Contrastive Token Learning with Similarity Decay for Repetition Suppression in Machine Translation

Huangyu Dai[1,*] **Ben Chen**[1,*,†] **Kaidi Chen**[1]
Ying Han[2] **Zihan Liang**[1] **Wen Jiang**[1]
[1]Alibaba Group, Hangzhou, China
[2]Zhejiang Gongshang University, School of Foreign Languages, Zhejiang, China

Abstract

For crosslingual conversation and trade, Neural Machine Translation (NMT) is pivotal yet faces persistent challenges with monotony and repetition in generated content. Traditional solutions that rely on penalizing text redundancy or token reoccurrence have shown limited efficacy, particularly for lengthy article and e-commerce descriptions with inherent redundancy, even with the advent of Large Language Models (LLMs). This paper investigates the underlying causes of textual repetition through the lens of information entropy, attributing the phenomenon to the elevated uncertainty within the input text. To address this, a novel algorithm named Contrastive Token Learning with Similarity Decay (CTSD) is introduced, which modulates the suppression of tokens dynamically, informed by varying attention weights and inter-token distances. Furthermore, an e-commerce dataset comprised of title texts of real online items is compiled and released and is susceptible to hallucination translations to benchmark the algorithm. Extensive evaluations demonstrate that CTSD significantly outperforms existing approaches in precision and generalizability. Additional online A/B testing underscores its practical value, showing marked user engagement and conversion improvements. Notably, this method has been implemented with full traffic on six multilingual sites of alibaba.com, the largest B2B e-commerce platform in the world.

1 Introduction

In recent years, the synergy of neural networks coupled with the increasing scale of parallel corpora has significantly propelled Neural Machine Translation (NMT) forward (Liu et al., 2020; Costa-jussà et al., 2022). Notably, the sophisticated reasoning abilities and specialized knowledge acquired by Large Language Models (LLMs) (Touvron et al., 2023; Bai et al., 2023) further contribute modern NMT systems towards achieving near-human-level performance (Lin et al., 2022; Zhu et al., 2023). However, the reliability of NMT in delivering accurate and coherent translations remains unstable, and unexpected errors such as omissions or nonsensical outputs are often encountered. This challenge persists across the spectrum, especially for complex textual materials like repetition-prone articles and e-commerce descriptions.

Typical NMT problems, commonly referred to as "hallucinations", can be categorized into two main types (Dale et al., 2022; Guerreiro et al., 2023b). The first involves repeating words or sentences, known as "oscillations", while the second pertains to generating content not supported by the source, termed "largely fluent". Of these two types, "oscillations" are particularly intolerable for leading to repetition with low coherence and accuracy, making NMT limited for multiple applications (Ji et al., 2023; Guerreiro et al., 2023a). Consequently, addressing oscillation (repetition generation) has emerged as a primary focus in current research, vital for improving reliability and usability in complex scenarios.

Previous methods mainly employed two strategies to suppress repetition generation. The first is the direct strategy interventions during the inference stage, such as n-gram not repeat, Contrastive Search (CS) (Su et al., 2022), and Penalized Sampling (PS) (Keskar et al., 2019). These techniques focus on preventing repeated tokens to eliminate oscillations. However, they would disrupt the token distribution of output, leading to other errors. Consequently, recent methods focus on designing training objectives during the model training stage to better address hallucination problems (Welleck et al., 2020; Su et al., 2022; Jiang et al., 2022). Yet, these training objectives do not adequately explore the root of oscillations in transformer-based models, often using a direct intervention way to sup-

[*]Equal Contribution.
[†]Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3247–3261

Model	Type	Sentence
NLLB-1.3B	src_t	Baseball cap Manufacturer Custom plain Baseball hat Embroidered baseball cap for men
	tran_t	Baseballkappe Hersteller, Baseballhut, Stück für Stück, Baseballhut, Stück für Stück, Stück für Stück, ...
	opt_t	Baseballmütze Hersteller individuelle schlichte Baseballmütze bestickte Baseballmütze für Herren
mBART-large	src_t	1.8 Ton Mini Excavator Crawler Excavator Mini Bagger Cheap Price With Ce For Sale Epa Ce Mini Excavator
	tran_t	1,8 Tonnen Mini Bagger Bagger Bagger Bagger Bagger Bagger Bagger ...
	opt_t	1,8 Tonnen Mini Bagger Mini Bagger Preis mit Ce Zum Verkauf Epa Ce Mini Bagger
LLaMA2-7B	src_t	4 in 1 modern rotating multi game billiard pool table 7ft with air hockey 4 in 1 pool table 4 in 1 table game
	tran_t	4-in-1 moderne rotateürende multi-Spiel-Billard-Pool-Tisch 4-in-1-Tisch-Spiel, 4-in-1-Tisch-Spiel...
	opt_t	4-in-1 moderner drehbarer multi-spiel-billardtisch 7-fuss mit air-hockey 4-in-1-pooltisch 4-in-1-tischspiel
Qwen-7B	src_t	Excavator Machine electric Hydraulic Mini Small Micro Crawler Bagger Digger Mini Excavators
	tran_t	Abbaumaschine Elektrohydraulische Kleinmodell Mikro-Krabbenwerfer Mini-Bagger 迷你 挖掘机挖掘机 ...
	opt_t	Bagger Abbauger Elektro-hydraulik klein Mikro-Crawler-abbaugern minibaggerminibagger

Table 1: Examples of repetition generation in NMT. Src_t, tran_t, and opt_t are the abbreviations of source texts, translated texts with the original model, and optimized texts with the additional CTSD method.

press tokens that have appeared. Although they can effectively suppress word or sentence repetition, they also lead to a lower coherence and accuracy in translation (Post, 2018; Wan et al., 2022).

In this paper, we conduct an in-depth exploration of the fundamental reasons underlying textual repetition in machine translation, primarily utilizing the concept of information entropy. Our research reveals that this phenomenon largely stems from increased levels of uncertainty present within the input text. Repetitive token generation occurs because the information from previously generated tokens does not provide additional value (information entropy). To effectively deal with this issue, we propose an innovative algorithm called "Contrastive Token Learning with Similarity Decay" (CTSD). This innovative approach aims to dynamically adjust the suppression of tokens by analyzing the attention differences between different output tokens' embeddings and the distance between the inner tokens, thereby enhancing the accuracy and stability of the output. Meanwhile, our method can be applied to both specialized translation models and LLM without additional data preparation. The results show that our method can effectively improve the performance of models in translation tasks and prevent oscillation hallucinations. Compared with Contrastive Token Learning (CT), CTSD achieves improvements by 1% to 10% in translation quality on both the FLORES-200 and our proprietary e-commerce datasets.

In addition, a comprehensive evaluation by experts reveals that CTSD exceeds current methods in both precision and generalizability. Online A/B tests further highlight its practicality, as evidenced by substantial gains in user engagement with higher click-through and conversion rates and final gross merchandise volume. Importantly, this method has been successfully deployed across eight multilingual websites with full traffic of alibaba.com, the world's largest B2B e-commerce platform.

2 Related Work

2.1 Multilingual Neural Machine Translation

Multilingual NMT has advanced significantly from its early focus on two-language systems. The pioneering work of Dong et al. (2015) expanded NMT into a one-to-many framework by sharing encoders across four language pairs. This development sparked a surge in research on NMT systems capable of handling multiple languages (Johnson et al., 2017; Chu and Dabre, 2019; Yang et al., 2021). At first, the research was mainly focused on improving multilingual NMT's capabilities on rich-resource languages through specific components and more diverse training data (Escolano et al., 2020; Fan et al., 2021; Liu et al., 2020). Now, more research has turned to low-resource languages. Tars et al. (2021) improved the capabilities of low-resource languages by simultaneously training different languages of the same language family. Pan et al. (2021) enhanced the translation quality of non-English language directions through data augmentation and contrastive learning. NLLB Team developed a conditional compute model based on a Sparsely Gated Mixture of Experts (MoE) to improve low-resource language translation quality (Costa-jussà et al., 2022).

With the increasing scale of parameters and training corpus, LLMs (GPT-3, BLOOM, and LLaMA included in open source models (Brown et al., 2020; Workshop et al., 2022; Touvron et al., 2023), ChatGPT, GPT-4 and Claude included in closed source models (OpenAI, 2022; Achiam et al., 2023; An-

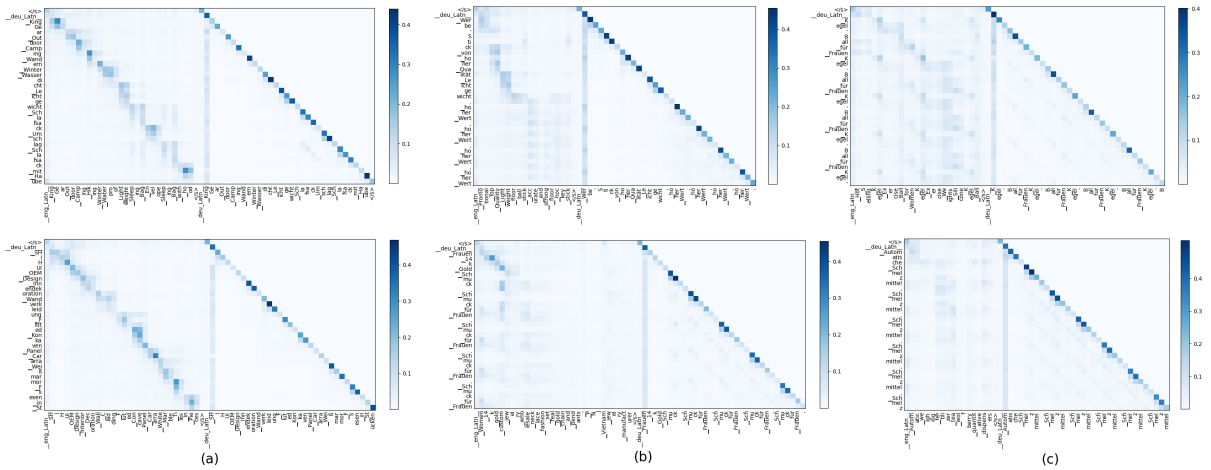

Figure 1: ALTI+ results for En-De translation examples. (a) normal result. (b) middle appearing repetition result, and (c) total repetition result. The contribution values of all tokens in each row have been normalized.

thropic, 2022)) have gained unexpected complex reasoning and emergent abilities in the face of unseen tasks, enabling it to handle various tasks, like text summarization, QA system, and free dialogue (Wei et al., 2022). However, the professional evaluation found that many large models still cannot surpass state-of-the-art translation engines like NLLB and Google Translate in professional translation (Zhu et al., 2023). The recently emerged translation-specialized LLMs attempt to conduct more specialized data training to reduce this gap (Xu et al., 2023; Chen et al., 2024). However, they still inevitably generate repetition when faced with a complex text translation.

2.2 Repetition Suppression

With the advent of LLMs, repetition suppression methods have received significant attention. Currently, there are two mainstream types of methods: decoding methods and training methods. Decoding methods initially gained popularity because no further tuning is needed. Commonly used methods include PS and CS. Keskar et al. (2019) implemented PS, using a temperature coefficient to reduce the likelihood of historical tokens, reducing the probability of producing oscillatory hallucinations. Su et al. (2022) proposed CS to suppress historically generated tokens by computing the cosine similarity between the embedding of historical tokens and the current token.

While decoding methods successfully suppress oscillation hallucination, they face challenges of reduced generation quality and increased inference cost. Therefore, research is shifting towards designing training objectives for more accurate and stable translations. Welleck et al. (2020) proposed unlikelihood training (UL), suppressing repetition through unlikelihood loss. Su et al. (2022) adopted contrastive training, emphasizing distinctions between different tokens to prevent monotonous repetition. Jiang et al. (2022) introduced the CT loss, which selectively suppresses tokens on a negative token list without impacting irrelevant tokens. CT has been theoretically proven advantageous over traditional cross-entropy (CE) and UL loss, emerging as the most effective algorithm of oscillations suppression to date (Sun et al., 2023; Guan et al., 2023).

With the emergency of LLMs, reinforcement learning (RL) methods such as Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) have surfaced as new strategies for reducing hallucinations (Schulman et al., 2017; Rafailov et al., 2024). By training LLMs with preference data, they can instruct outputs that are close to human expectations, effectively lowering the chance of hallucinations.

3 Methodology

3.1 Hallucination Analysis

A well-known theoretical analysis of the repetition problem in text generation simplifies predicting the next word into a first-order Markov chain (Fu et al., 2021). It assumes that the currently generated token is only affected by the same token generated at the previous moment. Under this assumption, the entire generation sequence forms a directed cycle when the model generates a word that has already been generated. For the sentence "I like

it and guess he knows I like it because ...", this theory insists that the second generation of "I like it" is mainly affected by the former so that the next predicted token is most likely "and". Methods like CT and PS employ this idea to prevent directed cycles and reduce repetitive text generation.

However, this theory ignores the input text and previously translated text, contrary to the model based on the cross-attention mechanism. Consider the title of an item: "Best Selling New Arrival Outdoor Shapewear Dress Women's Dresses Built-in Shapewear Maxi Dress". Global-level suppression of repetitive generation can lead to the replacement of "shapewear" and "dress" in the latter part of the title translation with other words, thus deviating from the original meaning. To more accurately analyze the impact of each former token on the next predicted word, we perform visual analysis through the ALTI+ method (Ferrando et al., 2022). ALTI+ calculates the contribution of each previous token to the generation of the current token by computing the Manhattan distance between the previous token and the newly generated token representation in each layer. For a comprehensive comparison, we show En-De translation with three columns: (a) normal result, (b) middle appearing repetition result, and (c) total repetition result, respectively, in Figure 1.

We can distinctly observe that: 1) The generation of each token is primarily influenced by the input text and the nearest neighboring tokens, with tokens at relatively farther positions exerting minimal impact; 2) Identical repeated words are affected by tokens in the same position, and the longer the generated text, the weaker the influence of the corresponding tokens in the input text. This explains why global suppression of repeated words, although effective in suppressing repetition, leads to poorer translation outcomes, particularly in decoder-only LLMs, where a forcibly replaced repetitive token results in subsequent tokens deviating increasingly from the original meaning.

Here, we attempt to elucidate the underlying causes of text repetition generation from the perspective of information entropy. In the transformer mechanism, the generation of each token is influenced by all preceding tokens. Repetitive token generation occurs because the information from previously generated tokens does not provide additional value. For instance, when predicting the (n+1)-th token, the information from tokens 0 to (n-1) is identical to that from tokens 0 to n, causing the model to generate the n-th word repetitively.

To substantiate this, we have calculated the embeddings of each token and visualized it using T-SNE in Figure 2. We can observe that the 2D vectors of repeated tokens are clustered, and two identical tokens generated continuously are closely together. The average cosine similarity of two adjacent "her" tokens (embedding of the last layer) in the first picture is 0.85, which is much higher than two adjacent different tokens. In the second picture, the cosine similarity of two adjoining "mittel" tokens is 0.94, but the similarity between two tokens appearing before and after is only 0.33. This shows that when repeated tokens are generated, the text's information entropy is almost unchanged. Therefore, to suppress repetitive generation while maintaining accuracy and stability, we should selectively focus on each former token and adaptively attend to the changes in information entropy with each token generation.

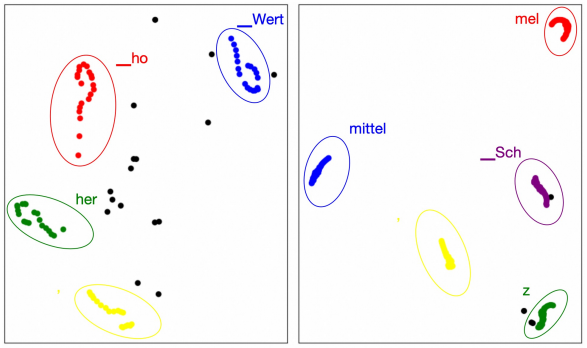

Figure 2: T-SNE results of different generated tokens. (a) middle appearing repetition result and (b) total repetition result.

3.2 Learning-Based Solution

Models trained with traditional CE loss are prone to hallucinations when facing out-of-domain data. To address the issue in translation models, several new training objectives (loss functions) have been designed to suppress negative tokens more effectively. Welleck et al. (2020) proposed unlikelihood training specifically designed to penalize the likelihood of negative tokens. The token-level unlikelihood training objective (UL-T) at time step t is defined as:

$$\mathcal{L}_{UL}^{t} = - \sum_{y_t^- \in C^t} \log \left(1 - p\left(y_t^- \mid \mathbf{y}_{<t}, \mathbf{x}\right)\right) \quad (1)$$

where $\mathbf{x}$ is the source text, $\mathbf{y}_{<t}$ is the translation text generated before time step t, C^t is the set of

Dataset	Model	Method	SacreBLEU↑	Rouge-L↑	COMET↑	rep-2↓	rep-3↓	rep-w↓	rep-r↓	div↑
FLORES-200	Ground Truth	-	-	-	-	0.43	0.1	0.03	0.01	1.00
	NLLB-1.3B	CE	31.77	0.558	0.840	0.69	0.22	0.04	0.02	0.99
		CT	30.51	0.547	0.841	0.33	**0.10**	0.02	0.01	1.00
		CTSD	**32.14**	**0.561**	**0.843**	**0.53**	0.13	**0.03**	**0.01**	**1.00**
	mBART-large	CE	27.87	0.499	0.818	0.56	0.12	0.04	0.01	0.99
		CT	27.27	0.499	0.816	0.58	0.21	0.03	0.01	0.99
		CTSD	**28.04**	**0.508**	**0.820**	**0.56**	**0.12**	**0.03**	**0.01**	**0.99**
	LLaMA2-7B	CE	19.65	0.439	0.826	4.29	3.84	0.05	0.02	0.88
		CT	19.42	0.439	0.827	0.90	**0.24**	0.04	0.02	0.99
		CTSD	**19.94**	**0.443**	**0.827**	**0.86**	0.25	**0.04**	**0.02**	**0.99**
	Qwen-7B	CE	18.99	0.426	0.776	0.26	0.12	0.02	0.01	1.00
		CT	19.53	0.435	0.780	0.28	0.07	0.02	0.01	1.00
		CTSD	**19.72**	**0.441**	**0.784**	**0.31**	**0.09**	**0.02**	**0.01**	**1.00**
	GPT-3.5-Turbo	-	3.86	0.065	0.735	0.46	0.12	0.03	0.01	0.99
	GPT-4-Turbo	-	3.81	0.061	0.739	0.40	0.09	0.03	0.01	0.99
E-Commerce	NLLB-1.3B	CE	6.71	0.178	0.575	36.17	37.21	0.13	0.10	0.24
		CT	7.16	0.182	0.600	0.82	0.21	**0.05**	0.03	0.99
		CTSD	**7.59**	**0.192**	**0.602**	**0.75**	**0.19**	**0.05**	**0.02**	**0.99**
	mBART-large	CE	16.99	0.357	0.658	23.68	17.08	0.35	0.40	0.54
		CT	17.23	0.380	0.687	18.95	13.18	0.29	0.31	0.63
		CTSD	**17.67**	**0.391**	**0.694**	**12.66**	**6.13**	**0.29**	**0.31**	**0.79**
	LLaMA2-7B	CE	19.06	0.436	0.747	**0.59**	**0.04**	0.06	0.02	0.99
		CT	20.72	0.455	0.753	0.82	0.14	0.06	0.02	0.99
		CTSD	**21.11**	**0.460**	**0.757**	0.80	0.12	**0.06**	**0.02**	**0.99**
	Qwen-7B	CE	24.14	0.457	0.734	0.73	0.12	0.05	0.02	0.99
		CT	24.01	0.457	0.730	0.73	0.22	0.05	0.02	0.99
		CTSD	**24.58**	**0.462**	**0.741**	**0.73**	**0.12**	**0.05**	**0.02**	**0.99**
	GPT-3.5-Turbo	-	5.28	0.112	0.579	0.80	0.20	0.03	0.02	0.99
	GPT-4-Turbo	-	3.84	0.081	0.582	0.57	0.08	0.02	0.01	0.99

Table 2: Translation quality and repetition rate (rep-2 and rep-3 are percentages) of NLLB-1.3B, mBART-large, LLaMA2-7B, and Qwen-7B models under different training methods and different datasets. (The model with the repetition rate closest to the ground truth on the FLORES-200 dataset is considered to have the best performance)

previous negative tokens at time step t. This approach focuses on decreasing the probability of generating already-produced tokens, aiming to break the directed cycle observed in NMT models.

Additionally, contrastive learning loss $\mathcal{L}_{CL}$ has been proposed as an effective training objective (Su et al., 2022), which encourages models to learn isotropic token representations through a similarity penalty. The $\mathcal{L}_{CL}$ is defined as:

$$\mathcal{L}_{CL}^{t} = \frac{1}{t-1}\sum_{i=1}^{t-1}\max\{0, \rho - s\left(h_{y_i}, h_{y_i}\right) + s\left(h_{y_i}, h_{y_{t-i}}\right)\} \tag{2}$$

where $\rho \in [-1, 1]$ is a pre-defined margin, h_{y_i} is the embedding of token y_i, $s\left(h_{y_i}, h_{y_{t-i}}\right) = (h_{y_i}^{\top} h_{y_j}) \setminus (\|h_{y_i}\| \cdot \|h_{y_{t-i}}\|)$ is cosine similarity. This loss function is designed to increase the distance between representations of distinct tokens, thereby creating a more discriminative and diverse model representation space.

Recently, CT loss has presented (Jiang et al., 2022) and the formulation for time step t is defined as:

$$\mathcal{L}_{CT}^{t} = \log\left(1 + \sum_{y_t^- \in S_N^t} \exp\left(h_t^T W_{y_t^-} - h_t^T W_{y_t}\right)\right) \tag{3}$$

where h_t is the hidden state, y_t means the positive token at time step t. W_{y_t} denotes the embedding for token y_t, S_N^t is the set of the previous N tokens.

The research shows that CT loss is the optimal loss function in effectively suppressing oscillations hallucination, as it suppresses negative tokens while enhancing positive tokens. Despite its effectiveness, its somewhat rough selection of negative tokens sometimes leads to suboptimal results. Therefore, based on the previously analyzed repeatability principle of ALTI+ and T-SNE, we propose the CTSD loss, an optimization of the original CT loss, significantly improving the accuracy,

Model	Dataset	Method	SacreBLEU↑	rep-2↓
Qwen-1.8B	FLORES-200	CE	5.79	1.88
		CTSD	**5.88**	**0.19**
	E-Commerce	CE	18.83	1.69
		CTSD	**19.00**	**0.71**
Qwen-14B	FLORES-200	CE	21.47	0.32
		CTSD	**21.80**	**0.22**
	E-Commerce	CE	26.42	1.42
		CTSD	**26.46**	**0.72**

Table 3: Translation quality and repetition rate of Qwen-1.8B and Qwen-14B models under different training methods and different datasets.

stability, and overall effectiveness of the model output. This method dynamically suppresses previously generated tokens by designing two attenuation factors. The first attenuation factor uses cosine similarity to measure the similarity in the context where the hallucination token's attention is very similar to the previous token. Additionally, an exponential-decay attenuation factor is designed to weaken the suppression of distant tokens, considering that the contribution of generated tokens is inversely related to the distance between tokens. Through this approach, the model can more accurately handle generation tasks, further enhancing its performance.

Finally, CTSD loss for time step t is defined as:

$$\mathcal{L}^t_{CTSD} = \log\left(1 + \sum_{y_t^- \in S_N^t} \alpha_d \alpha_s \exp\left(h_t^T W_{y_t^-} - h_t^T W_{y_t}\right)\right) \tag{4}$$

where $\alpha_d = e^{\frac{t_- - t}{T}}$, t_- represents the time when y_t^- is generated, T is the temperature coefficient that controls decay. $\alpha_s = \frac{atten_{t_-}^{\mathrm{T}} atten_t}{\|atten_{t_-}\| \|atten_t\|}$, $atten_{t_-}$ represents attention distribution between y_t^- and encoder embedding.

To more intuitively demonstrate the role of the attenuation factor, we display the weight matrix for a typical translated sentence (constructed from the attenuation factor of each generated token) in Figure 3. Figure 3(a) shows the cosine similarity between the contributions of the input tokens. Figure 3(b) illustrates the exponential decay matrix. It can be observed that under normal translation conditions, the overall attention similarity between tokens is generally low. However, some unrelated tokens that are far apart exhibit high similarity. Therefore, an additional exponential-decay attenuation factor is necessary to suppress these extraneous similarities further.

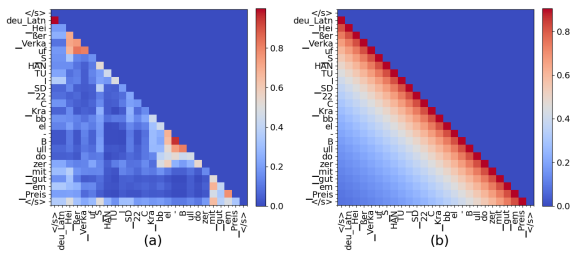

Figure 3: Attenuate factor of different generated tokens. (a) attention similarity and (b) exponential decay matrix.

4 Experiments

4.1 Experiments Setup

The experiments aim to evaluate whether CTSD can suppress hallucinations in specialized translation models and LLMs while maintaining stability. We integrate several baseline methods, including traditional CE loss, decoding-based methods like CS and PS, and training-based methods such as UL at Token-level (UL-T), CL, and CT.

For specialized models like NLLB-1.3B and mBART-large, extensive comparative analysis and experimentation suggest that a batch size of 64 and a fixed learning rate of 5×10^{-5} provide an effective balance between stability and performance. For decoder-only LLMs such as LLaMA2-7B and Qwen-7B, the LoRA method is employed with parameters $r = 8$ and $\alpha = 16$. These models utilize a batch size of 32 and a learning rate of 2×10^{-5} to enhance translation capabilities consistently. In our experiments, the CTSD method uses $T = 10$, $N = 5$ for specialized models and $T = 5$, $N = 10$ for LLMs. Further details on the hyperparameter ablation experiments can be found in the appendix.

Our datasets comprise the open-source general dataset WMT16, along with some e-commerce translation data for training and the FLORES-200 devtest dataset for evaluation. Furthermore, a novel evaluation dataset comprising e-commerce texts susceptible to hallucination translations is compiled and released to benchmark our algorithm. This dataset is an English-German Parallel Corpus encompassing 3,500 authentic titles from alibaba.com. Each text segment has undergone meticulous translation and verification by human experts.

Evaluations of NMT performance include many metrics. Accuracy-related measures include SacreBLEU and Rouge-L, which evaluate the precision of lexical choices, and COMET, which assesses semantic similarity. Repetition-related metrics include rep-2, rep-3, div, rep-w, rep-r, and uniq-1.

Model	Method	SacreBLEU↑	Rouge-L↑	COMET↑	rep-2↓	rep-3↓	rep-w↓	rep-r↓	div↑	uniq-1↑
NLLB-1.3B	-	0.71	0.098	0.280	93.29	92.65	0.91	0.95	0.00	7355
	PS	6.47	0.173	0.571	3.69	3.54	0.03	0.02	0.89	13955
	CS	5.03	0.139	0.482	**0.18**	**0.04**	**0.01**	**0.00**	**1.00**	**18378**
	CE	6.71	0.178	0.575	36.17	37.21	0.13	0.10	0.24	13060
	UL-T	7.01	0.183	0.578	34.91	35.95	0.13	0.09	0.26	12547
	CL	6.84	0.181	0.578	26.51	27.23	0.11	0.08	0.38	12881
	CT	7.16	0.182	0.600	0.82	0.21	0.05	0.03	0.99	13670
	CTSD	**7.59**	**0.192**	**0.602**	0.75	0.21	0.05	0.02	0.99	12658
Qwen-7B	-	4.14	0.133	0.599	5.66	5.22	0.04	0.03	0.85	14072
	PS	3.29	0.113	0.593	0.97	0.39	0.03	0.02	0.98	**14265**
	CS	4.09	0.135	0.599	2.84	2.19	0.04	0.03	0.93	13973
	CE	24.14	0.457	0.734	0.73	0.12	0.05	0.02	0.99	11303
	UL-T	24.13	0.459	0.736	0.97	0.43	0.05	0.02	0.98	11252
	CL	24.55	0.460	0.740	0.74	0.12	0.05	0.02	0.99	11103
	CT	24.01	0.457	0.730	0.73	0.22	0.05	0.02	0.99	11227
	CTSD	**24.58**	**0.462**	**0.741**	**0.73**	**0.12**	**0.05**	**0.02**	**0.99**	11130

Table 4: Translation quality and repetition rate of NLLB-1.3B and Qwen-7B under different repetition suppression methods during training or inference stage on the e-commerce hallucination dataset.

Specifically, rep-2, 3, w, and r focus on lexical repetition, whereas div and uniq-1 emphasize lexical diversity.

For detailed information on the construction of the hallucination dataset and the definitions and calculation methods of NMT metrics, please refer to the Appendix.

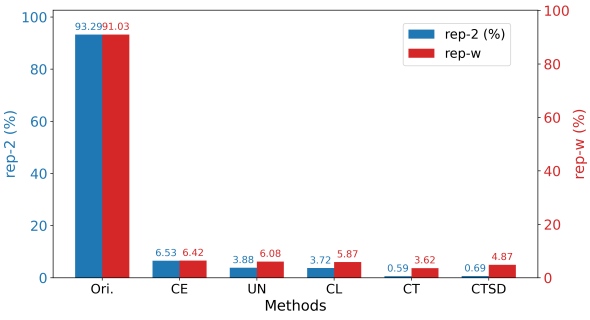

Figure 4: Top 1% repeatability metrics among 1 million items, titles from online e-commerce websites.

4.2 Evaluation Results

As shown in Table 2, CTSD consistently improves translation quality across all models while maintaining meager repetition rates.

For specialized translation models, the CT loss underperforms CE loss in the non-hallucination dataset, while CTSD significantly enhances performance on both e-commerce hallucination and general datasets. NLLB-1.3B and mBART-large showed notable improvements of +13.1% and +4.0% in SacreBLEU and +4.7% and +5.47% in COMET, respectively, substantially reducing repetition rates. For LLMs prompted for translation tasks, CTSD demonstrated significant improvements, particularly on the e-commerce hallucination dataset. LLaMA2-7B achieved +10.76% in SacreBLEU and +5.50% in Rouge-L compared to the CT model. Additionally, closed-source models like ChatGPT and GPT-4 scored lower in SacreBLEU but acceptable in COMET, with decent translation capabilities and strong hallucination suppression, while showing weaker professionalism for specialized tasks.

Experiments with the Qwen-1.8B and Qwen-14B models (Table 3) show that CTSD effectively maintains translation accuracy across different LLM sizes, emphasizing its robust enhancement of LLM translation capabilities regardless of hallucination tendencies.

To further verify the effectiveness of CTSD as a repetition suppression method, we conducted experiments comparing different methods, summarized in Table 4. Although the decoding method significantly improved the hallucination dataset (709.86% average increase in SacreBLEU and 88.04% in COMET for NLLB-1.3B), its translation quality still lagged behind training methods. Among the training methods, CTSD stands out on both specialized translation models and LLMs, maintaining a meager repetition rate, indicating it is a general and efficient repetition suppression method.

Model	Method	SacreBLEU↑	Rouge-L↑	COMET↑	rep-2↓	rep-3↓	rep-w↓	rep-r↓	div↑	uniq-1↑
Qwen-7B	Ori.	4.14	0.133	0.599	5.66	5.22	0.04	0.03	0.85	14072
	ConvDPO	1.87	0.082	0.509	5.77	2.73	0.04	0.08	0.90	33832
	TransDPO	0.22	0.026	0.278	49.46	39.15	0.45	0.58	0.21	24292
	TransDPOLoRA	0.67	0.032	0.356	30.11	25.24	0.28	0.31	0.40	13702
	CovDPO+CE	21.53	0.419	0.711	0.75	0.12	0.05	0.02	0.99	10826
	CovDPO+CTSD	21.71	0.421	0.708	0.74	0.21	0.05	0.02	0.99	10778

Table 5: Translation quality and repetition rate of Qwen-7B under DPO and training methods.

Comparison	Metric	AR	DE	TH	HI	HE	PT
CTSD vs. Baseline	LQR-3	+22.22%	+6.91%	+56.51%	+85.12%	+11.72%	+40.22%
	LQR-4	+54.10%	+22.82%	+156.96%	+257.47%	+44.89%	+38.85%
CTSD vs. Google Translate	LQR-3	+9.66%	+7.77%	+22.81%	+11.63%	+8.28%	+6.15%
	LQR-4	+14.93%	+0.10%	+50.00%	+3.33%	+6.51%	+38.45%

Table 6: LQR-3, LQR-4 rates of CTSD compared to baseline and Google Translate under human evaluations.

Metric	PV	UV	CTR	CVR	GMV	RPM
AR	+0.74%	+0.56%	+0.38%	+2.06%	+2.96%	+2.21%
P-value	0.001	0.03	0.002	0.018	0.034	0.031
DE	+0.67%	+0.31%	+0.44%	+1.82%	+0.26%	+0.63%
P-value	0.015	0.002	0.023	0.021	0.013	0.017

Table 7: Online A/B testing results in www.alibaba.com

In order to demonstrate the benefits of CTSD for normal title translations, we translated approximately 1 million e-commerce titles on alibaba.com using models trained with different methods. By filtering the top 1% of repeated titles through the rep-w metric, the final repetition rates of various models are shown in Figure 4. It is evident that the CT and CTSD methods outperformed other baselines, with the rep-2 average decreasing by 920.31% and rep-w by 51.24%, respectively. Compared to CT, the repetition rate of CTSD is slightly higher, which aligns with the nature of word stacking in e-commerce titles. This demonstrates that CTSD can suppress oscillation hallucinations and preserve the natural repetitive characteristics of e-commerce titles in the meantime.

In our final experiment, we evaluated the DPO method's effectiveness in mitigating model oscillation hallucinations, as shown in Table 5. ConvDPO utilized general preference data, while TransDPO and TransDPOLoRA employed private-domain preference data. Within the dataset of translation preferences, the "chosen answer" represented authentic e-commerce translations, and the "rejected answer" represented base model translations. The results clearly demonstrate that whether using generic or private-domain data, DPO fails to address hallucinations effectively for e-commerce translations. Moreover, DPO followed by LoRA fine-tuning for sub-tasks is significantly less effective than direct LoRA fine-tuning of NMT tasks. In summary, while DPO is commonly used for suppressing hallucinations in LLMs, it is evidently ineffective against oscillation in e-commerce contexts. Our results emphatically underline that CTSD is superior for this specific challenge.

4.3 Online E-commerce Experiments

We implemented the CTSD algorithm on the specialized translation model of the www.alibaba.com website, which uses an encoder-decoder structure with 48 layers and approximately 1.1B parameters. We selected six high-traffic language websites (AR - Arabic, DE - German, TH - Thai, HI - Hindi, HE - Hebrew, and PT - Portuguese) to translate item titles and descriptions. These sites serve millions of users, generating nearly 20 million daily page views (PVs). A/B tests were conducted with models fine-tuned on an e-commerce dataset, comparing CTSD and non-CTSD models. Each user saw translation text from only one model to ensure fairness.

First, to ascertain the impact on online translation accuracy, we conducted a pre-procedure expert evaluation. We randomly selected 2,000 items and used two models to translate the corresponding titles and descriptions. These translations were randomly distributed to prevent order effects from influencing the evaluation. Experts were engaged to rate each item based on translation accuracy and smoothness, using a 5-point scale. Only transla-

tions of professional terms that were completely accurate received a rating of 3 or above, while ratings of 4 and 5 required all translations to be both accurate and easy to understand. Each translated text was evaluated by three experts, with LQR-3 (or 4) indicating that at least two experts rated the same translated text above 3 points (or 4 points). Table 6 shows that the CTSD-trained model significantly improved translations across all six languages, particularly for languages with fewer training data (AR, TH, HI, HE). Additionally, the CTSD model outperformed Google Translate overall, with the Fleiss Kappa mean value exceeding 0.6, demonstrating high consistency among raters.

For online evaluations, we assessed business indicators such as page view (PV), retained user (UV), click-through rate (CTR), average conversion rate (CVR), gross merchandise volume (GMV), and revenue per mille (RPM). The online A/B experiments in AR and DE (Table 7) showed that the new translation model improved title translation quality, leading to greater product attention and significant enhancements in all indicators, especially GMV and RPM, which enhanced by 2.96% and 2.21% on Arabic sites, respectively.

5 Conclusion

In conclusion, this study addresses the critical challenge of repetition generation in NMT. By analyzing and visualizing the underlying causes from the lens of information entropy, we propose one novel method, which can dynamically modulate token suppression to reduce the redundancy of some generated words. Extensive experiments on offline general and e-commerce datasets and rigorous online A/B tests have verified its performance in improving translation quality and handling oscillation hallucinations.

6 Limitations

While our CTSD method has shown significant improvements in reducing repetition and enhancing translation quality, there are some limitations to consider. Firstly, the optimal settings for temperature coefficient and decay factor may vary across models and datasets. Automatic tuning for these hyperparameters needs further investigation. Secondly, the additional computations for attention similarities and decay factors during training have not been thoroughly analyzed. Assessing the trade-off between performance gains and computational costs is necessary, especially for resource-limited environments. Addressing these limitations in future work can enhance the robustness and applicability of the CTSD method, contributing to more reliable NMT systems.

7 Ethics Statement

In this work, we employed publicly released and private e-commerce domain datasets to train our machine translation models. Public datasets have been reviewed for ethical concerns, and our inspections found no significant moral issues, such as violent or offensive content. The e-commerce datasets are anonymized and collected with proper consent, following data protection regulations. We also intend to share our source code with clear instructions to encourage ethical use. Despite these precautions, machine translation can sometimes produce unexpected outputs. We will implement mechanisms to reduce such risks and advise users to follow ethical guidelines to prevent misuse.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2022. Introducing claude.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 159 of *NIPS'20*, pages 1877–1901, Vancouver, BC, Canada.

Kaidi Chen, Ben Chen, Dehong Gao, Huangyu Dai, Wen Jiang, Wei Ning, Shanqing Yu, Libin Yang, and

Xiaoyan Cai. 2024. General2specialized llms translation for e-commerce. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 670–673.

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *arXiv preprint arXiv:1906.07978*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

Carlos Escolano, Marta R Costa-jussà, José AR Fonollosa, and Mikel Artetxe. 2020. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *arXiv preprint arXiv:2006.01594*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Javier Ferrando, Gerard I Gállego, Belen Alastruey, Carlos Escolano, and Marta R Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. *arXiv preprint arXiv:2205.11631*.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12848–12856.

Jian Guan, Zhenyu Yang, Rongsheng Zhang, Zhipeng Hu, and Minlie Huang. 2023. Generating coherent narratives by learning dynamic and discrete entity states with a contrastive framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12836–12844.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Shaojie Jiang, Ruqing Zhang, Svitlana Vakulenko, and Maarten de Rijke. 2022. A simple contrastive learning objective for alleviating neural text degeneration. *arXiv preprint arXiv:2205.02517*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Jessy Lin, Geza Kovacs, Aditya Shastry, Joern Wuebker, and John DeNero. 2022. Automatic correction of human translations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–507.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

OpenAI. 2022. Introducing chatgpt.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

Maali Tars, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. *arXiv preprint arXiv:2105.13065*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text degenerate with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2021. Multilingual agreement for multilingual 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 2: Short Papers)*, pages 233–239, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Details of E-Commerce Dataset

The e-commerce dataset comprises 3,500 authentic product titles sourced from Alibaba.com, a leading global B2B e-commerce platform. The selection process focused on identifying titles prone to triggering oscillation hallucinations in baseline models. Specifically:

1. **Initial Collection and Translation:** A substantial corpus of product titles was initially gathered. These titles were subsequently translated using several baseline translation models, specifically NLLB-1.3B, mBART-large, LLaMA2-7B, and Qwen-7B.

2. **Repetition Rate Analysis:** To quantitatively assess the propensity for oscillation hallucinations, each translated title was evaluated using the repetition rate (rep-w) metric. Titles exhibiting high rep-w scores across multiple translation models were earmarked for further analysis.

3. **Human Verification:** The shortlisted titles, identified based on their elevated rep-w scores, underwent a rigorous translation and verification process by human experts. These experts performed meticulous translations and cross-verifications to establish high-fidelity ground truth translations.

B Explanation of Metrics

BLEU (Bilingual Evaluation Understudy) metric is used to assess the accuracy of translations by quantifying the degree of similarity between machine-generated translations and reference translations. The BLEU score is computed using the following formula:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (5)$$

where p_n denotes the precision of n-grams and w_n represents the weights, typically assigned as $w_n = \frac{1}{N}$ (equally weighted). The Brevity Penalty (BP) is incorporated to mitigate the tendency of generating excessively short translations and is defined as:

Table 8: The impact of hyperparameters W, N, and T in CTSD on translation quality and repeatability.

Model	Weight	PredToken	T	SacreBLEU↑	Rouge-L↑	COMET↑	rep-2↓	rep-3↓	rep-w↓	rep-r↓	div↑
NLLB-1.3B	0.1	10	5	7.04	0.187	0.558	61.41	62.02	0.15	0.15	0.16
	0.5	5	5	7.58	0.193	0.585	50.56	51.2	0.18	0.18	0.11
	0.5	5	10	7.78	0.197	0.595	43.86	44.61	0.15	0.15	0.17
	1.0	2	5	**8.19**	0.202	0.616	24.37	24.73	0.10	0.09	0.42
	1.0	5	5	7.99	0.201	0.606	34.98	35.23	0.14	0.13	0.27
	1.0	10	5	8.15	0.202	0.614	26.08	26.32	0.11	0.10	0.39
	2.0	10	5	**8.19**	**0.203**	**0.622**	**12.66**	**12.21**	**0.07**	**0.06**	**0.67**
Qwen-7B	0.005	10	5	23.93	0.451	0.737	**0.62**	**0.09**	**0.04**	0.02	0.99
	0.01	5	5	23.70	0.457	**0.740**	0.73	0.14	0.05	0.02	0.99
	0.01	5	10	**24.35**	0.457	**0.740**	0.75	0.21	0.05	0.02	0.99
	0.02	5	5	24.22	**0.460**	0.739	0.72	0.13	0.05	0.02	0.99
	0.02	10	5	23.97	0.459	0.738	0.67	0.12	0.05	0.02	0.99
	0.02	20	5	22.64	0.437	0.728	0.78	0.22	0.05	0.02	0.99
	0.1	10	5	22.37	0.432	0.717	3.84	3.48	0.05	0.02	0.89

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{6}$$

where c is the length of the candidate translation, and r is the length of the reference translation.

SacreBLEU (Standardized BLEU) serves as an enhanced variant of the original BLEU metric, which introduces a suite of standardized calculation parameters (including tokenizer definitions and n-gram ranges) to ensure the comparability of BLEU scores across different implementations and systems.

Rouge-L (Recall-Oriented Understudy for Gisting Evaluation) is a metric used particularly for summarization and translation quality evaluation. Unlike SacreBLEU, which emphasizes precision, Rouge-L evaluates the longest common subsequence (LCS), effectively assessing how well the generated translation covers the reference translation. The Rouge-L score is computed as follows:

$$\text{Rouge-L} = F_1 = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{7}$$

where precision is defined as the ratio of the LCS length to the length of the candidate translation, while recall is the ratio of the LCS length to the length of the reference translation. The parameter β is utilized to control the balance between precision and recall, with it commonly set to 1 to signify equal importance between the two metrics.

COMET (Cross-lingual Optimized Metric for Evaluation of Translation) is an advanced neural-based metric developed using multilingual transformer models. In this work, we employ the *Unbabel/wmt22-comet-da* model to calculate the COMET metric. COMET leverages representations derived from source sentences, references, and candidate translations to generate a comprehensive quality score. The calculation process of COMET can be delineated as follows:

1. **Embedding Representation:** COMET uses pre-trained models, such as BERT or XLM-R, to generate contextual embeddings for reference, machine-generated, and source text.

2. **Similarity Scoring:** COMET calculates the similarity between these embeddings.

3. **Regression Model:** These similarity scores are fed into a regression model trained on human-annotated translation quality data and return a final quality score.

Unlike traditional evaluation metrics, COMET captures contextual nuances and semantic similarity, providing a more holistic assessment of translation quality.

Rep-n is a metric used to quantify the repetition within the generated text at the n-gram level. The calculation of rep-n involves stripping any leading or trailing whitespace from each text, computing the unique and total n-grams of size n for each text, and updating the counts. The repetition proportion for each n-gram size n is then calculated as follows:

$$\text{rep-n} = 1 - \frac{U_n}{T_n} \tag{8}$$

where U_n is the number of unique n-grams and T_n is the total number of n-grams for a given n. This metric helps assess the text's repetition and can be

Model	Method	SacreBLEU↑	Rouge-L↑	COMET↑	rep-2↓	rep-3↓	rep-w↓	rep-r↓	div↑
NLLB-1.3B	PS	6.47	0.173	0.571	3.69	3.54	0.03	0.02	0.89
	CS	5.03	0.139	0.482	**0.18**	**0.04**	**0.01**	**0.00**	**1.00**
	CE	**6.71**	**0.178**	**0.575**	36.17	37.21	0.13	0.10	0.24
	CE + PS	6.64	0.176	0.572	2.33	1.86	0.03	0.02	0.91
	CE + CS	6.03	0.157	0.532	**0.18**	0.05	**0.01**	**0.00**	**1.00**

Table 9: Translation Quality and Repetition Rate of NLLB-1.3B with and without Training Under Repetition Suppression Methods in the Inference Stage.

applied across different n-gram sizes to evaluate the generated translations comprehensively.

Rep-w is calculated by the proportion of current tokens occurring within the previous w tokens, expressed as:

$$\text{rep-w} = \frac{1}{|\mathcal{D}|}\sum_{s\in\mathcal{D}}\frac{1}{|s|}\sum_{t=1}^{|s|}\mathbf{1}\left[s_t \in s_{t-w-1:t-1}\right] \quad (9)$$

where $\mathcal{D}$ represents the result set, s represents generated sentences in $\mathcal{D}$. We selected w = 8 for the rep-w metric in this paper, which means we consider the previous eight words. This is a standard setting used in most papers and evaluations in the field. Furthermore, we analyzed public translation datasets comprising about 193 million sentences. Our findings show that for nearly 95% of the data, the interval between identical words within a single sentence is less than 8.

Rep-r stands for the ratio of the repetition snippet in a sentence measured by length, defined as:

$$\text{rep-r} = \frac{1}{|s|}\left|\left\{ i \mid (s_i = s_j \wedge s_{i+1} = s_{j+1}, \exists j \neq i) \vee (s_i = s_k \wedge s_{i-1} = s_{k-1}, \exists k \neq i)\right\}\right| \quad (10)$$

where s represents a generated sentence, $|s|$ is the length of the sentence s, s_i is the i-th token in the sentence s, $\wedge$ and $\vee$ are logical "and" and "or" respectively. Rep-r can effectively assess text repetitiveness without the need to set any hyperparameters, making it highly useful for monitoring hallucination in machine translation.

Div (diversity) is a metric used to analyze the lexical diversity of generated sentences. It considers n-grams (2, 3, and 4-grams, as most terms no longer than 4 words.) and measures the proportion of unique n-grams to the total number of n-grams. A higher 'div' score indicates less repetition and greater diversity in the generated text.

Uniq-1 (unique unigrams) counts the unique tokens in the entire dataset. This metric reflects the overall vocabulary diversity of the model's translations. A higher 'uniq-1' score suggests that the model uses a more diverse set of words across all translations.

C Additional Visualization Details and Analysis

C.1 Extended Analysis of ALTI+

To provide more rigorous statistical evidence supporting our initial observations in Section 3.1, we conducted an additional large-scale analysis. We analyzed 500 translation samples with oscillation hallucinations and 500 samples without hallucinations. For each output token of each sample, we calculated a contribution vector using the ALTI+ method, representing how much each input token influences it.

Type	10%	30%	50%	70%	90%
Hallucination	0.35	0.57	0.63	0.78	0.93
Non-Hallucination	0.01	0.01	0.03	0.08	0.11

Table 10: Contribution Similarity Percentiles for 500 Sentences. (The 10% (30%-90%) columns represent the similarity of the bottom 10% (30%-90%) of sentences based on median similarity. "Hallucination" refers to sentences with oscillation hallucinations, and "Non-Hallucination" refers to sentences without hallucinations)

Our statistical analysis corroborates our initial observations and hypotheses, indicating that in cases of hallucination, each input token contributes similarly to each output token, resulting in repetitive generation (see Table 10).

C.2 In-Depth Clarification of Figures

ALTI+ Analysis Method Figure 1 illustrates the ALTI+ analysis method. The x-axis represents input tokens (English source tokens for the encoder

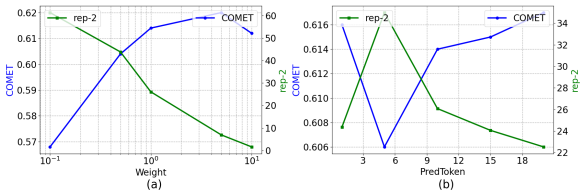

Figure 5: The impact of hyperparameters W and N on the translation quality and reproducibility of the NLLB-1.3B model.

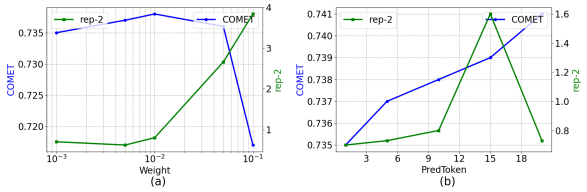

Figure 6: The impact of hyperparameters W and N on the translation quality and reproducibility of the Qwen-7B model.

and previously generated German tokens for the decoder), and the y-axis shows the generated German tokens at each decoding step. Each cell in the heatmap indicates the contribution of the input embedding vector to the output embedding vector in each Multi-Head Attention (MHA) layer. The contribution is calculated using the Manhattan distance between vectors. For example, in the first image of Figure 1, the first generated German token "King" on the vertical axis is mainly generated by the input English tokens "King", "Ge", and "ar".

Visualization of CTSD Loss Figure 3(a) illustrates the cosine similarity between the contribution degrees of input tokens (denoted as α_s in the CTSD loss). For example, consider the similarity of attention between the first German output token "Hei" on the horizontal axis and the German output token "Verka" on the vertical axis. This similarity, located at coordinate (2, 4), is determined by calculating the cosine similarity between two vectors: one representing the contribution of input tokens to "Hei" and the other representing the contribution of input tokens to "Verka".

Figure 3(b) depicts token distances (denoted as α_d in the CTSD loss). For instance, if the output distance between the output token "Hei" on the horizontal axis and the output token "Verka" on the vertical axis is 2, then the value at coordinate (2, 4) in the corresponding heatmap is $e^{-2/T}$.

D Additional Experimental Results

D.1 CTSD Ablation Experiment

We discuss the impact of the three hyperparameters W, N, and T within the CTSD algorithm on the translation quality and repetition performance across various models, where W represents the ratio of CE loss with CTSD loss, N denotes the number of previous tokens to focus on. T represents the temperature coefficient of α_d.

The analysis presented in Table 8 and Figure 5 demonstrates that for specialized translation models like NLLB-1.3B, a moderate increase in T can lead to an improvement in translation quality with repetition rates exhibiting the opposite trend. On the other hand, when W increases, the translation quality first increases and then decreases while the repetition rate continues to decline. However, as N changes, there is an initial decrease in translation quality followed by an increase that improves and then deteriorates, and the repetition rate consistently declines. These results suggest that incorporating a suitable CTSD loss into specialized model training and setting a larger window can effectively reduce repetition rates and enhance model performance. However, it is essential to note that an excessively high weight on the CTSD loss can disrupt the original training direction of the model. Moreover, larger windows require increased computational demands, which poses a trade-off between accuracy and training duration. However, the results for Qwen-7B exhibit a distinct pattern. As shown in Figure 6, when W increases, the repetition rate initially decreases and then rises. This implies that increasing CTSD loss continuously during large model training diminishes the translation quality and induces new oscillatory hallucinations.

D.2 Analysis of Hybrid Suppression Methods

We also investigated hybrid suppression methods that combine training and decoding approaches. The results suggest that integrating fine-tuning with decoding methods often leads to over-suppression, which adversely affects translation quality. In contrast, concentrating solely on suppressing oscillation hallucinations during the training phase yields better outcomes.

Table 9 demonstrates that while methods such as CE + PS and CE + CS reduce repetition rates, they do so at the expense of translation quality, as evidenced by lower SacreBLEU and COMET

scores compared to methods that focus exclusively on one stage (e.g., CE). For instance, CE alone achieves the highest SacreBLEU score of 6.71 and a COMET score of 0.575. In contrast, hybrid methods like CE + PS and CE + CS exhibit lower scores and are susceptible to over-suppression.

Unless an absolute zero tolerance for oscillation hallucinations is required, employing methods that concentrate on a single suppression stage is more advantageous to maintain higher translation quality.