comparable quality between D-MT and ND-MT candidates when assessed by lexical metrics. For TER, which measures edit distance (where higher values indicate greater lexical differences), we observe distinct patterns across ND-MT systems. All pre-trained LLM-based ND-MT systems exhibit positive delta values, reflecting increased lexical variation compared to their D-MT counterparts. In contrast, other systems show minimal differences, indicating closer lexical similarity to D-MT. For the reference-free metric GLVS, the values reflect the lexical diversity of MT systems, where lower scores indicate higher diversity; deterministic MT systems consistently score 100. The substantial GLVS values demonstrate that non-deterministic MT systems generate diverse lexical representations while maintaining quality, as evidenced by the modest gaps in other lexical metrics (except for TER). This conclusion is further supported by the large lexical standard deviation values of GLVS shown in Figure 1b. A significant advantage of GLVS is its applicability to common scenarios, eliminating the need for reference translations. For a better understanding, we provide the baseline results in Appendix B.

ND-MT can keep the semantic equivalence We observe that the performance gaps (Figure 1c) for semantic-based metrics are substantially smaller than those for lexical-based metrics (Figure 1a), with differences below 10 percentage points for most MT systems, except for specific cases like llama2-pre-7 and llama2-chat-70. The average standard deviation values (Figure 1d) remain below 10 percentage points across all metrics, demonstrating strong semantic equivalence under non-deterministic settings. For a better understanding, we provide the baseline results in Appendix B.

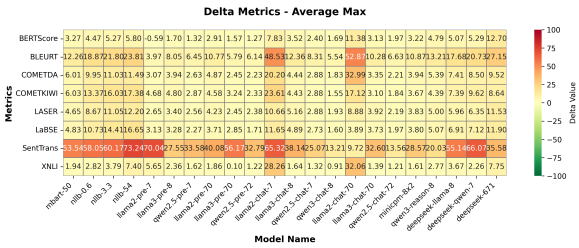
ND-MT has the potential to provide better candidates than D-MT We further explore the potential of non-deterministic MT systems in providing high-quality candidates. We select the best candidate according to each metric from the candidate group for each source, then compute the average maximum values across the dataset. Note that in real-world scenarios, references are unavail-

able; therefore, this analysis simulates the ideal performance potential of non-deterministic MT systems rather than actually selecting the "best" candidate based on specific metrics. Figure 2 demonstrates overall improved performance for non-deterministic MT systems, revealing their substantial potential for generating higher-quality candidates. This finding validates prior work on data augmentation and candidate selection using non-deterministic MT systems.

Generality of ND-MT potential in addressing multi-modality Finally, we evaluate the generality of ND-MT potential across different language pairs. We test ⟨German, English⟩ and ⟨Russian, English⟩ in both directions with five state-of-the-art LLM-based MT models (Touvron et al., 2023; Qwen et al., 2025; Yang et al., 2025). The results in Figure 3 exhibit similar trends to those observed in Figure 1, leading us to conclude that modern ND-MT systems demonstrate significant potential for generating diverse candidates under semantic equivalence, effectively addressing multi-modality limitations. Our experimental evidence indicates that modern MT systems learn translation through semantic equivalence and lexical diversity, positioning them as viable alternatives to D-MT systems. Future research can unlock the full potential of ND-MT systems in generating higher-quality translation candidates.

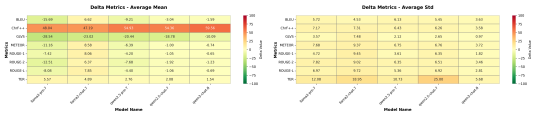


(a) Lexical Metrics–Delta Max Values

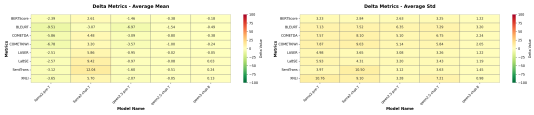


(b) Semantic Metrics–Delta Max Values

Figure 2: Delta max values on WMT23 EN-ZH measured by lexical and semantic metrics,respectively. The delta value is computed with deterministic results on the same dataset under greedy decoding.



(a) Lexical Metrics–Delta Mean Values on EN-DE (b) Lexical Metrics–Delta Mean Values on EN-RU



(c) Semantic Metrics–Delta Mean Values on EN-DE (d) Semantic Metrics–Delta Mean Values on EN-RU

Figure 3: Delta Mean values on WMT23 EN-DE and WMT23 EN-RU measured by lexical and semantic metrics, respectively.

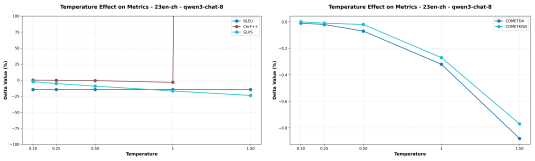
3.4 Temperature Constraints on ND-MT Potential

While we have demonstrated the potential of ND-MT in addressing multimodality challenges, the quality of generated candidates depends critically on the temperature parameter. We further investigate the effect of temperature on the performance of ND-MT. Unlike previous fine-grained studies aimed at identifying optimal parameters for generating the best single candidate, we examine how temperature influences the overall potential of ND-MT. We conduct experiments on WMT23 EN-ZH using five models (Touvron et al., 2023; Qwen et al., 2025; Yang et al., 2025), with qwen3-chat-8 serving as a representative example, as all models exhibit similar trends. We list all the results and detailed analysis for metrics and models in Appendix C

We evaluate both lexical diversity and semantic equivalence using the same metrics from Section 3.3. For lexical analysis, we select GLVS as a reference-free metric, and BLEU and ChrF++ as reference-based metrics that measure effects at the lexical and character levels, respectively. For semantic analysis, we choose COMET20DA and COMET22KIWI as reference-based and reference-free metrics, respectively. Figure 4 presents the results. GLVS shows a decreasing trend as temperature increases, indicating that lexical diversity grows with temperature, which aligns with the general purpose of raising temperature: making a broader range of lexical items more probable. Notably, ChrF++ exceeds 100 at higher temperatures, indicating its unreliability for evaluation on ND-MT. The semantic metrics exhibit a monotonic decreasing trend, indicating that **as temperature**

increases, ND-MT maintains lexical diversity while sacrificing semantic equivalence. In practical applications, the acceptable degree of semantic degradation depends on the specific use case and the baseline semantic quality. Our observations align with previous findings that non-deterministic systems show weaker performance than deterministic systems on certain downstream tasks (Song et al., 2025).

In summary, to harness the potential of ND-MT, temperature values must be carefully calibrated to maintain both lexical diversity and semantic equivalence when addressing multi-modality challenges. Additionally, the effects of specific temperature settings should be evaluated in advance to align with application requirements. Our experimental evidence reveals that semantic equivalence decreases while lexical diversity increases with rising temperature, providing valuable guidance for determining optimal temperature configurations in future ND-MT research and applications.



(a) Temperature Effect on Lexical Metrics (b) Temperature Effect on Semantic Metrics

Figure 4: The temperature effect from qwen3-chat-8 model on WMT23 EN-ZH dataset on GLVS, BLEU (Papineni et al., 2002), ChrF++ (Papineni et al., 2002) of lexical metrics and COMETDA (Rei et al., 2020), COMETKIWI (Rei et al., 2022a) of semantic metrics.

4 The Under-Explored Space of ND-MT on the Evaluation Scheme

4.1 Challenges of the Current D-MT evaluation Scheme on ND-MT

In Section 3.4, we demonstrate the potential of ND-MT to address multi-modality challenges by providing lexically diverse candidates while maintaining semantic equivalence within the candidate set. This raises an important question: how should we evaluate current and future ND-MT systems? The prevailing generate-once evaluation paradigm relies on established metrics that have been validated through human assessment. However, this paradigm is primarily suited for D-MT for two key reasons: 1) The multi-modality challenge represents a fundamental limitation that affects both the design and measurement capabilities of existing metrics. For instance, lexical-based metrics such as BLEU and ChrF++ allow multiple references during evaluation, yet this assumes the availability of such references, which is often impractical to obtain. Conversely, semantic-based metrics leverage large-scale supervised training to mitigate multi-modality issues; however, their effectiveness remains constrained by the scale of the training data and computational resources. Nevertheless, larger models demonstrate improved evaluation performance. 2) The non-deterministic nature of ND-MT, which generates numerous candidates, renders traditional human evaluation impractical, particularly given that ND-MT performance is temperature-dependent.

In this section, we investigate the under-explored domain of ND-MT evaluation frameworks. First, we examine an intuitive approach that directly applies evaluation rankings from D-MT, revealing significant inconsistencies. Second, we evaluate current metrics using group-based measurements and identify the bucket effect in ND-MT that influences ranking determination. Finally, we propose the *ExpectoSample* strategy to identify reliable metrics for selecting robust ND-MT systems across varying sampling sizes.

4.2 The Inconsistent Evaluation Results between ND-MT and D-MT

One intuitive approach is to directly apply the ranking from deterministic MT systems to their non-deterministic counterparts. We evaluate this approach by computing Spearman's ρ and Kendall's τ across five aggregation methods: *min*, *max*, *mean*, *random*, and *std*. Specifically, we hypothesize that higher-ranked MT systems possess stronger capabilities for generating high-quality candidates; consequently, we expect *std* to exhibit high negative correlation (i.e., higher-ranked MT systems should produce lower *std* values).

The results in Tables 2 and 7 present correlations for lexical-based and semantic-based metrics, respectively. While most metrics demonstrate moderate to strong correlations exceeding 0.5 for both Spearman's ρ and Kendall's τ (with TER being a notable exception), the observed gaps suggest that D-MT evaluation rankings provide limited reliability when applied to ND-MT systems. Furthermore, the weak correlations for *std* suggest that assessing the robustness of ND-MT systems requires evaluation frameworks that extend beyond traditional

Table 2: Correlation Results of Lexicon-based Metrics on WMT23 EN-ZH for 22 ND-MT Systems.

Strategy	BLEU	METEOR	ROUGE	TER	chrF++
Kendall's τ / p-value					
Min	.69/.00	.68/.00	.70/.00	.19/.22	.69/.00
Max	.69/.00	.70/.00	.71/.00	.27/.08	.70/.00
Mean	.67/.00	.68/.00	.69/.00	.32/.04	.72/.00
Random	.69/.00	.69/.00	.68/.00	.18/.26	.71/.00
Std	-.09/.57	-.47/.00	-.56/.00	.30/.05	-.02/.91
Spearman's ρ / p-value					
Min	.87/.00	.87/.00	.87/.00	.28/.21	.87/.00
Max	.86/.00	.87/.00	.88/.00	.34/.12	.86/.00
Mean	.86/.00	.87/.00	.87/.00	.33/.13	.88/.00
Random	.87/.00	.87/.00	.86/.00	.20/.37	.88/.00
Std	-.13/.57	-.60/.00	-.70/.00	.35/.11	.00/.99

deterministic approaches.

4.3 Buckets Effect of ND-MT

Table 3: Correlation Analysis of MT Evaluation Metrics Across Sampling Sizes with Lexical Metrics on WMT23 EN-ZH with Five SOTA ND-MT Systems

Size	strategy	BLEU		GLVS		ChrF++	
		ρ	τ	ρ	τ	ρ	τ
20	Max	.70	.60	1.0	1.0	.90	.80
	Mean	.90	.80	.90	.80	1.0	1.0
	Min	1.0	1.0	1.0	1.0	1.0	1.0
	Rand.	.90	.80	.90	.80	1.0	1.0
	Std	.70	.60	.90	.80	1.0	1.0
50	Max	.70	.60	.90	.80	.90	.80
	Mean	.90	.80	.90	.80	1.0	1.0
	Min	1.0	1.0	1.0	1.0	1.0	1.0
	Rand.	.90	.80	.90	.80	1.0	1.0
	Std	.70	.60	1.0	1.0	.90	.80

ρ = Spearman's correlation; τ = Kendall's tau. All correlations significant at $p < 0.10$.

To further investigate reliable evaluation frameworks, we conduct experiments across different sampling sizes $(10, 20, 50)$ while maintaining constant temperature values for five state-of-the-art ND-MT models. For evaluation metrics, we employ BLEU, ChrF++, and GLVS as lexical-based metrics, and COMET20DA and COMET22KIWI as semantic metrics. The results are presented in Tables 3 and **??**. A key observation is the bucket effect: the minimum-score aggregation method for ND-MT systems provides stable ranking evaluations across all sampling sizes and metrics. Encouragingly, our findings demonstrate that controlled sampling sizes—rather than arbitrarily large samples—can yield reliable evaluations with existing metrics. However, metric selection requires careful consideration, as evidenced by the exceptional behavior of TER discussed in Appendix D.

4.4 ExpectoSample: Selecting Reliable Metrics and Robust ND-MT Systems

To address the challenge of identifying reliable evaluation metrics and robust ND-MT systems, we propose the *ExpectoSample* strategy and use the mean value due to the real-usage consideration. This approach examines ranking correlations across sampling sizes $10, 20, 50$ based on the principle that reliable metrics should produce consistent system rankings regardless of sampling size, while robust ND-MT systems should maintain stable performance characteristics across different sample counts.

Our analysis reveals that metrics that maintain the same ranking across all sample size pairs can be considered reliable for ND-MT evaluation, while systems that produce consistent relative rankings under these metrics can be identified as robust ND-MT systems. We identify that ChrF++, COMET20DA, and COMET22KIWI are reliable metrics from Tables 3 and **??**. Future work can utilize this strategy to filter the reliable metrics for further robust ND-MT selection.

5 Conclusion

In this work, we systematically investigate ND-MT systems, revealing their significant potential in addressing the long-standing multi-modality challenge in MT. Through comprehensive experiments across 22 systems in six language directions, we demonstrate that ND-MT can generate lexically diverse candidates while maintaining semantic equivalence. However, we find that this potential is temperature-constrained: only low temperature settings preserve both lexical diversity and semantic equivalence. Our investigation also reveals critical challenges in evaluating ND-MT systems. We demonstrate that traditional D-MT evaluation schemes yield inconsistent rankings when applied to ND-MT and reveal the bucket effect, where minimum-score candidates consistently influence system rankings across varying sampling sizes. Finally, we propose the ExpectoSample strategy for identifying reliable metrics and robust ND-MT systems for real-world ND-MT usage.

6 Limitations

While our work provides a systematic investigation into ND-MT, several limitations warrant acknowledgment. First, our experiments focus primarily on SOTA modern MT systems mainly on open-

sourced models, and our findings may not generalize to other types of MT systems like closed-source MT systems. Second, our temperature analysis is constrained to a specific range of values, and the optimal temperature settings may vary across different model families, language pairs, or domain-specific applications. Third, our evaluation framework relies on existing automatic metrics (both lexical and semantic), which themselves have known limitations in capturing nuanced aspects of translation quality, such as cultural appropriateness, style consistency, and accuracy in domain-specific terminology.

Additionally, while we propose the Expecto-Sample strategy for identifying reliable metrics and robust systems, our experiments are limited to sampling sizes of $10, 20, 50$. Larger sampling sizes or different sampling strategies might reveal additional patterns or insights. Furthermore, our analysis of the bucket effect and ranking consistency does not include human evaluation due to the impracticality of assessing numerous candidates across multiple systems and sampling sizes. Human judgment would provide valuable validation of our automatic evaluation findings, particularly regarding whether the lexical diversity we observe translates to genuinely useful translation alternatives for end users. Finally, our investigation covers six language directions, which, while diverse, represent only a fraction of the world's languages, and our findings may not fully capture the challenges specific to low-resource languages or linguistically distant language pairs.

7 Ethical Statement

Our research on non-deterministic machine translation raises several ethical considerations that warrant careful attention. First, the non-deterministic nature of ND-MT systems, which generate multiple diverse candidates for a single source sentence, introduces potential risks in high-stakes applications such as legal document translation, medical information dissemination, or official communications. While lexical diversity can be beneficial in creative or informal contexts, deploying ND-MT systems without appropriate safeguards in critical domains could lead to inconsistent or ambiguous translations that may have serious consequences. Additionally, we use open-source LLMs that may inadvertently generate outputs containing personal information from their training data. We emphasize that practitioners must carefully assess the suitability of ND-MT for their specific use cases and implement appropriate quality control mechanisms.

Second, the temperature-constrained nature of ND-MT systems presents transparency challenges. Users of MT systems may not be aware that different temperature settings can significantly affect translation quality and semantic equivalence. This lack of transparency could undermine user trust, particularly when systems produce semantically divergent outputs at higher temperatures. Developers deploying ND-MT systems have a responsibility to clearly communicate these limitations to end users and provide appropriate controls or defaults that prioritize semantic accuracy. Additionally, the evaluation challenges we identify—particularly the unreliability of traditional D-MT evaluation schemes for ND-MT—highlight the need for careful system comparison and selection. Misleading performance claims based on inappropriate evaluation methods could harm users who rely on MT systems for important communications.

Finally, we acknowledge that our released code, data, and evaluation results could potentially be misused to develop MT systems without adequate quality assurance or to make unfounded claims about system capabilities. We encourage researchers and practitioners who utilize our resources to do so responsibly, with appropriate consideration for the limitations we have identified and the potential impacts on end users across diverse linguistic and cultural communities.

8 The Use of AI Assistant

The writing process of this paper incorporated stylistic and grammar suggestions from Claude Sonnet 4.5, supervised by the authors, without any content generation or fabrication.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024*, pages 225–237. Association for Computational Linguistics.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture,

Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" llm settings. *Preprint*, arXiv:2408.04667.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Guangsheng Bao, Zhiyang Teng, Hao Zhou, Jianhao Yan, and Yue Zhang. 2023. Non-autoregressive document-level machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14791–14803, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. 2025. YESciEval: Robust LLM-as-a-judge for scientific question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13749–13783, Vienna, Austria. Association for Computational Linguistics.

Michael Ginn and Alexis Palmer. 2025. LLM dependency parsing with in-context rules. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 186–196, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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.

Esther Ploeger, Huiyuan Lai, Rik Van Noord, and Antonio Toral. 2024. Towards tailored recovery of lexical diversity in literary machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 286–299, Sheffield, UK. European Association for Machine Translation (EAMT).

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, Albuquerque, New Mexico. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. 2024. Mitigating the language mismatch and repetition issues in LLM-based machine translation via model editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15681–15700, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

A Model Statistics

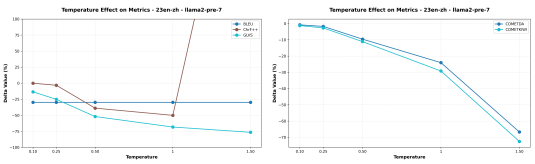
We show the detailed information on Table 4about the modern MT systems used in this paper.

B Evaluation Results on D-MT

We show the original evaluation results on D-MT systems for the <English, Chinese> pairs on WMT23 EN-ZH datasets in the lexical-based metrics Table 5 and semantic-based metrics Table 6. For other results, we will release them at a later time.

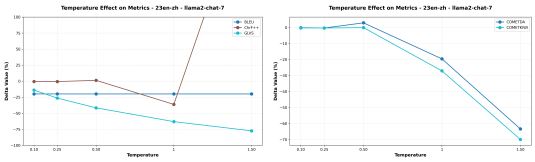
C Temperature Effect on ND-MT

We demonstrate the temperature effect on ND-MT across five state-of-the-art systems: llama2-pre-7 (Touvron et al., 2023), llama2-chat-7 (Touvron et al., 2023), qwen2.5-pre-7 (Qwen et al., 2025), qwen2.5-chat-7 (Qwen et al., 2025), and qwen3-chat-8 (Yang et al., 2025). Figures 5,6,7,8, and9 show a decreasing trend in semantic equivalence as temperature increases. However, for specific datasets, the optimal temperature is not always the greedy setting (e.g., llama2-chat-7 (Touvron et al., 2023)), indicating that ND-MT can exceed D-MT performance and that optimal temperature selection is non-trivial. Furthermore, we observe a general decrease in GLVS scores at higher temperatures, indicating a reduction in lexical diversity as models tend to generate natural language content rather than faithful translations. In contrast, BLEU (Papineni et al., 2002) scores remain nearly unchanged across temperatures, demonstrating BLEU's failure to detect variations in lexical diversity. These findings provide key references for future work on temperature selection and the evaluation of lexical diversity and semantic equivalence in ND-MT systems.



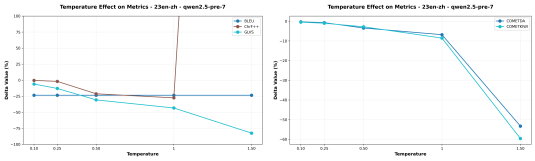
(a) Temperature Effect on Lexical Metrics (b) Temperature Effect on Semantic Metrics

Figure 5: The temperature effect from llama2-pre-7 model on WMT23 EN-ZH dataset on GLVS, BLEU (Papineni et al., 2002), ChrF++ (Papineni et al., 2002) of lexical metrics and COMETDA (Rei et al., 2020), COMETKIWI (Rei et al., 2022a) of semantic metrics.



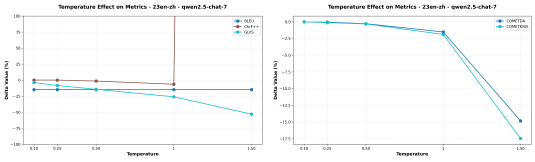
(a) Temperature Effect on Lexical Metrics (b) Temperature Effect on Semantic Metrics

Figure 6: The temperature effect from llama2-chat-7 model on WMT23 EN-ZH dataset on GLVS, BLEU (Papineni et al., 2002), ChrF++ (Papineni et al., 2002) of lexical metrics and COMETDA (Rei et al., 2020), COMETKIWI (Rei et al., 2022a) of semantic metrics.



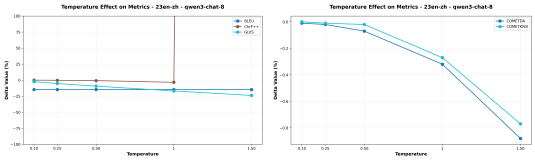
(a) Temperature Effect on Lexical Metrics (b) Temperature Effect on Semantic Metrics

Figure 7: The temperature effect from qwen2.5-pre-7 model on WMT23 EN-ZH dataset on GLVS, BLEU (Papineni et al., 2002), ChrF++ (Papineni et al., 2002) of lexical metrics and COMETDA (Rei et al., 2020), COMETKIWI (Rei et al., 2022a) of semantic metrics.



(a) Temperature Effect on Lexical Metrics (b) Temperature Effect on Semantic Metrics

Figure 8: The temperature effect from qwen2.5-chat-7 model on WMT23 EN-ZH dataset on GLVS, BLEU (Papineni et al., 2002), ChrF++ (Papineni et al., 2002) of lexical metrics and COMETDA (Rei et al., 2020), COMETKIWI (Rei et al., 2022a) of semantic metrics.



(a) Temperature Effect on Lexical Metrics (b) Temperature Effect on Semantic Metrics

Figure 9: The temperature effect from qwen3-chat-7 model on WMT23 EN-ZH dataset on GLVS, BLEU (Papineni et al., 2002), ChrF++ (Papineni et al., 2002) of lexical metrics and COMETDA (Rei et al., 2020), COMETKIWI (Rei et al., 2022a) of semantic metrics.

Table 4: Overview of Language Models Used in Experiments

Model Name	Architecture	Parameters (B)	Dense/MoE
Encoder-Decoder Models			
mBART	Encoder-Decoder NMT	0.68	Dense
NLLB-200	Encoder-Decoder NMT	0.6, 3.3, 54.5	Dense
Decoder-Only Models: Llama Family			
Llama 2	Decoder-Only Pre	7, 70	Dense
Llama 2	Decoder-Only Chat	7, 70	Dense
Llama 3	Decoder-Only Pre	7, 70	Dense
Llama 3	Decoder-Only Chat	7, 70	Dense
Decoder-Only Models: Qwen Family			
Qwen 2.5	Decoder-Only Pre	7, 72	Dense
Qwen 2.5	Decoder-Only Chat	7, 72	Dense
Qwen 3	Decoder-Only Chat	8	Dense
Qwen 3	Decoder-Only Reason	8	Dense
Decoder-Only Models: DeepSeek Family			
DeepSeek (Llama-based)	Decoder-Only Reason	8	Dense
DeepSeek (Qwen-based)	Decoder-Only Reason	7	Dense
DeepSeek-R1	Decoder-Only Reason	671	MoE
Other Decoder-Only Models			
MiniCPM	Decoder-Only Chat	16	MoE

NMT = Neural Machine Translation; Pre = Pre-trained; Chat = Instruction-tuned (Chat); Reason = Reasoning; MoE = Mixture of Experts; B = Billions of parameters.

D Buckets Effect

We demonstrate the Buckets effect on ND-MT systems across all lexical metrics in Table 8. The results indicate that the worst-performing candidate consistently determines system ranking across different sampling sizes, exhibiting the highest ranking correlation compared to other aggregation methods (min, max, mean, random). Specifically, the TER metrics do not show consistency across all group-based metrics. This indicates the unreliability of TER in measuring the ND-MT.

Table 5: The Original Lexical-based Metrics Results of D-MT Models on 23 EN-ZH

Model	BLEU	MET	R-1	R-2	R-L	chrF	TER
NMT							
mBART-50	31.41	46.90	55.98	27.72	52.98	25.34	145.50
NLLB-600M	26.02	36.81	48.02	24.31	45.18	21.03	108.01
NLLB-3.3B	26.34	36.84	48.20	25.72	45.50	21.78	123.47
NLLB-54B	24.17	33.72	45.35	24.43	42.90	20.44	111.43
Pre-trained (7-8B)							
Llama2-7B	28.32	45.71	56.11	27.11	52.70	24.54	102.84
Llama3-8B	37.60	54.77	63.07	35.48	59.67	31.63	103.47
Qwen2.5-7B	44.00	61.46	67.89	42.31	64.33	36.69	**98.25**
Pre-trained (70-72B)							
Llama2-70B	40.53	58.64	65.89	39.13	62.29	33.74	101.76
Llama3-70B	44.19	61.89	68.08	42.43	64.47	37.38	99.27
Qwen2.5-72B	**48.49**	**65.85**	**70.80**	**46.98**	**67.62**	**40.36**	98.38
Chat (7-8B)							
Llama2-C-7B	15.39	29.23	34.64	13.51	32.14	13.56	756.12*
Llama3-C-8B	35.58	53.24	61.11	33.77	57.51	30.00	108.84
Qwen2.5-C-7B	39.43	58.02	64.53	37.28	61.03	32.90	104.77
Qwen3-C-8B	41.97	60.52	66.00	40.16	62.76	34.91	99.35
Chat (70-72B)							
Llama2-C-70B	16.04	36.90	33.75	15.36	31.13	16.13	2169.34*
Llama3-C-70B	42.13	59.99	66.07	40.17	62.78	35.48	99.09
Qwen2.5-C-72B	45.88	63.89	69.13	44.30	65.77	38.20	103.38
Reasoning							
MiniCPM-8x2	42.30	59.68	66.36	40.51	63.02	34.47	107.09
Qwen3-R-8B	40.39	57.86	63.56	38.57	60.60	33.66	1551.24*
DS-Llama-8B	33.61	50.95	59.34	31.33	55.51	27.84	183.59
DS-Qwen-7B	30.35	49.12	57.25	28.19	53.27	25.76	132.72
DS-671B	26.77	43.46	50.87	24.19	47.99	23.77	114.91

MET=METEOR; R-1/2/L=ROUGE-1/2/L; chrF=chrF++; C=Chat; R=Reason; DS=DeepSeek.
*Exceptionally high TER values indicate potential issues. Lower TER is better; higher is better for other metrics.

Table 6: The Original Semantic-based Metrics Results of D-MT Models on 23 EN-ZH

Model	KIWI	BLE	BERT	COMET	LASER	LaBSE	SentT	XNLI
NMT								
mBART-50	75.24	58.97	86.32	80.81	82.44	84.09	13.00	**97.80**
NLLB-600M	67.08	52.74	83.58	75.36	78.41	77.28	10.75	97.08
NLLB-3.3B	66.06	52.11	83.13	75.72	75.82	73.78	10.62	96.07
NLLB-54B	63.48	49.76	81.85	74.75	72.16	69.06	9.79	92.69
Pre-trained (7-8B)								
Llama2-7B	71.96	54.39	84.63	78.77	79.31	79.89	14.02	93.96
Llama3-8B	77.12	61.15	87.46	83.73	81.88	84.24	14.81	97.32
Qwen2.5-7B	79.47	64.03	88.73	85.97	82.27	85.18	15.16	98.03
Pre-trained (70-72B)								
Llama2-70B	76.61	60.99	87.87	83.18	82.57	85.36	14.27	97.76
Llama3-70B	78.63	63.72	88.74	85.40	82.94	85.72	15.08	97.89
Qwen2.5-72B	**80.40**	**66.25**	**89.80**	**86.94**	82.82	86.15	14.85	**98.32**
Chat (7-8B)								
Llama2-C-7B	58.58	35.67	76.58	64.75	78.35	77.17	32.93*	76.42
Llama3-C-8B	78.08	59.95	86.68	84.01	80.69	83.29	15.34	97.93
Qwen2.5-C-7B	79.27	62.08	87.75	85.40	81.92	85.11	15.20	98.05
Qwen3-C-8B	80.70	63.53	88.35	86.30	82.31	85.77	14.23	98.18
Chat (70-72B)								
Llama2-C-70B	57.36	32.23	70.85	53.04	71.95	76.19	41.27*	69.03
Llama3-C-70B	80.02	63.42	88.36	86.07	81.88	84.79	14.60	98.03
Qwen2.5-C-72B	80.57	65.13	89.22	86.78	**82.76**	**85.99**	14.60	98.11
Reasoning								
MiniCPM-8x2	78.71	62.86	88.29	84.94	82.07	84.74	13.37	97.92
Qwen3-R-8B	79.21	62.38	87.25	84.62	81.37	84.26	15.43	96.77
DS-Llama-8B	75.79	57.68	85.53	81.64	80.77	82.21	13.42	96.21
DS-Qwen-7B	74.05	55.00	84.91	80.43	81.36	83.10	16.74	97.55
DS-671B	76.54	54.51	79.85	81.05	75.02	77.37	13.66	92.34

KIWI=COMETKIWI; BLE=BLEURT; BERT=BERTScore; COMET=COMETDA; SentT=SentTrans; C=Chat; R=Reason; DS=DeepSeek.
*Anomalous SentTrans values. Higher is better for all metrics. Best results in **bold**.

Table 7: Correlation Results of Semantic-based Metrics on WMT23 EN-ZH for 22 ND-MT Systems.

Strategy	BERTScore	BLEURT	COMETDA	COMETKIWI
Kendall's τ / p-value				
Min	.58/.00	.64/.00	.74/.00	.77/.00
Max	.57/.00	.59/.00	.67/.00	.70/.00
Mean	.63/.00	.66/.00	.73/.00	.77/.00
Random	.63/.00	.67/.00	.73/.00	.77/.00
Std	-.57/.00	-.64/.00	-.71/.00	-.76/.00
Spearman's ρ / p-value				
Min	.72/.00	.81/.00	.87/.00	.89/.00
Max	.74/.00	.78/.00	.83/.00	.85/.00
Mean	.79/.00	.84/.00	.87/.00	.89/.00
Random	.79/.00	.85/.00	.87/.00	.89/.00
Std	-.69/.00	-.79/.00	-.87/.00	-.89/.00

Table 8: Correlation Analysis of MT Evaluation Metrics Across Sampling Sizes and Selection Strategies on WMT23 EN-ZH with Five SOTA ND-MT Systems

Size	Strategy	BLEU		GLVS		METEOR		ROUGE-1	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ
20	Max	.70/.19	.60/.23	1.0/.00	1.0/.02	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Mean	.90/.04	.80/.08	.90/.04	.80/.08	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Min	1.0/.00	1.0/.02	1.0/.00	1.0/.02	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Random	.90/.04	.80/.08	.90/.04	.80/.08	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Std	.70/.19	.60/.23	.90/.04	.80/.08	1.0/.00	1.0/.02	.82/.09	.74/.08
50	Max	.70/.19	.60/.23	.90/.04	.80/.08	.90/.04	.80/.08	1.0/.00	1.0/.02
	Mean	.90/.04	.80/.08	.90/.04	.80/.08	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Min	1.0/.00	1.0/.02	1.0/.00	1.0/.02	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Random	.90/.04	.80/.08	.90/.04	.80/.08	1.0/.00	1.0/.02	1.0/.00	1.0/.02
	Std	.70/.19	.60/.23	1.0/.00	1.0/.02	.97/.00	.95/.02	.82/.09	.74/.08

Size	Strategy	ROUGE-2		ROUGE-L		TER		chrF++	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ
20	Max	.90/.04	.80/.08	.90/.04	.80/.08	.90/.04	.80/.08	.90/.04	.80/.08
	Mean	1.0/.00	1.0/.02	1.0/.00	1.0/.02	.90/.04	.80/.08	1.0/.00	1.0/.02
	Min	1.0/.00	1.0/.02	1.0/.00	1.0/.02	.40/.50	.40/.48	1.0/.00	1.0/.02
	Random	1.0/.00	1.0/.02	1.0/.00	1.0/.02	.90/.04	.80/.08	1.0/.00	1.0/.02
	Std	.92/.03	.88/.05	.82/.09	.74/.08	.90/.04	.80/.08	1.0/.00	1.0/.02
50	Max	.80/.10	.60/.23	1.0/.00	1.0/.02	.90/.04	.80/.08	.90/.04	.80/.08
	Mean	1.0/.00	1.0/.02	1.0/.00	1.0/.02	.90/.04	.80/.08	1.0/.00	1.0/.02
	Min	1.0/.00	1.0/.02	1.0/.00	1.0/.02	.90/.04	.80/.08	1.0/.00	1.0/.02
	Random	1.0/.00	1.0/.02	1.0/.00	1.0/.02	.80/.10	.60/.23	1.0/.00	1.0/.02
	Std	.92/.03	.89/.04	.76/.13	.67/.12	.90/.04	.80/.08	.90/.04	.80/.08

ρ = Spearman's correlation coefficient; τ = Kendall's tau coefficient. Values shown as coefficient/p-value.