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

Weichuan Wang[1,2], **Mingyang Liu**[1], **Linqi Song**[1,2]*, **Chen Ma**[1,2]*
[1]City University of Hong Kong
[2]City University of Hong Kong Shenzhen Research Institute
{weicwang2-c, mingyaliu8-c}@my.cityu.edu.hk
{linqsong, chenma}@cityu.edu.hk

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 multimodality 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 state-of-the-art ND-MT systems using both lexical-based and semantic-based metrics at varying sampling sizes. The results reveal a Buckets Effect across these systems: the ranking of ND-MT systems is dominated by the worst-quality candidate translation, as shown by automatic evaluation metrics. To mitigate this issue, we propose ExpectoSample, a strategy that first identifies reliable metrics and then enables robust ND-MT system selection for real-world.

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 benefits of the non-deterministic (ND) properties (Atil et al., 2025; Song et al., 2025) of LLMs on their potential for flexible and customized chat-box applications (DeepSeek-AI, 2025; OpenAI et al., 2024; Yang et al., 2025). Recent studies have progressed fine-grained exploration and analysis of this property, primarily focusing on deterministic tasks such as question answering (Song et al., 2025; Kuhn et al., 2023). However, the impact of ND on machine translation, a complex and ND task, remains under-explored both on the generation and evaluation aspects. To bridge this research gap, we examine modern **N**on-**D**eterministic **Machine Translation** (**ND-MT**) systems for both the potential benefits and the evaluation challenges associated with non-determinism.

We first examine one of the most prominent challenges in MT: multimodality (Papineni et al., 2002; Bao et al., 2023), a phenomenon where a single source sentence may correspond to multiple valid translation candidates due to contextual ambiguity. This challenge is particularly problematic for automatic evaluation due to the scarcity of comprehensive reference sets (Papineni et al., 2002; Popović, 2015; Rei et al., 2022b). While previous research has employed human assessment (Kocmi et al., 2024, 2023, 2025) to mitigate this issue, such approaches are increasingly unscalable due to the prohibitive cost of constant re-evaluation necessitated by 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 have the potential to satisfy both principles simultaneously. This is observed in practical applications where the same translation prompt yields diverse yet acceptable outputs. In this work, we sys-

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 7677–7701

tematically investigate the potential of ND-MT for addressing multimodality, focusing on its ability to provide both lexical diversity and semantic equivalence. Our analysis covers 22 modern MT systems across six language directions under a uniform temperature setting (0.5) (see Appendix A for details). Additionally, we introduce a reference-free, group-level lexical metric termed the Group Lexical Variance Score (GLVS) to mitigate the bias resulting from limited references when evaluating ND-MT systems. We employ both lexical and semantic metrics to quantify the impact of non-determinism on lexical variability and meaning preservation, respectively. Our results demonstrate significant lexical diversity alongside nearly identical semantic content when compared to D-MT systems using the same underlying models. Furthermore, we investigate how the value of temperature affects the final system performance of ND-MT. The results reveal that while various temperature settings can induce lexical diversity, semantic equivalence is only preserved at lower temperatures. Consequently, we characterize modern MT systems as temperature-constrained ND-MT systems.

However, the evaluation of ND-MT remains under-explored. We first investigate the feasibility of directly adopting performance from D-MT counterparts using current evaluation schemes (Kocmi et al., 2024, 2023, 2025). To bypass the high cost of human assessment, we explore this direct adaptation through existing automatic evaluation metrics (Papineni et al., 2002; Popović, 2015; Rei et al., 2022b) that exhibit high correlation with human judgment. Specifically, we apply these via group-level measurements: *min*, *max*, *mean*, *random*, and *std* (standard deviation), capturing the group-level performance of ND-MT systems. The resulting inconsistent rankings demonstrate the unreliability of traditional D-MT evaluation schemes when applied to the ND-MT. Furthermore, we examine how sampling size ($\{10, 20, 50\}$) affects evaluation results with five state-of-the-art ND-MT systems at a fixed temperature (0.5). The results reveal a strong Buckets Effect: for each source, the lowest-quality candidate largely determines system rankings across sample sizes. This highlights the inherent risk in evaluating non-deterministic systems, as the minimum performance quality is stochastically hidden and cannot be predicted before generation. To mitigate this, we propose the ExpectoSample strategy, which first identifies reliable metrics and subsequently selects robust ND-MT systems.

Our contributions are threefold: (1) We demonstrate that ND-MT systems effectively address the multimodality challenge by providing lexical diversity while maintaining semantic equivalence under specific temperature constraints. (2) We uncover the Buckets Effect in ND-MT evaluation—where system rankings are predominantly determined by the lowest-quality candidates, and propose the ExpectoSample strategy to mitigate this challenge by identifying reliable metrics for robust system selection. (3) We conduct a systematic investigation of 22 ND-MT systems across six language directions using 11,947 source cases. To support future research, we release our complete code, dataset, and evaluation results at [1].

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 (Vaswani et al., 2017) on multilingual text then fine-tuned on bilingual text (Team et al., 2022), and decoder-only architectures pre-trained on multilingual text without specific fine-tuning requirements (Brown et al., 2020a). From the 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 prompts (Vilar et al., 2023). Different LLM-based MT approaches exhibit distinct characteristics: pre-training-only MT systems typically use few-shot methods (Brown et al., 2020b; 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).

[1] https://github.com/weichuanW/TC-DN-MT

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 under-explored.

2.2 Non-determinism of LLMs

Previously, substantial effort was focused on deterministic tasks such as sentiment classification (Zhang et al., 2024) and parsing (Ginn and Palmer, 2025), with most attention directed toward utilizing deterministic capabilities from LLMs. In recent years, the non-determinism (ND) property of LLMs has emerged as a significant area of interest and have been leveraged to satisfy customized user requirements (Tseng et al., 2024). Most models now implement non-determinism as a default property (DeepSeek-AI, 2025; Yang et al., 2025), enabling LLMs to provide a variety of reasonable outputs for the same prompt to enhance 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, research of ND of LLMs in complex tasks like MT remains underexplored.

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 multimodality 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.

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 by selecting state-of-the-art architectures, including both encoder-decoder and decoder-only models of varying scales. We then generate multiple translation candidates and evaluate them using automatic metrics via group-level measurements. Our analysis reveals that ND-MT effectively addresses the multi-modality challenge by providing lexical diversity while maintaining semantic equivalence under specific temperature constraints. Based on these observations, we characterize modern MT systems as temperature-constrained ND-MT systems.

3.1 Experimental Preparation

3.1.1 ND-MT Systems

We consider both encoder-decoder and decoder-only architectures for modern MT. 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 and 8B parameters) and large-scale (70, 72 and 671B parameters) variants across pre-trained, instruction-tuned, and reasoning types when available. The detailed information are provided in Appendix A.

3.1.2 Dataset Statistics

In this work, we adopt sentence-level MT and leverage existing, well-established open-source datasets to study both ND-MT and their counterpart D-MT systems. Specifically, we use the latest WMT data from 2023–2024 [2] across six translation directions

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

Table 1: Dataset Statistics Information

Source	Translation Direction	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

(ZH↔EN, EN↔DE, EN↔RU), covering three language pairs. We identify ⟨English, Chinese⟩ translation as particularly valuable for investigation due to substantial differences in language families and choose it as our primary experimental setting to explore the potential of ND-MT. Further evaluation on ⟨English, German⟩ and ⟨English, Russian⟩ are made to demonstrate the general potential of ND-MT across diverse language pairs. We present detailed statistics in Table 1.

3.1.3 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. These metrics use the reference as an anchor to show the lexical diversity.

Semantic-based Methods For semantic equivalence, we employ COMETKIWI(Rei et al., 2022a) and COMETDA (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) A primary drawback of current lexical-based metrics is their reliance on gold-standard references (Papineni et al., 2002; Popović, 2015; Lin, 2004; Snover et al., 2006), which are typically unavailable for ND-MT systems in real-world scenarios. While some in-group lexical evaluation methods (Zhu et al., 2018) adapt the principles of BLEU (Papineni et al., 2002) or ChrF++ (Popović, 2015) to measure overlapping between candidates, these approaches often lack discriminative power. Specifically, when candidates are highly similar or significantly different, these metrics tend toward extreme values (zero or one), failing to capture nuanced lexical diversity. Conversely, simple word-counting strategies (Liu et al., 2022) mitigate these extreme cases but fail to account for the lexical relationships, such as shared vocabulary between candidates. To address these limitations, we propose the Group Lexical Variance Score (GLVS) to quantify lexical diversity by establishing a group-level vocabulary and computing the frequency distribution of unique tokens across the candidate set, capturing both individual variance and inter-candidate relationships.

The computation of the Group Lexical Variance Score (GLVS) proceeds in three stages:

1. Candidate Tokenization Each candidate c_i in the generated set $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ is tokenized into a sequence of words $\mathcal{W}_i = \{w_1, w_2, \ldots, w_l\}$. To focus on lexical variety, we define $\mathcal{W}_i^U$ as the set of unique words for candidate c_i.

2. Collective Vocabulary Frequency We define the collective word pool $\mathcal{V}_{total}$ as the vocabulary containing all words across all N candidates. Let M be the total word count in $\mathcal{V}_{total}$. For each unique word w, its relative frequency $f(w)$ is calculated as:

$$f(w) = \frac{\text{count}(w, \mathcal{V}_{total})}{M} \tag{1}$$

3. GLVS Computation and Aggregation For each individual candidate c_i, we calculate a candidate-level score by summing the relative frequencies of its unique constituent words:

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

To characterize the overall behavior of an ND-MT system for a specific source sentence, we aggregate these individual scores at the group level using the mean (μ_{GLVS}):

$$\mu_{GLVS} = \frac{1}{N} \sum_{i=1}^{N} GLVS(c_i) \tag{3}$$

In this framework, μ_{GLVS} serves as an inverse proxy for lexical diversity. A low μ_{GLVS} value indicates that candidates are primarily composed of

words that appear rarely within the group, thereby signaling high lexical diversity. Conversely, a high value suggests lexical redundancy, where candidates converge on a narrow set of common terms.

Furthermore, other group-level measurements, like the standard deviation, can provide additional analytical depth. For instance, a high standard deviation indicates that the ND-MT system is unstable in its generation quality, producing candidates with significantly varying degrees of lexical diversity.

3.1.4 Experimental Settings

Decoding Strategy We employ greedy decoding as the deterministic baseline and sampling-based decoding in the non-deterministic setting with adjustable temperature, generating K candidates per source. We use temperature 0.5 and sampling size 10, motivated by a previous study (Kuhn et al., 2023), to investigate the potential of ND-MT.

Group-based Measurements For each source, we compute their group-based measurements: *min*, *max*, *mean*, *random*, and *std* (standard deviation) for various metrics and aggregate them on the dataset level to capture the system performance of ND-MT.

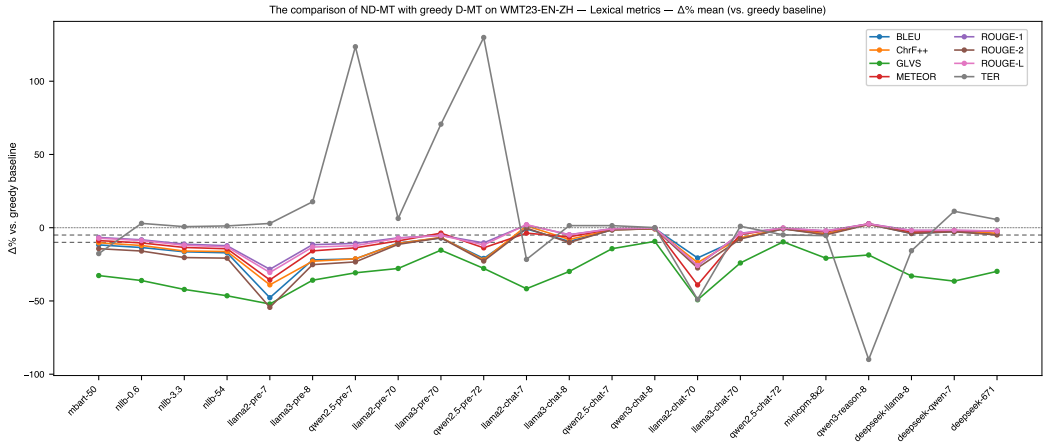
3.2 The Potential of ND-MT to Solve Multimodality

To explore the potential of ND-MT in addressing the multimodality 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 such as top-p (Holtzman et al., 2020), top-k (Noarov et al., 2025). We run the corresponding D-MT systems as baselines to enable direct comparison, where each system generates only one candidate.

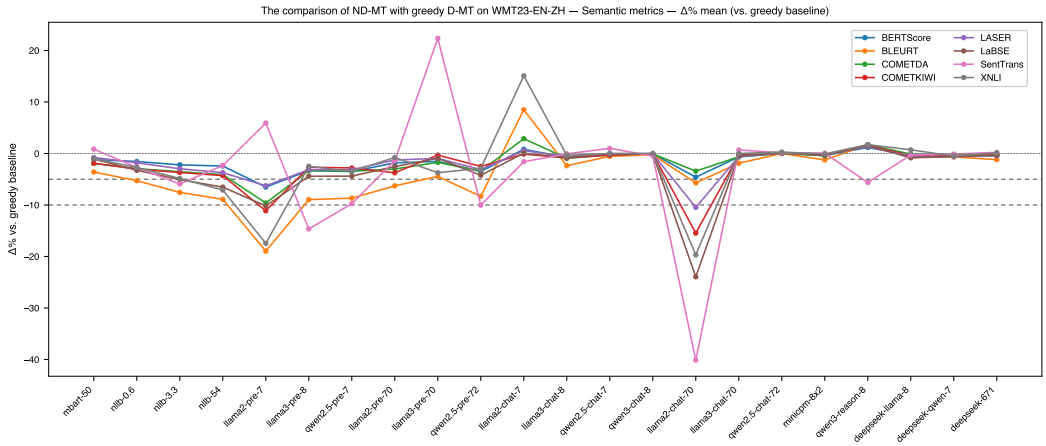
ND-MT provides salient lexical diversity Figure 1a illustrates the lexical diversity captured by various evaluation metrics. We first observe that traditional reference-based metrics (Papineni et al., 2002; Popović, 2015; Banerjee and Lavie, 2005; Lin, 2004; Snover et al., 2006) detect only weak lexical diversity, with the majority of delta values falling below 10%. In contrast, our proposed GLVS metric identifies at least 5% lexical variation across all models, with most values exceeding 10%, indicating substantial lexical diversity. While standard metrics can detect diversity in specific cases—such as Llama-2-7b (pre-trained) (Touvron et al., 2023) and Llama-2-70b-chat (instruct-tuned) (Touvron et al., 2023), they fail to do so for the majority of models. Furthermore, the trends identified by GLVS and reference-based metrics are sometimes divergent. For instance, the significant lexical diversity of Llama-2-7b-chat is captured by GLVS but remains undetected by all other metrics. This highlights the inherent insensitivity of reference-based metrics, which suffer from an over-reliance on limited gold-standard references and an inability to account for potential references caused by multimodality. Finally, GLVS values can be used to reflect the magnitude of lexical diversity across different ND-MT systems. Because deterministic MT systems score 100 and we utilize delta values, lower GLVS scores indicate higher diversity. Besides, Although reference-based metrics are weak on detecting lexical diversity, their small delta values demonstrate that ND-MT systems generate high-quality results like D-MT. A primary advantage of GLVS is its practical utility in real-world deployment, as it eliminates the need for reference translations. In the main body, we focus on the overview analysis, and we list the concrete analysis and more results in Figure 5a and Appendix C.

ND-MT maintains the semantic equivalence Figure 1b illustrate the mean delta results for semantic metrics. We first find that 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 Llama-2-7b (pre-trained) (Touvron et al., 2023) and Llama-2-70b-chat (instruct-tuned) (Touvron et al., 2023). Apart from that, the Std delta results from Figure 5b almost remain below 10 percentage points across all metrics, demonstrating the MT systems maintain strong semantic equivalence under non-deterministic settings. In-depth discussion and baseline results are provided in Appendix C.

ND-MT has the potential to provide better candidates than D-MT We further investigate the capacity of ND-MT systems to generate superior-quality candidates. Specifically, we identify the best-performing candidate within each group according to each metric and compute the average maximum scores across the dataset. It is important to note that in real-world deployment, reference translations are typically unavailable; consequently, practical systems must rely on selection



(a) Mean Delta Results of WMT23 En→Zh on Lexical Metrics.



(b) Mean Delta Results of WMT23 En→Zh on Semantic Metrics.

Figure 1: Mean Delta results for lexical and semantic metrics on WMT23 En→Zh ($T = 0.5$, 10 candidates). Delta results are calculated relative to greedy decoding on identical data and models. Thresholds of -5 and -10 (dotted line) are included to indicate levels of significance.

strategies such as Minimum Bayes Risk (MBR) decoding (Müller and Sennrich, 2021) with auxiliary ranking functions (González-Rubio and Casacuberta, 2013). Therefore, these *Max Delta* results should be interpreted as the theoretical potential of ND-MT to produce high-quality translations. As illustrated in Figure 6, ND-MT systems demonstrate a consistent and substantial performance gain on both lexical and semantic, revealing their capacity to generate higher-quality outputs than deterministic baselines. Furthermore, our findings validate the existence of a high quality upper bound, suggesting that deterministic selection methods (like MBR) have significant room to improve final translation results.

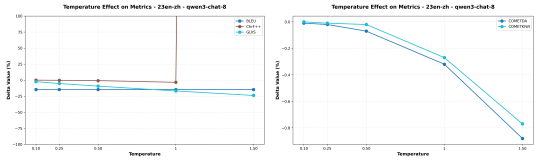
Generality of ND-MT in Addressing Multimodality Finally, we evaluate the generality of ND-MT potential across different language pairs. We test ⟨German, English⟩ and ⟨Russian, English⟩ in both directions with five state-of-the-art LLM-based MT models (Touvron et al., 2023; Qwen et al., 2025; Yang et al., 2025). The results in Figure 7 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 while maintain semantic equivalence, effectively addressing multimodality 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.

3.3 Temperature Constraints on ND-MT

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 D

We evaluate both lexical diversity and semantic equivalence using the same metrics from Section 3.1.3. For lexical analysis, we select GLVS as a reference-free metric, and BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2015) as reference-based metrics that measure effects at the lexical and character levels, respectively. For semantic analysis, we choose COMETDA (Rei et al., 2020) and COMETKIWI (Rei et al., 2022a) as reference-based and reference-free metrics, respectively. Figure 2 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 multimodality 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 2: The temperature effect for 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 Limitations of the Current D-MT Evaluation Scheme on ND-MT

In Section 3.3, we demonstrate the potential of ND-MT to address multimodality 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 multimodality 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 multimodality issues. However, their effectiveness remains constrained by the scale of the training data and computational resources. 2) Evaluating ND-MT systems using humans is increasingly impractical. Because these systems generate a high volume of translation candidates, and because those outputs change significantly depending on the temperature setting, researchers would need to conduct

a massive number of evaluations to reach a reliable conclusion. This creates a *bottleneck* where human assessment becomes far too slow and expensive to be sustainable.

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, which reveals significant inconsistencies. Second, we evaluate current metrics using group-based measurements and identify the Buckets Effect in ND-MT that influences ranking determination. Finally, we propose the *ExpectoSample* strategy to identify reliable metrics for selecting robust ND-MT systems.

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 9 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 deterministic approaches.

4.3 Buckets Effect of ND-MT

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 COMETDA and COMETKIWI as semantic metrics. The results are presented in Tables 3 and 11. A key observation is the ***Buckets Effect***: the worst-cases of ND-MT systems determine the system ranking across all sampling sizes and metrics. Our findings demonstrate that a controlled sampling sizes rather than arbitrarily large samples—can yield reliable evaluations with existing metrics. However, the Buckets Effect indicates the difficult for evaluation the real system performance of ND-MT since the worst-case is unknown in advance and hard to find. However, we notice that finding reliable metrics may possible but need careful filtering (concrete discussed about the metrics in Appendix F).

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

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 $p < 0.10$.

4.4 ExpectoSample: Identifying Reliable Metrics For Selecting Robust Systems

The Buckets Effect presents significant challenges for the reliable evaluation and practical deployment of ND-MT systems. However, our empirical analysis (Tables 3 and 11) reveals that certain robust metrics, specifically ChrF++ (Popović, 2015), COMET-20-DA (Rei et al., 2020), and COMET-22-Kiwi (Rei et al., 2022a) maintain consistent system rankings under both *mean* and *random* sam-

pling settings. This consistency is essential for real-world since the computation cost is limited for sampling evaluation.

Motivated by these findings, we propose **ExpectoSample**, a two-stage framework designed to first filter reliable metrics and subsequently select stable ND-MT systems. The framework is built on the principle that a truly reliable metric should produce consistent system rankings regardless of sample size, while a robust ND-MT system should maintain stable performance across varying sample counts. The ExpectoSample strategy consists of the following two steps: **1) Metric Reliability Filtering:** We examine the system-ranking correlations across a set of increasing sampling sizes $\mathcal{X} = \{X_1, X_2, X_3\}$, where $2X_1 \leq X_2$ and $2X_2 \leq X_3$. Given a set of ND-MT systems and candidate metrics, we retain only those metrics whose *mean* ranking correlation (e.g., Spearman's ρ) remains above a stability threshold ϵ (where $0 < \epsilon \leq 1$) across all sampling pairs. **2) System Selection and Deployment:** Utilizing the filtered reliable metrics from Step 1, we rank the available ND-MT systems using a small sample size X_k (typically $X_k \leq 10$). The top-k performing systems are then identified as the most reliable candidates for production-level usage.

Notably, the set of ND-MT systems used for metric filtering in Step 1 does not need to be identical to the systems evaluated in Step 2. This decoupling enhances the generalizability of the ExpectoSample strategy, as metrics proven reliable on a reference benchmark can be confidently applied to rank new or unseen systems in real-world deployment.

5 Discussion and Future Directions

Exploring the Temperature-Constrained Nature of ND-MT While we focused on temperature due to its accessibility in API-based models, other parameters like *top-k* (Noarov et al., 2025) and *top-p* (Holtzman et al., 2020) warrant further investigation to achieve a fine-grained understanding of non-determinism. This study establishes a general framework for assessing ND-MT potential, though future work should evaluate how diverse sampling strategies interact with these findings.

Upper Bounds for MBR and Re-ranking Strategies Our results highlight a significant performance gap between mean and maximum candidate quality, providing a theoretical justification for the success of Minimum Bayes Risk (MBR) (Müller and Sennrich, 2021) and re-ranking methods (González-Rubio and Casacuberta, 2013). While these selection techniques aim to capture high-quality candidates, our analysis of the Buckets Effect suggests that their success is inherently bounded by the underlying system's generation capability.

Limitations in Automated and Human Evaluation We omitted LLM-as-a-Judge (Kim, 2025; Kocmi and Federmann, 2023) due to inherent biases of LLMs on preferring LLMs' generations (Ye et al., 2025), relying instead on established lexical and semantic metrics to ensure objective evaluation. Future research can develop cost-effective human-in-the-loop (Schroeder et al., 2025) protocols that can reliably assess non-deterministic outputs without the prohibitively high costs of full human assessment.

The Buckets Effect: A Metric for System-Level Improvement The Buckets Effect demonstrates that ND-MT performance is governed by systemic differences and vocabulary-level lower bounds rather than isolated successes on specific cases. This phenomenon suggests that meaningful progress in MT, such as through RLHF (Ouyang et al., 2022; Lee, 2025) or instruction tuning (Rios, 2025) should be measured by shifts in the entire performance distribution rather than single-output improvements.

6 Conclusion

In this work, we systematically investigate the potential and evaluation challenges of ND-MT. Our findings reveal their significant potential to address the long-standing multimodality challenge in MT by generating candidates with pronounced lexical diversity while maintaining semantic equivalence under specific temperature constraints. Furthermore, we identify critical evaluation challenges unique to ND-MT, specifically the inconsistency of system rankings compared to deterministic counterparts and the emergence of the *Buckets Effect*. We propose *ExpectoSample*, a robust strategy designed to filter for reliable evaluation metrics and identify robust ND-MT systems. Ultimately, this research provides both empirical evidence and a methodological framework for the more rigorous assessment and deployment of ND-MT.

Acknowledgments

This work is supported by the Early Career Scheme (No.CityU 21219323) and the General Research Fund (No.CityU 11220324) of the University Grants Committee (UGC), the NSFC Young Scientists Fund (No.9240127), the Donation for Research Projects (No.9229164 and No.9229216). Additional support is provided by the Research Grants Council of the Hong Kong SAR under Grant GRF 11217823, 11216225, and Collaborative Research Fund C1042-23GF. This work is also funded by the National Natural Science Foundation of China under Grant 62371411 and the InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies.

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 Buckets 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. Apart from that, 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. Finally, while the Buckets Effect is robustly validated through empirical evidence, establishing a formal theoretical framework remains a necessary step for its broader generalization.

Ethical Statement

Our research on non-deterministic machine translation may raise 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 ade-

quate 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.

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 Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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. 2020b. 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.

Jesús González-Rubio and Francisco Casacuberta. 2013. Improving the minimum Bayes' risk combination of machine translation systems. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text

degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.

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.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, 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.

Jieh-Sheng Lee. 2025. Instructpatentgpt: training patent language models to follow instructions with human feedback. *Artif. Intell. Law*, 33(3):739–782.

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.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.

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.

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.

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. 2025. Foundations of top-*k* decoding for language models. *CoRR*, abs/2505.19371.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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.

Miguel Rios. 2025. Instruction-tuned large language models for machine translation in the medical domain. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 162–172, Geneva, Switzerland. European Association for Machine Translation.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795, Vienna, Austria. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

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.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Model	Variant	Params (B)	Type
Encoder-Decoder Models			
mBART	NMT	0.68	Dense
NLLB-200	NMT	0.6, 3.3, 54	Dense
Decoder-Only: Llama Family			
Llama 2	Base	7, 70	Dense
Llama 2	Chat	7, 70	Dense
Llama 3	Base	8, 70	Dense
Llama 3	Chat	8, 70	Dense
Decoder-Only: Qwen Family			
Qwen 2.5	Base	7, 72	Dense
Qwen 2.5	Chat	7, 72	Dense
Qwen 3	Chat	8	Dense
Qwen 3	Reasoning	8	Dense
Decoder-Only: DeepSeek Family			
DeepSeek (Llama)	Reasoning	8	Dense
DeepSeek (Qwen)	Reasoning	7	Dense
DeepSeek-R1	Reasoning	671	MoE
Other Decoder-Only Models			
MiniCPM	Chat	16	MoE

Abbreviations: Base: Pre-trained only; **SFT**: Instruction-tuned; **Reasoning**: Reinforcement Learning tuned.

Table 4: Specifications of language models used in our experiments. We report the model family, specific training variant (Base, Chat, or Reasoning), parameter counts, and the underlying architecture type (standard Dense versus Mixture-of-Experts models).

A Model Statistics and Implementation

We summarize the evaluated systems in Table 4. The table specifies the model family, parameter count, and architecture type. Additionally, we distinguish between model variants (Base, Chat, Reasoning) to indicate the corresponding machine translation prompts applied during inference. Apart from that, all models were obtained from the Hugging Face Hub under open-source licenses.[3] Evaluation metrics were implemented using the Hugging Face `evaluate` library.[4] Specifically, for the semantic metrics, we utilize the `Unbabel/wmt22-comet-da`[5] and `Unbabel/wmt22-cometkiwi-da`[6] checkpoints to represent COMETDA (Rei et al., 2020) and COMETKIWI (Rei et al., 2022a), respectively.

[3] https://huggingface.co/models

[4] https://github.com/huggingface/evaluate

[5] https://huggingface.co/Unbabel/wmt22-comet-da

[6] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

B Prompting Strategies

To ensure fair evaluation across varying architectures, we tailor our prompting strategies to the specific training stage of each model, as detailed in Table 5.

Encoder-Decoder (NMT). For standard NMT systems like NLLB-200 and mBART, we adhere to the official default configurations. Specifically, we append the designated language identification tokens (e.g., `eng_Latn`, `zho_Hans`) to the input sequence to specify the translation direction. This tokenization process is automated using the standard preprocessing pipelines provided by the Hugging Face library (Wolf et al., 2019) [7].

Base Models (Pre-trained). For pre-trained decoder-only models, we employ **5-shot in-context learning** (ICL) to stabilize generation (Wang et al., 2024). As detailed in Table 5, the prompt consists of five fixed parallel demonstrations (e.g., English: "The weather is beautiful today" → Chinese: "今天天气很好。"), followed by the input query. This format effectively primes the model to adhere to the expected output structure: `English: [Input] Chinese:`. To ensure high quality, the demonstration examples were manually curated from LLM-generated candidates.

Chat and Reasoning Models. For instruction-tuned (Chat) and reasoning-optimized variants (including DeepSeek-R1), we use a structured zero-shot prompt. To avoid the common issue of chat models generating conversational filler (e.g., "Sure, here is the translation..."), we explicitly constrain the output with the instruction: *"Only provide the translation, no explanations."* The input is formatted as a single user message containing the source text and the target language specification.

C Evaluation Results on D-MT

C.1 Original Evaluation Results

We present the evaluation results for Deterministic Machine Translation (D-MT) on the WMT23 English→Chinese (EN-ZH) dataset. Table 6 summarizes the performance using standard lexical-based metrics, while Table 7 details the corresponding results for semantic-based metrics.

LLMs outperform traditional NMT baselines. As shown in Table 6, large language models

[7] https://huggingface.co/

Variant	Strategy	Input Template / Format
NMT	Language Tokens	`<src_lang_token> [Source Sentence]` *Example:* `zho_Hans Hello world`

Base	5-Shot In-Context	`Translate the following English sentences to Chinese:`
		`English: The weather is beautiful today.`
		`Chinese: 今天天气很好。`
		`... (3 other examples) ...`
		`English: <text>`
		`Chinese:`

Chat / Reasoning	Zero-Shot Instruction	`Translate the following <src_lang> text to <tgt_lang>. Only provide the translation, no explanations:`
		
		`<text>`

Table 5: Overview of prompting strategies. We use standard tokens for **NMT**, a fixed 5-shot template with human-curated LLM demonstrations for **Base** models, and constrained zero-shot instructions for **Chat/Reasoning** models. In practice, generic language placeholders in the templates are instantiated with the specific translation direction (e.g., "English" to "Chinese").

(LLMs) significantly surpass dedicated NMT systems. Among NMT baselines, mBART-50 is the strongest performer (31.41 BLEU), yet it is easily overtaken by modern 7B-scale LLMs. For instance, Llama3-8B achieves 37.60 BLEU, and Qwen2.5-7B reaches 44.00 BLEU, demonstrating that general-purpose pre-training is highly effective for translation even without task-specific architectural bias.

Scaling and Architecture Dominance. Performance scales consistently with model size. The Qwen2.5-72B model achieves state-of-the-art results across nearly all metrics, setting the benchmark at **48.49 BLEU** and **86.94 COMET**. Notably, the Qwen family consistently outperforms Llama models of comparable size (e.g., Qwen2.5-7B surpasses Llama3-8B by +6.4 BLEU), likely due to its stronger multilingual pre-training corpus.

The "Chat" Alignment Tax. Comparing Base models to their Chat variants reveals a mixed impact of instruction tuning. For the Llama 2 family, the Chat versions suffer a catastrophic performance drop (e.g., Llama2-7B drops from 28.32 to 15.39 BLEU), accompanied by exploding TER scores (756.12), indicating severe repetition or formatting issues. However, newer models like Llama 3 and Qwen 2.5 show minimal degradation—or even slight improvements—in their Chat variants, suggesting that modern alignment techniques (RLHF) have become more robust for translation tasks.

Reasoning Models Struggle with Form. Surprisingly, reasoning-optimized models (e.g., DeepSeek-R1, Qwen3-Reasoning) underperform compared to standard dense models. Despite its massive scale, DeepSeek-R1 (671B) achieves only 26.77 BLEU, lower than the 7B Base models. The semantic metrics in Table 7 confirm this trend (COMET 81.05 vs. 86.94 for Qwen2.5-72B). This suggests that "reasoning" reinforcement learning, while powerful for logic, may introduce verbosity or structural deviations that are penalized in standard translation evaluation.

API Instability Impacts Reasoning Models. Contrary to expectations based on parameter scale, reasoning-optimized models (e.g., DeepSeek-R1, DS-Qwen) significantly underperform standard dense models. Our error analysis reveals that this is largely due to **API instability** rather than inherent model capability. We observed frequent occurrences of empty responses caused by network timeouts or API overloading during inference. These null outputs are penalized heavily by lexical metrics—resulting in low BLEU scores (e.g., 26.77 for DeepSeek-R1)—and distort semantic evaluations, highlighting the reliability challenges of deploying API-based models for large-scale benchmarks.

C.2 Ranking Inconsistency and Metric Reliability

To assess the reliability of automated evaluation, we computed the relative rankings of all 18 models across the metrics, as detailed in **Table 8**. While top-tier models show some stability, the overall analysis reveals critical weaknesses in relying on single-generation outputs.

Dominance vs. Disagreement. At the top of the leaderboard, metrics largely align: **Qwen2.5-72B**

Table 6: 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.

achieves the Rank #1 position across nearly all categories (see **Table 8a** for BLEU and **Table 8b** for XNLI), confirming its status as the current state-of-the-art. However, outside the top rank, significant contradictions emerge. For instance, the **mBART-50** baseline ranks high on semantic embedding metrics (Rank #4 in LASER, **Table 8b**) but falls to the bottom tier on lexical overlap (Rank #16 in BLEU, **Table 8a**). This implies that while the model captures semantic intent, its surface realization diverges from the reference, a nuance that lexical metrics punish disproportionately.

The Fragility of Single-Metric Evaluation. Crucially, no two metrics produce an identical ranking order. We observe extreme divergence in the Chat-tuned models:

- **Llama2-Chat-70B** is ranked as the **best** model (Rank #1) by SentTrans, yet is rated as the **worst** (Rank #23) by COMET and BLEU (see **Table 8**).
- **NLLB-600M** is ranked #11 in TER (better than Llama2-Chat), yet #19 in COMET (worse than Llama2-Chat).

Table 7: Semantic-based metrics for D-MT models (En-Zh). We report COMETKIWI (KIWI), BLEURT (BLE), BERTScore (BERT), COMETDA (CMT), LASER (LSR), LaBSE (LBS), SentTrans (SNT), and XNLI.

Model	KIWI	BLE	BERT	CMT	LSR	LBS	SNT	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)								
L2-C-7B	58.58	35.67	76.58	64.75	78.35	77.17	32.93*	76.42
L3-C-8B	78.08	59.95	86.68	84.01	80.69	83.29	15.34	97.93
Q2.5-C-7B	79.27	62.08	87.75	85.40	81.92	85.11	15.20	98.05
Q3-C-8B	80.70	63.53	88.35	86.30	82.31	85.77	14.23	98.18
Chat (70-72B)								
L2-C-70B	57.36	32.23	70.85	53.04	71.95	76.19	41.27*	69.03
L3-C-70B	80.02	63.42	88.36	86.07	81.88	84.79	14.60	98.03
Q2.5-C-72B	80.57	65.13	89.22	86.78	**82.76**	**85.99**	14.60	98.11
Reasoning								
MiniCPM	78.71	62.86	88.29	84.94	82.07	84.74	13.37	97.92
Q3-R-8B	79.21	62.38	87.25	84.62	81.37	84.26	15.43	96.77
DS-Llama	75.79	57.68	85.53	81.64	80.77	82.21	13.42	96.21
DS-Qwen	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

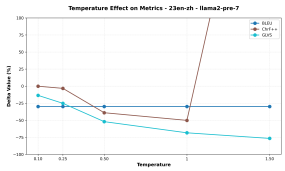
Metrics: KIWI=COMETKIWI; BLE=BLEURT; BERT=BERTScore; CMT=COMETDA; LSR=LASER; LBS=LaBSE; SNT=SentTrans.
Models: L=Llama; Q=Qwen; DS=DeepSeek; C=Chat; R=Reasoning.
*Anomalous SentTrans values. Best results in **bold**.

This misalignment underscores the danger of evaluating D-MT systems based on a single deterministic generation. Since greedy decoding represents only one point on the probability curve, it is susceptible to "lucky" or "unlucky" stylistic choices that metrics weight differently. This finding motivates our shift toward Non-Deterministic (ND-MT) evaluation to capture the model's full capability rather than a single, potentially biased output.

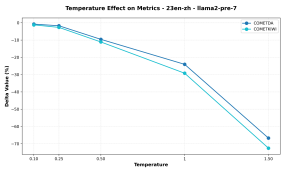
D Temperature Effect on ND-MT

We analyze the impact of temperature sampling on Non-Deterministic Machine Translation (ND-MT) across five state-of-the-art systems, including the Llama 2 family (Pre/Chat-7B) (Touvron et al., 2023) and the Qwen family (2.5-Pre/Chat-7B, 3-Chat-8B) (Qwen et al., 2025; Yang et al., 2025).

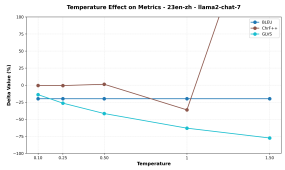
Semantic Equivalence vs. Temperature. As illustrated in **Figures 3 and 4**, there is a general



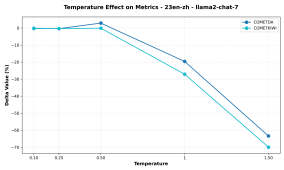
(a) Llama2-Pre-7B (Lexical)



(b) Llama2-Pre-7B (Semantic)



(c) Llama2-Chat-7B (Lexical)



(d) Llama2-Chat-7B (Semantic)

Figure 3: Temperature effect on Llama 2 models. While lexical metrics (left) remain stable, semantic metrics (right) show degradation at high temperatures.

downward trend in semantic metric scores as temperature increases, indicating that higher randomness often degrades translation fidelity. However, the optimal temperature is not always the greedy setting ($T = 0$). For instance, in specific datasets with `llama2-chat-7`, non-zero temperatures yield marginal improvements, suggesting that ND-MT can occasionally surpass deterministic baselines (D-MT) when tuned correctly.

Lexical Diversity and Metric Sensitivity. Regarding lexical analysis, we observe a distinct divergence between metrics. While BLEU (Papineni et al., 2002) scores remain nearly invariant across the temperature range, GLVS scores exhibit significant volatility (see **Figure 3(a) and (c)**). This demonstrates that BLEU fails to capture the subtle variations in lexical selection introduced by sampling. The shift in GLVS at higher temperatures suggests that models drift toward generating more "natural" language content rather than adhering strictly to the source fidelity, a nuance that standard lexical metrics overlook. These findings highlight the necessity of multi-dimensional evaluation for ND-MT systems.

Table 9: Correlation Results of Semantic-based Metrics on WMT23 EN-ZH for Five SOTA ND-MT Systems.

Strategy	BERT	BLEURT	COMETDA	KIWI
Kendall's τ / p-value				
Min (Worst)	.58/.00	.64/.00	.74/.00	.77/.00
Max (Best)	.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 (Worst)	.72/.00	.81/.00	.87/.00	.89/.00
Max (Best)	.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

KIWI=COMETKIWI.

E Case Study: The Impact of Decoding Temperature

Table 12: Quantitative analysis of character ratio vs. temperature.

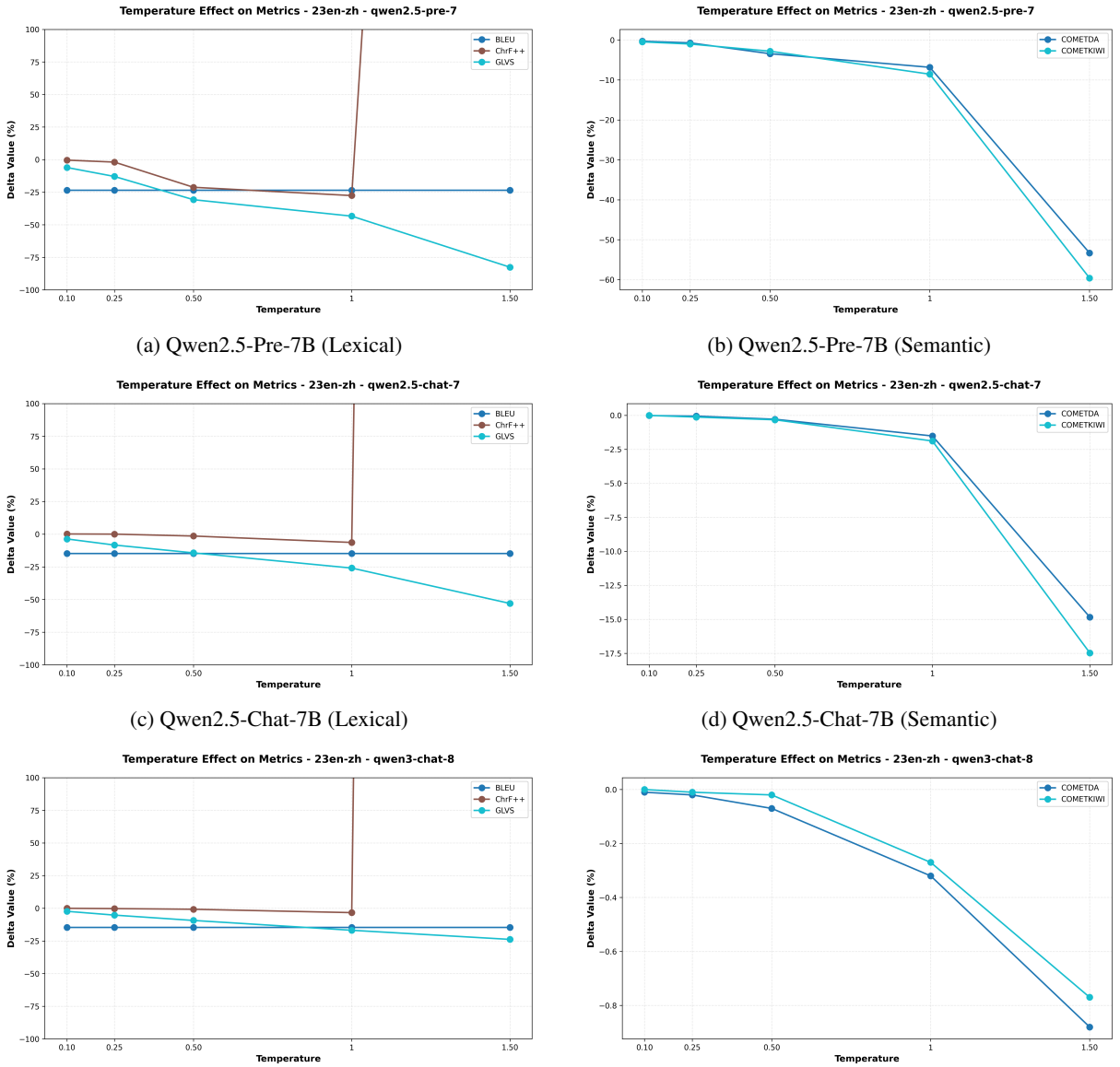
Temperature (T)	Mean Character Ratio ↓
0.10	0.9541
0.25	0.9467
0.50 (Default)	0.7588
1.00	1.2258
1.50	45.6444

Table 13: Qualitative examples showing temperature-induced collapse.

T	Candidate Translation (Excerpt)
0.10	黑客版的星士在崩是因它在函表上求了一函在束的端。
0.25	黑客版的星Wars 在Crash 是因它在vtable的尾求函。
0.50	《黑客版的杰尼卡特之夜在崩溃中因为它在执行一个函数...》
1.00	「差了！我在《绝地当amentalion》中发现了一个大错误！」
1.50	*The hacked up version of Jedi Knight was crashing because it was calling a function off the end of a vtable. China Catholicstructor))*{}.....

Generation Length Analysis To investigate the empirical effects of decoding temperature on generation stability, we perform a case study using Llama-2-7b-chat on the WMT23 En→Zh translation task. We employ the character-count ratio (length of translation divided by length of source) as a proxy metric for model stability.

As shown in Table 12, the character ratio remains relatively stable at lower temperatures but under-



(a) Qwen2.5-Pre-7B (Lexical)

(b) Qwen2.5-Pre-7B (Semantic)

(c) Qwen2.5-Chat-7B (Lexical)

(d) Qwen2.5-Chat-7B (Semantic)

(e) Qwen3-Chat-8B (Lexical)

(f) Qwen3-Chat-8B (Semantic)

Figure 4: Temperature effect on the Qwen family. Comparing Pre-trained vs. Chat variants, Qwen models show consistent sensitivity to sampling temperature across both lexical (left) and semantic (right) metrics.

goes a dramatic "explosion" as T approaches 1.5. This signifies a total loss of structural alignment and linguistic coherence.

Qualitative Analysis of Generation Collapse We illustrate this phenomenon using a specific source sentence from the dataset: *"The hacked up version of Jedi Knight was crashing because it was calling a function off the end of a vtable."*

As shown in Table 13. At $T = 1.5$, the ND-MT system fails to maintain translation constraints. The output diverges into a mixture of garbled multilingual fragments and echoing English source tokens. This observation validates our characterization of modern MT systems as **temperature-constrained**: while they offer valuable lexical diversity at moderate settings, they lack the inherent robustness to remain semantic equivalence.

F The Buckets Effect in ND-MT

We introduce the "Buckets Effect" hypothesis to characterize ND-MT performance: just as a bucket's capacity is determined by its shortest plank, an ND-MT system's overall reliability is best approximated by its worst-case output. We validate this hypothesis by analyzing the correlation between different aggregation strategies (Min, Max, Mean, Random, Std) and the final system ranking across the five state-of-the-art ND-MT systems.

Worst-Case Performance Determines Ranking. Table 10 presents the correlation analysis across lexical metrics. The results strongly validate the Buckets Effect. For accuracy metrics (BLEU, GLVS, METEOR, ROUGE), the **Min** strategy (representing the lowest/worst score) consistently achieves near-perfect correlations ($\rho \approx 1.0$) with the true system ranking. In contrast, the **Max** strategy (best-case sample) often shows weaker correlations (e.g., $\rho = 0.70$ for BLEU), suggesting that a model's "lucky" best generations are poor predictors of its overall capability. This trend extends to semantic-based metrics (COMETDA, KIWI) shown in Table 11. The **Min** strategy maintains perfect correlations ($\tau = 1.00, \rho = 1.00$) across both sample sizes ($N = 20$ and $N = 50$), indicating that the lower bound of generation quality is a robust indicator of system ranking. Conversely, the **Max** strategy proves unstable, dropping as low as $\tau = 0.40$ and $\rho = 0.60$ for COMETDA at $N = 20$, though it improves with larger sampling sizes. Furthermore, the strong correlation of standard deviation (**Std**) ($\rho \geq 0.89$) reinforces that system consistency—specifically the ability to minimize variance and avoid quality collapse—is more indicative of model superiority than peak performance.

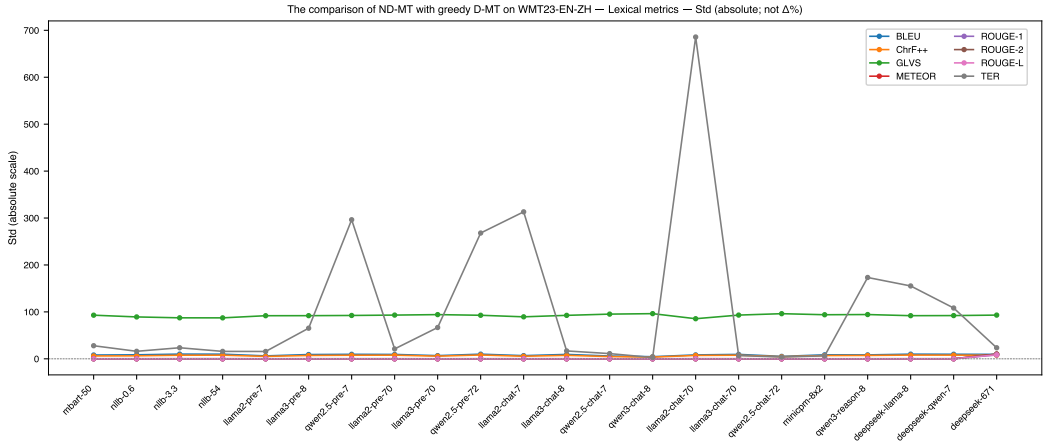
The Deceptive Nature of Best-Case TER. TER presents a unique case because it is an error metric (lower is better). Consequently, the **Min** strategy represents the *best-case* performance (lowest error), while the **Max** strategy represents the *worst-case* (highest error). As shown in Table 10 (Size 20), TER exhibits extreme divergence: its best-case performance (`Min`) loses predictive power ($\rho = 0.40$), whereas its worst-case performance (`Max`) remains highly predictive ($\rho = 0.90$). This confirms that the Buckets Effect holds universally: regardless of the metric's direction, the worst-case output is the true determinant of system quality.

G Other Supporting Figures

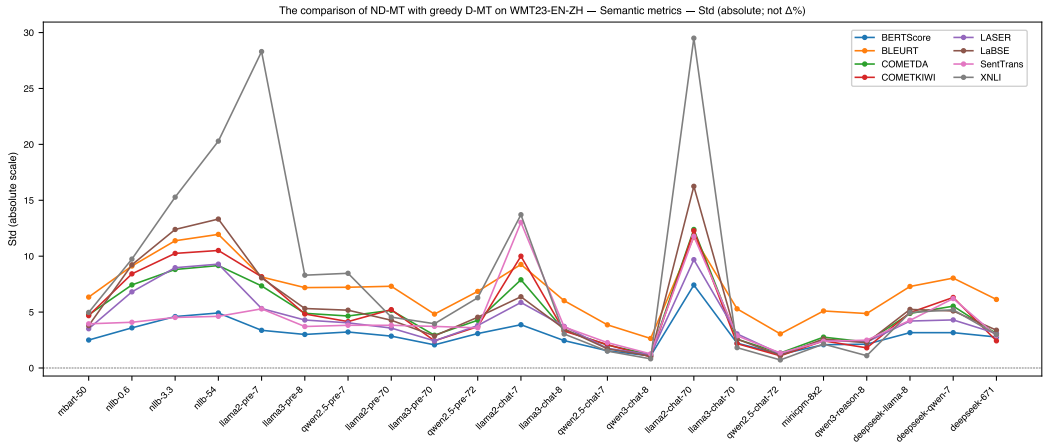
We report the other crucial supporting figures, including the ones to highlight the value of metrics on detecting the generation stability of ND-MT (See Figure 5); the ones to prove the potential of ND-MT on providing higher quality candidates than D-MT (See Figure 6); and the ones to show the generality of ND-MT on solving multimodality on diverse language directions (See Figure 7). For more details, please visit our github page.

H The Use of AI Assistant

The authors acknowledge the use of Claude Sonnet 4.5 and Gemini 3 solely for proofreading and polishing the language of this paper (e.g., improving grammar, clarity, and fluency). The writing process incorporated stylistic suggestions under the strict supervision of the authors. All technical ideas, methodology, experiments, analysis, and core content were conceived and produced entirely by the authors, without any AI-based content generation or fabrication.



(a) Std Delta Results of WMT23 En→Zh on Lexical Metrics.



(b) Std Delta Results of WMT23 En→Zh on Semantic Metrics.

Figure 5: Std Delta results for both lexical and semantic metrics on WMT23 En→Zh ($T = 0.5$, 10 candidates). Delta results are calculated relative to greedy decoding on identical data and models. Thresholds of -5 and -10 (dotted line) are included to indicate levels of significance.

Table 8: Rankings of D-MT Models across Lexical and Semantic Metrics. A rank of **1** indicates the best performance (Highest Score for all metrics, except Lowest Score for TER).

(a) Lexical Metric Rankings (Lower Rank # is Better)

Model	**BLEU**	**MET**	**R-1**	**R-2**	**R-L**	**chrF**	**TER**
NMT Baselines							
mBART-50	16	15	16	13	14	14	17
NLLB-600M	20	19	19	19	18	18	11
NLLB-3.3B	19	18	18	16	17	17	15
NLLB-54B	21	20	20	18	19	19	13
Pre-trained LLMs							
Llama2-7B	17	16	15	15	16	15	8
Llama3-8B	13	11	12	10	10	10	10
Qwen2.5-7B	5	5	4	4	4	5	**1**
Llama2-70B	9	7	7	7	7	8	7
Llama3-70B	4	4	3	3	3	4	4
Qwen2.5-72B	**1**	**1**	**1**	**1**	**1**	**1**	3
Chat Models							
L2-C-7B	23	23	22	23	22	23	20
L3-C-8B	14	12	13	11	11	11	12
Q2.5-C-7B	11	8	8	9	8	9	9
Q3-C-8B	8	6	6	6	6	6	5
L2-C-70B	22	17	23	22	23	22	21
L3-C-70B	7	9	5	5	5	3	2
Q2.5-C-72B	3	3	2	2	2	2	6
Reasoning							
MiniCPM	6	9	10	8	9	7	14
Q3-R-8B	10	10	11	12	12	13	19
DS-Llama-8B	15	13	14	14	13	12	18
DS-Qwen-7B	18	14	17	17	15	16	16
DS-671B	20	15	21	21	20	20	22

(b) Semantic Metric Rankings (Lower Rank # is Better)

Model	**KIWI**	**BLE**	**BERT**	**CMT**	**LSR**	**LBS**	**SNT**	**XNLI**
NMT Baselines								
mBART-50	15	13	14	15	4	12	18	8
NLLB-600M	19	19	19	19	18	19	19	11
NLLB-3.3B	20	20	20	18	20	22	20	15
NLLB-54B	21	21	21	20	21	23	21	18
Pre-trained LLMs								
Llama2-7B	17	18	18	17	17	17	12	17
Llama3-8B	14	11	12	12	9	11	8	10
Qwen2.5-7B	7	6	5	5	6	6	6	5
Llama2-70B	13	12	10	11	5	7	11	9
Llama3-70B	10	7	4	6	2	5	7	7
Qwen2.5-72B	**1**	**1**	**1**	**1**	3	2	7	**1**
Chat Models								
L2-C-7B	22	22	22	22	18	20	2	20
L3-C-8B	12	14	13	10	16	13	5	6
Q2.5-C-7B	8	10	8	7	8	8	6	4
Q3-C-8B	2	8	6	4	5	4	10	3
L2-C-70B	23	23	23	23	22	21	**1**	21
L3-C-70B	6	9	5	5	10	9	9	5
Q2.5-C-72B	3	3	2	2	**1**	**1**	9	4
Reasoning								
MiniCPM	9	8	7	8	7	9	16	7
Q3-R-8B	8	9	11	9	11	10	3	13
DS-Llama-8B	13	15	15	14	14	15	15	14
DS-Qwen-7B	16	16	16	16	12	14	4	9
DS-671B	11	17	21	15	20	18	14	19

Table 10: Correlation Analysis of Lexical Metrics across Sampling Sizes (20, 50) for Five SOTA ND-MT Systems. The "Worst Case" strategy consistently predicts system ranking. Note that for accuracy metrics (BLEU, etc.), **Min** is the worst case. For the error metric **TER**, **Max** is the worst case.

Size	Strategy	**BLEU** (↑)		**GLVS** (↑)		**METEOR** (↑)		**ROUGE-1** (↑)	
		ρ / τ	p-val	ρ / τ	p-val	ρ / τ	p-val	ρ / τ	p-val
20	Max (Best)	.70 / .60	.19	1.0 / 1.0	.00	1.0 / 1.0	.00	1.0 / 1.0	.00
	Mean	.90 / .80	.04	.90 / .80	.04	1.0 / 1.0	.00	1.0 / 1.0	.00
	Min (Worst)	**1.0 / 1.0**	.00	**1.0 / 1.0**	.00	**1.0 / 1.0**	.00	**1.0 / 1.0**	.00
	Random	.90 / .80	.04	.90 / .80	.04	1.0 / 1.0	.00	1.0 / 1.0	.00
	Std	.70 / .60	.19	.90 / .80	.04	1.0 / 1.0	.00	.82 / .74	.09
50	Max (Best)	.70 / .60	.19	.90 / .80	.04	.90 / .80	.04	1.0 / 1.0	.00
	Mean	.90 / .80	.04	.90 / .80	.04	1.0 / 1.0	.00	1.0 / 1.0	.00
	Min (Worst)	**1.0 / 1.0**	.00	**1.0 / 1.0**	.00	**1.0 / 1.0**	.00	**1.0 / 1.0**	.00
	Random	.90 / .80	.04	.90 / .80	.04	1.0 / 1.0	.00	1.0 / 1.0	.00
	Std	.70 / .60	.19	1.0 / 1.0	.00	.97 / .95	.00	.82 / .74	.09

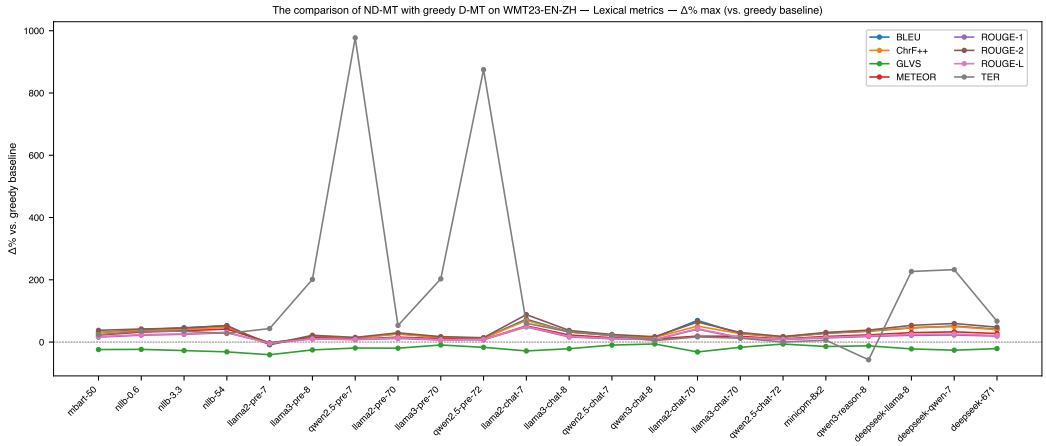
Size	Strategy	**ROUGE-2** (↑)		**ROUGE-L** (↑)		**TER** (↓)		**ChrF++** (↑)	
		ρ / τ	p-val	ρ / τ	p-val	ρ / τ	p-val	ρ / τ	p-val
20	Max	.90 / .80	.04	.90 / .80	.04	**.90 / .80 (Worst)**	.04	.90 / .80	.04
	Mean	1.0 / 1.0	.00	1.0 / 1.0	.00	.90 / .80	.04	1.0 / 1.0	.00
	Min	1.0 / 1.0	.00	1.0 / 1.0	.00	*.40 / .40 (Best)*	.50	1.0 / 1.0	.00
	Random	1.0 / 1.0	.00	1.0 / 1.0	.00	.90 / .80	.04	1.0 / 1.0	.00
	Std	.92 / .88	.03	.82 / .74	.09	.90 / .80	.04	1.0 / 1.0	.00
50	Max	.80 / .60	.10	1.0 / 1.0	.00	.90 / .80	.04	.90 / .80	.04
	Mean	1.0 / 1.0	.00	1.0 / 1.0	.00	.90 / .80	.04	1.0 / 1.0	.00
	Min	1.0 / 1.0	.00	1.0 / 1.0	.00	.90 / .80	.04	1.0 / 1.0	.00
	Random	1.0 / 1.0	.00	1.0 / 1.0	.00	.80 / .60	.10	1.0 / 1.0	.00
	Std	.92 / .89	.03	.76 / .67	.13	.90 / .80	.04	.90 / .80	.04

ρ = Spearman; τ = Kendall. Values are coefficient / p-value. For TER, Max is Worst Case, Min is Best Case.

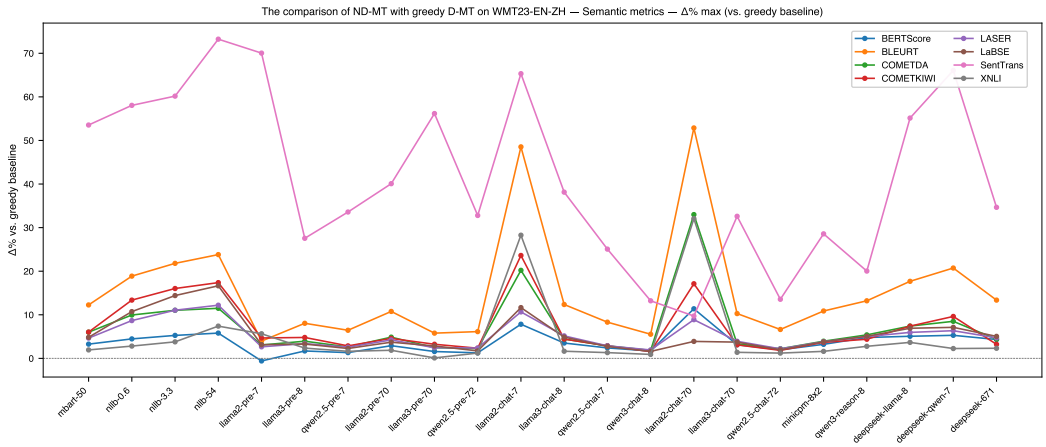
Table 11: Correlation Results of Semantic-based Metrics on WMT23 EN-ZH Comparing Sample Sizes $N = 20$ and $N = 50$.

Strategy	COMETDA		KIWI	
	$N = 20$	$N = 50$	$N = 20$	$N = 50$
Kendall's τ / p-value				
Min (Worst)	**1.0**/.02	**1.0**/.02	**1.0**/.02	**1.0**/.02
Max (Best)	.40/.48	.80/.08	1.00/.02	.80/.08
Mean	1.00/.02	1.00/.02	1.00/.02	1.00/.02
Random	1.00/.02	1.00/.02	1.00/.02	1.00/.02
Std	.84/.05	.95/.02	1.00/.02	.80/.08
Spearman's ρ / p-value				
Min (Worst)	**1.0**/.00	**1.0**/.00	**1.0**/.00	**1.0**/.00
Max (Best)	.60/.28	.90/.04	1.00/.00	.90/.04
Mean	1.00/.00	1.00/.00	1.00/.00	1.00/.00
Random	1.00/.00	1.00/.00	1.00/.00	1.00/.00
Std	.89/.04	.97/.00	1.00/.00	.90/.04

KIWI=COMETKIWI. N denotes sample size.

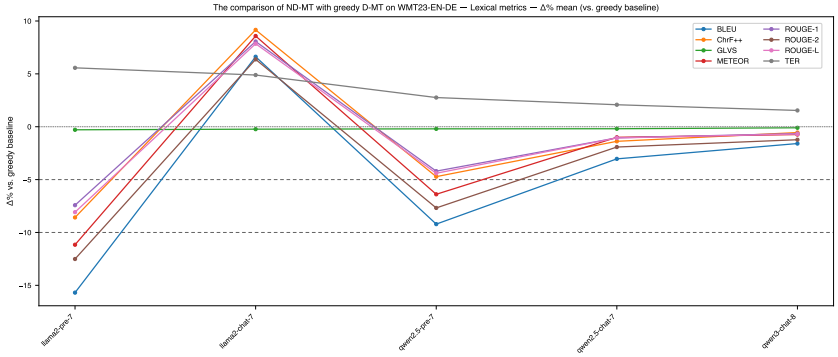


(a) Max delta on lexical metrics (WMT23 En→Zh).

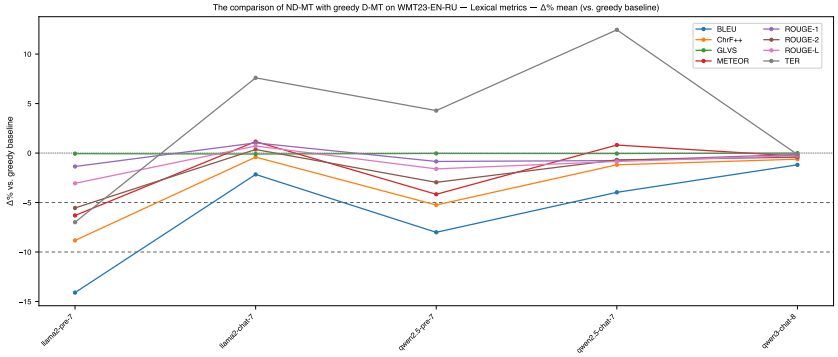


(b) Max delta on semantic metrics (WMT23 En→Zh).

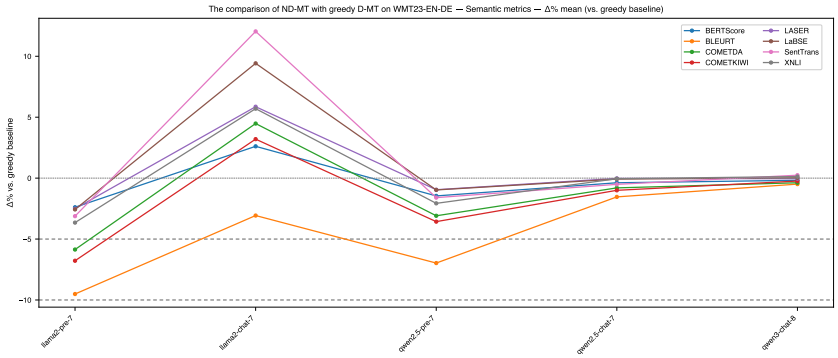
Figure 6: Max delta values on WMT23 En→Zh for lexical and semantic metrics (T=0.5, 10 candidates). Deltas are computed relative to greedy decoding on identical data and models.



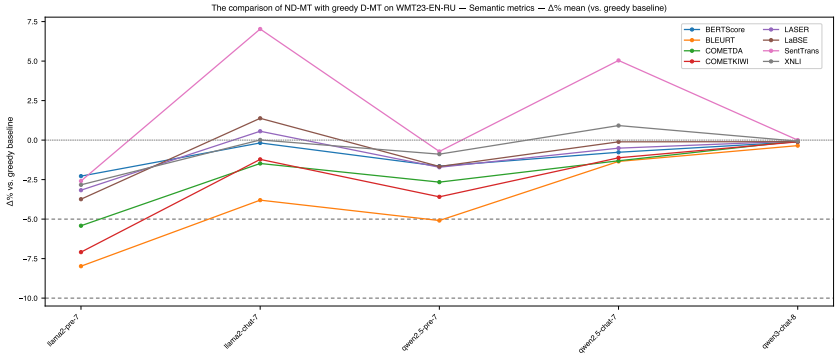
(a) Lexical metrics—delta mean (WMT23 En→De).



(b) Lexical metrics—delta mean (WMT23 En→Ru).



(c) Semantic metrics—delta mean (WMT23 En→De).



(d) Semantic metrics—delta std (WMT23 En→Ru).

Figure 7: Delta statistics on WMT23 En→De and En→Ru for lexical and semantic metrics (T=0.5, 10 candidates). Deltas are computed relative to greedy decoding on identical data and models.