

GRAD: GRADient Rectification based on Authentic Data for Low-Resource Neural Machine Translation with Back Translation

Zhuofei Liang, Xinglin Lyu, Lingling Mu, Hongfei Xu

Zhengzhou University, Henan, China

zfliangnlp@foxmail.com, {xllyu, iellmu}@zzu.edu.cn

Correspondence: Hongfei Xu hfxunlp@foxmail.com

Abstract

Synthetic data augmentation, especially back-translation with Large Language Models (LLMs), has become a standard strategy for low-resource Neural Machine Translation (NMT). However, jointly training on synthetic and authentic data may suffer from conflicts. We propose **GRADient Rectification based on Authentic Data (GRAD)**, an asymmetric gradient-level framework that treats authentic parallel data as an **Authentic Anchor**, and rectifies conflicting synthetic updates through orthogonal projection while preserving non-conflicting components. Furthermore, we leverage gradients of multiple authentic batch for comparison and use the most similar one for gradient rectification to alleviate information loss with the synthetic data. Experiments on six low-resource pairs across three LLM families (Qwen, Gemma, Llama) and different scales (4B/8B) show that GRAD consistently outperforms supervised fine-tuning, back-translation, and multi-stage baselines. Notably, on the lowest-resource En-Kk pair, GRAD achieves the highest COMET (**83.65**) among all evaluated systems. Further analysis shows that synthetic-induced gradient conflicts are localized, and parameter-wise rectification substantially outperforms global-wise rectification.

1 Introduction

Recent Large Language Models (LLMs) have substantially advanced Neural Machine Translation (NMT), enabling remarkable improvements in fluency, contextual understanding, and multilingual generalization (Brown et al., 2020; OpenAI et al., 2024; Yang et al., 2025; Team et al., 2025; Grattafiori et al., 2024). Despite these advances, the effectiveness of NMT systems remains heavily dependent on high-quality parallel corpora, which are often scarce, domain-restricted, or noisy in low-resource settings (Koehn and Knowles, 2017; Gu et al., 2018). Improving translation quality under

limited supervision therefore remains a fundamental challenge in multilingual NLP.

To alleviate data scarcity, synthetic data augmentation has become a standard strategy for low-resource NMT. Back-Translation (BT) (Sennrich et al., 2016; He et al., 2020) leverages target-side monolingual data to generate synthetic parallel corpora, and recent work has further improved synthetic data quality and diversity by using LLMs as translators (Wang et al., 2023; Jiao et al., 2023; de Gibert et al., 2025; Zebaze et al., 2025; Yang and Nicolai, 2023). However, scaling up synthetic data does not necessarily improve translation quality and may even degrade the performance (Edunov et al., 2018; Jiao et al., 2021; Arif, 2025; Ahmadi et al., 2025). Although LLM-generated translations are fluent, they may contain hallucinations, omissions, and semantic inconsistencies—errors that are particularly harmful when authentic supervision is already scarce.

When synthetic and authentic data are mixed during training, their gradients may pull model parameters toward conflicting optimization directions: authentic data primarily enforces semantic faithfulness, while noisy synthetic data may encourage stylistic fluency or even semantically incorrect generation patterns. Existing approaches address this through data-level interventions—quality filtering, curriculum learning, or multi-stage fine-tuning. But they cannot dynamically determine, at the parameter level, whether a synthetic update is semantically beneficial or harmful during training.

In this work, we revisit low-resource back-translation from the perspective of **gradient geometry**. We hypothesize that conflicts between synthetic and authentic supervision are reflected in the alignment between their gradients, and that authentic data should serve as a strict **Authentic Anchor** that rectifies unreliable synthetic updates. Based on this insight, we propose **GRADient Rectification based on Authentic Data (GRAD)**, an asymmetric

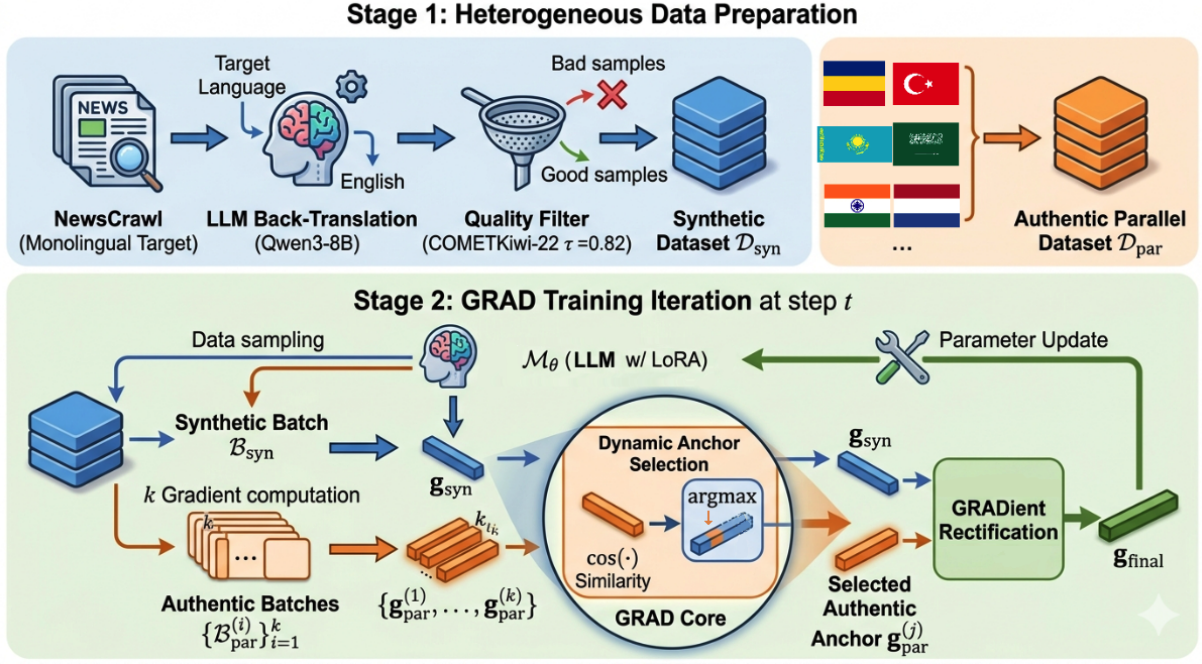


Figure 1: The overall architecture and training pipeline of our GRAD framework. **Stage 1 (Top)**: Heterogeneous data preparation, where back-translated synthetic data and authentic parallel corpora are collected and filtered. **Stage 2 (Bottom)**: GRAD training iteration at step t , illustrating the process of gradient computation ($\mathbf{g}_{syn}$ and $\mathbf{g}_{par}$), optimal authentic gradient anchor selection, and parameter updating with the rectified gradient $\mathbf{g}_{final}$.

optimization framework as illustrated in Figure 1. At each training step, GRAD computes gradients of both synthetic and authentic batches and measures their cosine similarity. When the gradients align, GRAD combines them to enrich the learning signal; when they conflict, GRAD projects the synthetic gradient onto the subspace orthogonal to the authentic gradient, removing only the conflicting component while preserving non-conflicting information.

Unlike symmetric gradient manipulation methods in multi-task learning (Yu et al., 2020; Wang et al., 2020), GRAD is fundamentally **asymmetric**: it does not treat synthetic and authentic data as equally reliable supervision sources, but instead anchors optimization to authentic supervision throughout training. We further show that synthetic-induced gradient conflicts are **highly localized** in the parameter space: fine-grained parameter-wise rectification consistently outperforms global and module-level strategies, indicating that conflicts emerge in specific optimization regions rather than uniformly across the model.

We conduct extensive experiments on six low-resource language pairs spanning Turkic, Indo-European, and Afro-Asiatic languages. GRAD consistently outperforms supervised fine-tuning, back-

translation, and multi-stage training baselines, and remains competitive against dedicated multilingual MT systems. Notably, on the lowest-resource En-Kk task, GRAD obtains the best COMET (**83.65**) among all compared systems. Further analysis reveals that batches triggering negative gradient similarity frequently contain hallucinations, suggesting that gradient disagreement implicitly captures unreliability in synthetic supervision. GRAD also generalizes across LLM families (Qwen, Gemma, Llama).

Our main contributions are as follows:

- We propose **GRAD**, an asymmetric gradient rectification framework that treats authentic parallel data as an **Authentic Anchor**, and dynamically filters harmful synthetic updates during low-resource NMT training.
- We propose to compare the gradients of multiple authentic batches with the gradient of the synthetic batch, and to use the most similar one to reduce the potential information loss.
- We show that synthetic-induced gradient conflicts are highly localized in the parameter space, and that fine-grained parameter-wise rectification consistently outperforms global and module-level strategies.

- Extensive experiments across six low-resource language pairs, three LLM families and two model scales demonstrate that GRAD consistently improves translation quality, attains the best performance on average and on the lowest-resource En-Kk task compared to baselines.

2 Our Method

2.1 Problem Formulation

We fine-tune a model $\mathcal{M}_\theta$ using two distinct datasets: the **Synthetic Back-Translation Dataset** $\mathcal{D}_{syn}$ and the **Authentic Parallel Dataset** $\mathcal{D}_{par}$.

At each training step t , we sample a synthetic batch $\mathcal{B}_{syn}$ and an authentic batch $\mathcal{B}_{par}$ and compute their gradients independently:

$$\mathbf{g}_{syn} = \nabla_\theta \mathcal{L}(\mathcal{B}_{syn}), \quad (1)$$

$$\mathbf{g}_{par} = \nabla_\theta \mathcal{L}(\mathcal{B}_{par}). \quad (2)$$

Our goal is to leverage the diversity and fluency benefits of synthetic back-translation data without interfering the optimization direction induced by authentic parallel data. i.e., authentic gradient serves as a reliable optimization anchor, while the synthetic gradient provides complementary learning signals.

To measure the geometric consistency between the two gradients, we compute their cosine similarity:

$$\rho = \cos(\mathbf{g}_{syn}, \mathbf{g}_{par}). \quad (3)$$

Note that parameters are **NOT** updated during gradient computation, ensuring that all gradients lie in the same tangent space.

2.2 GRADient Rectification based on Authentic Data

We dynamically modulate the update vector $\mathbf{g}_{final}$ according to the alignment between the synthetic gradient $\mathbf{g}_{syn}$ and the authentic anchor gradient $\mathbf{g}_{par}$.

Case 1: Signal Enrichment ($\rho > 0$). When the gradients are positively aligned, the synthetic data is considered consistent with the authentic optimization direction. In this case, we directly combine the two gradients to enrich the learning signal:

$$\mathbf{g}_{final} = \mathbf{g}_{syn} + \mathbf{g}_{par} \quad (4)$$

Case 2: Conflict Rectification ($\rho < 0$). A negative cosine similarity indicates that the synthetic gradient conflicts with the authentic optimization direction. To preserve the reliable semantic signal from authentic data, we remove the conflicting component of $\mathbf{g}_{syn}$ by projecting it onto the subspace orthogonal to $\mathbf{g}_{par}$:

$$\mathbf{g}_{syn}^\perp = \mathbf{g}_{syn} - \frac{\mathbf{g}_{syn} \cdot \mathbf{g}_{par}}{\|\mathbf{g}_{par}\|^2} \mathbf{g}_{par} \quad (5)$$

And the final update direction is:

$$\mathbf{g}_{final} = \mathbf{g}_{syn}^\perp + \mathbf{g}_{par} \quad (6)$$

Geometrically, the rectified gradient removes the component that directly opposes the authentic optimization direction while preserving the non-conflicting information carried by synthetic data.

2.3 Multi-Reference Gradient Rectification

A single authentic batch as the optimization anchor may not fully reflect beneficial optimization directions. To reduce the information loss of synthetic batches due to the variance of authentic batches, we extend GRAD to a multi-reference setting by sampling multiple authentic batches at each training step.

Specifically, for one synthetic batch $\mathcal{B}_{syn}$, we sample a set of k authentic batches $\{\mathcal{B}_{par}^{(i)}\}_{i=1}^k$ and independently compute their gradients:

$$\mathbf{g}_{par}^{(i)} = \nabla_\theta \mathcal{L}(\mathcal{B}_{par}^{(i)}), \quad i \in \{1, \dots, k\}. \quad (7)$$

We then compute the cosine similarity between the synthetic gradient and each authentic gradient:

$$\rho_i = \cos(\mathbf{g}_{syn}, \mathbf{g}_{par}^{(i)}). \quad (8)$$

Instead of treating all authentic gradients equally, GRAD dynamically selects the most relevant authentic anchor of the highest similarity:

$$j = \arg \max_i \rho_i. \quad (9)$$

The selected authentic gradient $\mathbf{g}_{par}^{(j)}$ is subsequently used as the anchor gradient in the rectification process described in Section 2.2.

This strategy allows GRAD to dynamically identify the authentic batch whose optimization direction is most compatible with the current synthetic signal, thereby alleviating information loss while resolving gradient conflicts.

ISO	Language	Script	Family
en	English	Latin	Indo-European
ar	Arabic	Arabic	Afro-Asiatic
nl	Dutch	Latin	Indo-European
tr	Turkish	Latin	Turkic
kk	Kazakh	Cyrillic	Turkic
ro	Romanian	Latin	Indo-European
gu	Gujarati	Gujarati	Indo-European

Table 1: Overview of the evaluated languages, including scripts and language families.

Language	Syn Size	Retention Rate	Par Origin	Par Size
En-Gu	672K	67.32%	WMT19	10K
En-Kk	571K	57.16%	WMT19	91K
En-Tr	629K	62.95%	WMT17	207K
En-Nl	737K	73.71%	IWSLT17	237K
En-Ar	294K	29.47%	IWSLT17	250K
En-Ro	816K	81.64%	WMT16	608K

Table 2: Statistics of the synthetic back-translation data and authentic parallel data. *Syn Size* denotes the number of synthetic sentence pairs retained after filtering, while *Retention Rate* reports the corresponding proportion relative to the initial 1M monolingual sentences. *Par Origin* and *Par Size* indicate the data source and the number of authentic parallel sentence pairs used for the translation direction respectively.

3 Experiments

3.1 Data

We conduct experiments on six English-centric translation directions covering a typologically diverse set of target languages. As summarized in Table 1, these languages span multiple language families, including Indo-European, Turkic, and Afro-Asiatic, as well as distinct writing systems such as Latin, Cyrillic, Arabic, and Gujarati. This diversity allows us to evaluate the robustness of our framework under different morphological and orthographic conditions.

Back-Translation via LLM Following the back-translation paradigm, we adopted Qwen3-8B to construct synthetic parallel data (D_{syn}). For each translation direction, we sampled 1M monolingual sentences on the target side from NewsCrawl and back-translated them into English using the LLM. To mitigate the risk of low-quality synthetic signals, we employed COMETKiwi-22 (Rei et al., 2022b) as a quality filter and discarded sentence pairs with quality scores below $\tau = 0.82$.

Parallel Data We collected 6 publicly available low-resource parallel corpora (<1M sentence pairs)

from previous WMT (En $\rightarrow$ Gu, Kk, Tr, Ro) and IWSLT (En $\rightarrow$ Nl, Ar) shared tasks.

The statistics of the resulted dataset are shown in Table 2.

3.2 Setup

Training Setup We adopted Qwen3-8B as the backbone model for our main experiments and fine-tuned it with LoRA (Hu et al., 2022) using a rank of $r = 256$. LoRA was applied to both the attention and MLP layers. We used the Adam optimizer with $\beta = (0.9, 0.98)$, $\epsilon = 10^{-9}$, without weight decay, and employing a constant learning rate of 1×10^{-5} . The models were fine-tuned using the BFloat16 mixed precision. For both synthetic and parallel datasets, we applied the same length constraint and retained only sentence pairs of at most 256 tokens. The effective batch size was 3072 target tokens facilitated via gradient accumulation. All models were trained for 100K steps. For GRAD, we set the number of authentic references to $k = 4$.

Evaluation We used a beam size of 4 for decoding with the best-performing model on the validation set saved in an interval of 1500 training steps. We evaluated with tokenized BLEU implemented by the sacreBLEU toolkit (Post, 2018) and the COMET-22 score (Rei et al., 2022a). The signature for BLEU was `nrefs:1|case:mixed|eff:no|tok:none|smooth:exp|version:2.5.1` for all language pairs except En-Ar, for which we used the FLORES-200 tokenizer (`tok:flores200`) with otherwise identical settings. For COMET, we used the `Unbabel/wmt22-comet-da` model across all language pairs. we used the official validation and test sets from WMT and IWSLT corresponding to each language pair.

Models for Comparison We compare our method against three categories: **General-purpose LLMs** including Gemma-3-4B-it (Team et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), LLaMAX3-8B-Alpaca (Lu et al., 2024) and Aya-Expanse-8B (Dang et al., 2024). **Dedicated MMT Models**, including GemmaX2-28-9B (Cui et al., 2025) and Hunyuan-MT-7B (Zheng et al., 2025).

3.3 Overall Performance

Table 3 reports translation performance across six low-resource language pairs. Our GRAD framework achieves the highest average BLEU (**28.94**) and COMET (**86.32**) among systems evaluated on

Models	En-Gu		En-Kk		En-Tr		En-Nl		En-Ar		En-Ro		Avg	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Gemma-3-4B-it	<u>25.75</u>	83.75	4.52	69.55	20.95	86.28	31.45	85.28	18.09	82.17	36.12	87.83	22.81	82.48
Llama-3.1-8B-Instruct	20.32	74.37	4.99	72.84	20.97	85.02	33.05	85.64	16.61	80.04	35.53	87.52	21.91	80.91
LLaMAX3-8B-Alpaca	23.31	78.78	6.58	<u>79.21</u>	17.61	84.49	30.17	83.95	16.09	79.35	32.39	85.97	21.03	81.96
Aya-expanse-8B	—	—	—	—	24.11	88.06	32.70	86.42	20.28	83.35	37.83	<u>89.15</u>	—	—
Hunyuan-MT-7B	20.56	86.40	0.61	52.09	22.76	90.25	21.07	85.61	14.81	84.71	31.80	89.17	18.60	81.37
GemmaX2-28-9B	—	—	—	—	30.79	<u>89.77</u>	34.92	87.37	20.90	83.54	—	—	—	—
Qwen3-8B Zero-Shot	19.14	79.49	4.85	73.32	20.29	85.77	30.90	84.97	17.10	81.62	33.21	86.72	20.92	81.98
Qwen3-8B SFT	19.21	78.49	2.17	67.91	24.01	88.24	<u>35.50</u>	86.38	22.85	83.52	36.55	88.92	23.38	82.24
Qwen3-8B BT	29.12	85.08	10.23	82.17	29.05	88.73	35.29	86.56	<u>23.09</u>	83.60	<u>40.38</u>	88.75	<u>27.86</u>	<u>85.82</u>
Ours (Qwen3-8B + GRAD)	29.98	<u>85.64</u>	11.91	83.65	<u>29.79</u>	89.07	36.03	<u>86.79</u>	23.39	<u>83.76</u>	42.55	88.98	28.94	86.32

Table 3: BLEU and COMET scores on six language pairs. **Bold** indicates the best result, and underline indicates the second best. “—” denotes language pairs not supported by the model. Average performances are only computed for systems evaluated on all six pairs.

all six pairs, demonstrating that explicit gradient-level rectification of synthetic and authentic supervision is a more effective training paradigm than conventional data mixing or sequential fine-tuning.

GRAD vs. SFT: avoiding low-resource collapse.

Standard SFT performs poorly in extremely low-resource settings. On En-Kk, SFT obtains only 2.17 BLEU, which are even lower than the zero-shot baseline (4.85), indicating severe overfitting to the limited authentic data. GRAD instead reaches **11.91 BLEU** and **83.65 COMET**, and delivers a similar BLEU gain on En-Gu (29.98 vs. 19.21). This shows that GRAD effectively leverages large-scale synthetic data while maintaining stable optimization under scarce parallel supervision.

GRAD vs. BT: consistent gains across all pairs.

Compared to standard back-translation, GRAD improves both BLEU and COMET on all the six language pairs, with the largest gains on En-Ro (+2.17 BLEU) and En-Kk (+1.68 BLEU / +1.48 COMET). The improvements span Turkic (Kazakh, Turkish), Indo-European (Romanian, Dutch, Gujarati), and Afro-Asiatic (Arabic) languages, suggesting that gradient-level coordination robustly adapts to heterogeneous synthetic supervision regardless of language family or writing system.

GRAD vs. dedicated MT systems. Despite only using the general Qwen3-8B model, GRAD remains highly competitive against strong translation-oriented baselines. It obtains the highest BLEU on five of six language pairs (En-Gu, En-Kk, En-Nl, En-Ar, and En-Ro), and the second-best BLEU on En-Tr (29.79 vs. GemmaX2-28-9B’s 30.79). On COMET, GRAD achieves the best scores on the most challenging En-Kk task (**83.65**), where all dedicated MT systems struggle: LLaMAX3-8B-Alpaca reaches only 79.21 and Hunyuan-MT-

7B collapses to 52.09. On the remaining language pairs, GRAD ranks consistently among the top systems while specialized baselines exhibit unbalanced behavior: Hunyuan-MT-7B, for instance, achieves the highest COMET on four tasks but collapses on En-Kk (0.61 BLEU / 52.09 COMET). This contrast suggests that resolving synthetic–authentic optimization conflicts may matter more for low-resource translation than scaling model sizes or multilingual pre-training alone.

3.4 Generalization across Model Families and Sizes

To verify that GRAD’s effectiveness is not specific to the Qwen3-8B model, we further evaluated it on two alternative LLM backbones: Gemma-3-4B-it and Llama-3.1-8B-Instruct. Table 4 shows the results on the low-resource En-Kk and medium-resource En-Ro tasks.

GRAD consistently improves over BT across all settings.

Across both backbones and both language pairs, GRAD outperforms standard back-translation on both BLEU and COMET metrics. On En-Kk, GRAD improves Gemma-3-4B-it by +1.16 BLEU / +1.98 COMET and Llama-3.1-8B-Instruct by +0.60 BLEU / +0.75 COMET. On En-Ro, both backbones gain +2.16 BLEU, with Llama-3.1-8B-Instruct additionally reaching the highest COMET score of 89.20. All these results together with the Qwen3 results in Table 3, indicate that gradient-level coordination between synthetic and authentic supervision generalizes across distinct tokenizers, pretraining corpora, and post-training recipes.

SFT instability further highlights the value of GRAD. A notable observation is that on the extremely low-resource En-Kk task, SFT even under

Methods	Gemma-3-4B-it				Llama-3.1-8B-Instruct			
	En-Kk		En-Ro		En-Kk		En-Ro	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Zero-Shot	4.52	69.55	36.12	87.83	4.99	72.84	35.53	87.52
SFT	2.74	69.75	36.74	89.03	1.80	67.37	36.06	88.94
BT	11.17	83.06	40.57	88.19	10.26	82.51	40.27	88.90
Ours (GRAD)	12.33	85.04	42.73	88.97	10.86	83.26	42.43	89.20

Table 4: Generalization of GRAD across different model architectures. GRAD consistently improves over standard back-translation on both low-resource and medium-resource translation tasks.

Granularity	En-Gu		En-Tr	
	BLEU	COMET	BLEU	COMET
Global-level	29.11	85.36	28.85	88.71
LoRA-level	29.87	85.48	29.49	88.94
Parameter-wise	29.98	85.53	29.79	89.07

Table 5: Effect of gradient rectification granularity. Finer-grained projection consistently improves performance, suggesting that optimization conflicts induced by synthetic data are highly localized within the parameter space.

performs the Zero-Shot baseline with both Gemma (2.74 vs. 4.52 BLEU) and Llama (1.80 vs. 4.99 BLEU). This indicates that directly fine-tuning on a small authentic corpus can collapse the model’s pretrained translation ability, whereas GRAD remains robust by anchoring optimization to authentic supervision while preserving the diversity of synthetic data. The contrast suggests that the benefits of GRAD stem not merely from access to more data, but from explicitly coordinating heterogeneous supervision through gradient geometry—a property that transfers across backbone families.

4 Analysis

We analyze why gradient rectification works after verifying the GRAD’s effectiveness across LLMs.

4.1 Granularity of Gradient Rectification

A key design choice in GRAD is the granularity at which gradient conflicts are measured and rectified. Since synthetic-induced conflicts may not be uniformly distributed across the parameter space, we compare three rectification strategies: **Global-level**, which concatenates all trainable gradients into a single vector and performs one unified projection, **LoRA-level**, which computes similarity and projection independently for each LoRA adapter module, and **Parameter-wise**, which operates inde-

pendently one very parameter tensor for the finest-grained rectification.

Table 5 shows that finer-grained rectification consistently yields better translation quality. Moving from Global-level to LoRA-level brings the largest jump (e.g., +0.76 BLEU on En-Gu and +0.64 BLEU on En-Tr), indicating that gradient conflicts already exhibit strong structural locality at the adapter-module level. Parameter-wise rectification further improves over LoRA-level on both language pairs and achieves the best BLEU (**29.98/29.79**) and COMET (**85.53/89.07**) scores, suggesting that conflicting and beneficial signals can coexist even within the same LoRA module.

This suggests that gradient conflicts induced by synthetic supervision are highly localized rather than globally consistent. Globally projecting the entire synthetic gradient risks over-suppressing beneficial signals in unrelated regions, whereas parameter-wise projection isolates only the conflicting components while preserving useful synthetic information. This indicates that the effectiveness of GRAD stems not merely from gradient projection itself, but from accurately identifying *where* conflicts emerge in the parameter space.

4.2 Ablation Study

To better understand which components drive the performance gains of GRAD, we conduct ablation experiments on **En-Kk**, **En-Tr**, and **En-Ro**, covering the full resource spectrum from extremely low-resource (En-Kk), through medium-resource (En-Tr), to relatively high-resource (En-Ro). This pairing allows us to verify whether each component of GRAD remains effective across different supervision conditions. Results are shown in Table 6.

Single-Source Training. We first compare against supervised fine-tuning using only authentic parallel data (*SFT Parallel Only*) or only syn-

Methods	En-Kk		En-Tr		En-Ro	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
SFT (Parallel Only)	2.17	67.91	24.01	88.24	36.55	88.92
SFT (Synthetic Only)	<u>10.54</u>	82.71	28.08	88.03	40.89	88.24
Two-Stage (Syn $\rightarrow$ Par)	4.69	76.89	26.75	88.70	38.95	88.82
PCGrad	10.38	<u>82.98</u>	<u>28.62</u>	<u>88.07</u>	<u>40.08</u>	<u>88.70</u>
GRAD (Full)	11.91	83.65	29.79	89.07	42.55	88.98

Table 6: Ablation results on English-Kazakh (En-Kk), English-Turkish (En-Tr), and English-Romanian (En-Ro) translation tasks, covering low-, medium-, and high-resource regimes. PCGrad uses $\mathbf{g}_{par}^{(j)}$ solely for gradient projection without adding it to the final gradient for parameter update. **Bold** numbers indicate the best results, and underlined numbers indicate the second best results for each language pair.

thetic back-translation data (*SFT Synthetic Only*). Parallel-only training performs poorly, especially on En-Kk (2.17 BLEU), confirming that the limited authentic corpus alone is insufficient for robust translation. Synthetic-only training substantially improves BLEU and COMET, demonstrating the value of large-scale synthetic supervision, yet it remains consistently suboptimal compared with GRAD across all three language pairs. This observation motivates GRAD’s core design, leveraging the richness of synthetic data while anchoring optimization to authentic supervision.

GRAD vs. Two-Stage Training. We next compare GRAD against the conventional two-stage strategy (*Synthetic $\rightarrow$ Parallel*). GRAD consistently outperforms two-stage training across all three language pairs (e.g., **11.91** vs. 4.69 BLEU on En-Kk, **42.55** vs. 38.95 BLEU on En-Ro). Strikingly, on En-Ro, two-stage training even underperforms synthetic-only fine-tuning (38.95 vs. 40.89 BLEU), indicating that sequential fine-tuning may actively harm the model, not merely fails to coordinate the two supervision sources. We attribute this gap to *optimization decoupling*: the synthetic stage absorbs both beneficial diversity and noisy supervision indiscriminately, and the subsequent parallel stage cannot identify which synthetic updates were harmful. In contrast, GRAD performs real-time gradient-level coordination, dynamically suppresses conflicting synthetic gradients while preserving constructive ones throughout training.

The Role of Gradient Rectification. To isolate the effect of conflicting projection itself, we evaluate *PCGrad* (Yu et al., 2020), which uses the authentic gradient solely as a projection anchor without adding $\mathbf{g}_{par}^{(j)}$ to the final update. On the lower-resource En-Kk and En-Tr tasks, this variant consistently improves over *SFT Synthetic Only*

Methods	En-Gu		En-Kk		En-Ro	
	BLEU	COMET	BLEU	COMET	BLEU	COMET
Single-Ref	29.33	85.44	10.69	82.98	42.08	88.61
Multi-Ref	29.98	85.64	11.91	83.65	42.55	88.98

Table 7: Effects of using single-reference versus multi-reference authentic supervision for GRAD.

(+0.27 COMET on En-Kk, +0.54 BLEU on En-Tr), confirming that orthogonal projection effectively removes harmful optimization directions when synthetic supervision is noisy. On the higher-resource En-Ro task, however, projection alone slightly underperforms synthetic-only training (40.08 vs. 40.89 BLEU), suggesting that when synthetic data is already relatively reliable, simply suppressing conflicting components is insufficient to drive further gains. The full GRAD framework recovers and surpasses this gap on En-Ro (**42.55** BLEU) and yields substantially larger improvements on the other two pairs. This pattern reveals that the authentic gradient plays a *dual role*: it not only filters conflicting synthetic updates as a geometric constraint, but also directly contributes a reliable optimization signal to the parameter update. Both components are necessary—neither rectification alone nor anchor injection alone fully captures the benefits of GRAD, validating our central *Authentic-Anchor* hypothesis.

Single-Reference vs Multi-Reference Rectification. Table 7 compares GRAD with single-reference and multi-reference authentic anchors. Multi-reference supervision consistently improves translation quality across all three evaluated pairs, with the largest gain on the low-resource En-Kk task (+1.22 BLEU, +0.67 COMET). We attribute this to reduced mini-batch gradient variance: sampling multiple authentic batches and selecting the most aligned anchor stabilizes the projection direc-

tion, which is particularly valuable when authentic data is scarce and individual batches are more susceptible to noise.

4.3 Case Study: What does Gradient Similarity Reflect?

To better understand the behavior of gradient alignment in GRAD, we conducted a human analysis on 75 randomly sampled training steps with negative cosine similarity ($\rho < 0$) and 75 steps with positive cosine similarity ($\rho > 0$).

Quantitative Insights. Our manual inspection reveals a clear difference in synthetic data quality between the two groups. Among the sampled negative-alignment cases ($\rho < 0$), we identify 12 synthetic translations containing clear hallucinations or factual inconsistencies, while only 5 of such cases are observed in the positive-alignment group ($\rho > 0$). Although the analysis is based on a limited sample, the trend is consistent across the inspected cases.

These observations suggest that negative gradient alignment is more frequently associated with problematic synthetic supervision signals. In particular, many negative-alignment examples contain severe semantic errors, incorrect entity substitutions, or fabricated content that conflicts with the authentic translation objective.

Qualitative Analysis. Table 8 presents representative examples illustrating the different behaviors captured by GRAD.

5 Related Work

Low-resource NMT. Limited parallel data remains a central challenge for low-resource neural machine translation, driving extensive research on low-resource settings (Liu et al., 2020; Lin et al., 2020; Li et al., 2022; Haddow et al., 2022; Yazar et al., 2023). Data augmentation is a widely used method for NMT, evolving from back-translation (Sennrich et al., 2016; He et al., 2020) to LLM-based synthesis (Wang et al., 2023; Jiao et al., 2023; OpenAI et al., 2024). While LLMs demonstrate remarkable fluency via in-context learning (Vilar et al., 2023; Garcia et al., 2023; Pei et al., 2025; de Gibert et al., 2025), they inevitably suffer from hallucinations and semantic drift (Bang et al., 2023; Ji et al., 2025; Borra et al., 2024). Unlike existing methods that rely on filtering to discard noisy samples (Guo et al., 2024; Zhang et al., 2024), we

Category	Text and Context Details
Case 1: Negative Alignment ($\rho < 0$) → Rectification	
Source 1	Күдіктілер - 32 жастағы ирандық және оның бауыры.
Syn BT 1	The victims - a 32-year-old Iranian and his brother
Source 2	Туыстары саяжайға келгенде ол жерде ешкім болмады.
Syn BT 2	When relatives came to the meadow, there was no one there.
Analysis	The synthetic translations contain severe factual and lexical errors. In Example 1, the back-translation incorrectly changes “suspects” [Күдіктілер] to “victims”, resulting in a semantic inversion. In Example 2, the word “dacha” [саяжайға] is mistranslated as “meadow”, producing an incorrect meaning. These problematic synthetic signals are accompanied by negative gradient similarity scores, indicating strong disagreement between synthetic and authentic optimization directions. In such cases, GRAD suppresses the conflicting synthetic update through gradient rectification.
Case 2: Positive Alignment ($\rho > 0$) → Enrichment	
Source 3	Бұл жоба аймақтың дамуына жаңа серпін береді.
Syn BT 3	This project will give a new impetus to the development of the region.
Analysis	The synthetic translation is fluent and semantically accurate, while preserving the formal style of the source sentence. In this case, the synthetic and authentic gradients exhibit positive alignment, allowing GRAD to retain the beneficial optimization signal.

Table 8: Case study of synthetic variations captured by GRAD on the Kazakh→English task. Red text highlights hallucinated or incorrect synthetic translations associated with negative gradient alignment.

propose to leverage synthetic data by explicitly rectifying its semantic deviations, enabling the model to benefit from synthetic fluency while adhering to authentic constraints.

Gradient Manipulation. Training on heterogeneous data sources resembles Multi-Task Learning (MTL), where conflicted gradients often lead to negative transfer (Ruder, 2017). Gradient surgery techniques, such as PCGrad (Yu et al., 2020), GradVac (Wang et al., 2020), and others (Zhu et al., 2025; Limarenko et al., 2026), mitigate this by projecting gradients to resolve conflicts. However, these methods operate under a **symmetric assumption**, treating tasks equivalently. In contrast, our **GRAD** is fundamentally **asymmetric**: we treat the parallel data as a strict “Authentic Anchor”. We project synthetic gradients only when they conflict with the anchor, making our approach uniquely

suited for aligning noisy synthetic data with high-quality authentic baselines.

6 Conclusion

We present **GRADient Rectification based on Authentic Data (GRAD)**, an asymmetric gradient-level optimization framework for low-resource neural machine translation with back-translation. Rather than treating synthetic and authentic supervision as equally reliable, GRAD uses authentic parallel data as an **Authentic Anchor** to dynamically rectify conflicting synthetic updates through online orthogonal projection, while preserving non-conflicting components. Furthermore, we leverage the gradients of multiple authentic batches for the comparison with the gradient of the synthetic batch, and use the most similar one to reduce the potential information loss. Extensive experiments on six low-resource language pairs demonstrate that GRAD consistently outperforms supervised fine-tuning, back-translation, and multi-stage training baselines, and generalizes across three LLM families (Qwen, Gemma, Llama) and model scales (4B/8B). Notably, GRAD attains the best COMET on the lowest-resource En-Kk pair (**83.65**). Further analysis reveals that synthetic-induced gradient conflicts are highly localized in the parameter space, and that fine-grained parameter-wise rectification substantially outperforms global projection strategies. More broadly, our results suggest that gradient geometry can serve as a general optimization principle for coordinating heterogeneous supervision sources of asymmetric reliability. We suggest that this perspective may benefit other synthetic-data-driven domains, such as code generation and mathematical reasoning, where high-quality and noisy supervision signals often coexist.

Limitations

Our experiments primarily focus on English-to-X directions in low-resource settings. While these scenarios are practically important, the effectiveness of GRAD in high-resource regimes or many-to-many translation settings remains under explored.

Acknowledgments

We thank anonymous reviewers for their insightful comments. Zhuofei Liang and Hongfei Xu are partially supported by the National Natural Science Foundation of China (Grant No. 62306284),

China Postdoctoral Science Foundation (Grant No. 2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386). Xinglin Lyu acknowledges the support of the China Postdoctoral Science Foundation (Grant No. GZC20260877).

References

- Sina Ahmadi, Razhan Hameed, and Rico Sennrich. 2025. [Literary translations and synthetic data for machine translation of low-resourced Middle Eastern languages](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 110–118, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Arwa Arif. 2025. [The saturation point of backtranslation in high quality low resource english gujarati machine translation](#). *Preprint*, arXiv:2506.21566.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenhao Li, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.
- Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. [MALTO at SemEval-2024 task 6: Leveraging synthetic data for LLM hallucination detection](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684, Mexico City, Mexico. 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, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. [Multilingual machine translation with open large language models at practical scale: An empirical study](#). 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 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. [Scaling low-resource MT via synthetic data generation with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. [The unreasonable effectiveness of few-shot learning for machine translation](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.
- 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.
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. [Universal neural machine translation for extremely low resource languages](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. [The curious decline of linguistic diversity: Training language models on synthetic text](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. [Survey of low-resource machine translation](#). *Computational Linguistics*, 48(3):673–732.
- Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ranzato. 2020. [Revisiting self-training for neural sequence generation](#). In *International Conference on Learning Representations*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O’Gara, Hua Xu, Brian Tse, and 7 others. 2025. [Ai alignment: A comprehensive survey](#). *Preprint*, arXiv:2310.19852.
- Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. [Alternated training with synthetic and authentic data for neural machine translation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834, Online. Association for Computational Linguistics.
- Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. [Is chatgpt a good translator? yes with gpt-4 as the engine](#). *Preprint*, arXiv:2301.08745.
- Philipp Koehn and Rebecca Knowles. 2017. [Six challenges for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.
- Pengfei Li, Liangyou Li, Meng Zhang, Minghao Wu, and Qun Liu. 2022. [Universal conditional masked language pre-training for neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6379–6391, Dublin, Ireland. Association for Computational Linguistics.
- Evgeny Alves Limarenko, Anastasiia Studenikina, Svetlana Illarionova, and Maxim Sharaev. 2026. [Gcond: Gradient conflict resolution via accumulation-based stabilization for large-scale multi-task learning](#). *IEEE Access*, 14:42086–42104.
- Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. [Pre-training multilingual neural machine translation by leveraging alignment information](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.

- Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. [LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, 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.
- Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. [Understanding in-context machine translation for low-resource languages: A case study on Manchu](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- 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, 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.
- Sebastian Ruder. 2017. [An overview of multi-task learning in deep neural networks](#). *Preprint*, arXiv:1706.05098.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvenc, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- 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.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- Sifan Wang, Yujun Teng, and Paris Perdikaris. 2020. [Understanding and mitigating gradient pathologies in physics-informed neural networks](#). *Preprint*, arXiv:2001.04536.
- 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](#). *arXiv preprint arXiv:2505.09388*.
- Wayne Yang and Garrett Nicolai. 2023. [Neural machine translation data generation and augmentation using chatgpt](#). *Preprint*, arXiv:2307.05779.
- Bilge Ka an Yazar, Durmu   zkan  ahin, and Erdal Kili . 2023. [Low-resource neural machine translation: A systematic literature review](#). *IEEE Access*, 11:131775–131813.
- Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. [Gradient surgery for multi-task learning](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc.
- Armel Randy Zebaze, Beno t Sagot, and Rachel Bawden. 2025. [TopXGen: Topic-diverse parallel data generation for low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22358–22381, Suzhou, China. Association for Computational Linguistics.
- Dongxu Zhang, Varun Gangal, Barrett Lattimer, and Yi Yang. 2024. [Enhancing hallucination detection through perturbation-based synthetic data generation in system responses](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13321–13332, Bangkok, Thailand. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-mt technical report](#). *Preprint*, arXiv:2509.05209.

Shijie Zhu, Hui Zhao, Tianshu Wu, Pengjie Wang, Hongbo Deng, Jian Xu, and Bo Zheng. 2025. [Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning](#). In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25*, page 204–212, New York, NY, USA. Association for Computing Machinery.