G-DIG: Towards Gradient-based DIverse and hiGh-quality Instruction Data Selection for Machine Translation

Xingyuan Pan[†] **Luyang Huang** **Liyan Kang**[†] **Zhicheng Liu**
Yu Lu **Shanbo Cheng**[‡]
ByteDance Research
xypan00@gmail.com {huangluyang, chengshanbo}@bytedance.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable abilities in general scenarios. Instruction finetuning empowers them to align with humans in various tasks. Nevertheless, the *Diversity* and *Quality* of the instruction data remain two main challenges for instruction finetuning. With regard to this, in this paper, we propose a novel gradient-based method to automatically select high-quality and diverse instruction finetuning data for machine translation. Our key innovation centers around analyzing how individual training examples influence the model during training. Specifically, we select training examples that exert beneficial influences on the model as high-quality ones by means of Influence Function plus a small high-quality seed dataset. Moreover, to enhance the diversity of the training data we maximize the variety of influences they have on the model by clustering on their gradients and resampling. Extensive experiments on WMT22 and FLORES translation tasks demonstrate the superiority of our methods, and in-depth analysis further validates their effectiveness and generalization.[1]

1 Introduction

Large Language Models (LLM) have revolutionized the field of Natural Language Processing with their strong abilities in general language understanding and generation (OpenAI, 2023; Achiam et al., 2023). To enable this strong ability, instruction finetuning has been proposed to better align language models (Wei et al., 2021; Chung et al., 2022; Ouyang et al., 2022). Significant progress has been made on collecting extensive instruction finetuning data for better aligning LLMs to produce helpful responses (Chung et al., 2022).

We argue that *Diversity* and *Quality* of the instruction data present a pair of challenges for instruction finetuning. Zhou et al. (2023) have demonstrated that a model trained with a limited, carefully curated dataset composed of high-quality and diverse examples outperforms models trained on larger, more extensive datasets during instruction finetuning. Subsequently, various methods have been proposed to automatically select high-quality and diverse training data from the large pool of instruction finetuning datasets (Chen et al., 2023a; Cao et al., 2023). Yet, these methods often rely on another model to decide quality or diversity, neglecting the inherent model behavior and strong ability of the LLM itself.

To this end, we propose *G-DIG*, a novel Gradient-based method to automatically select Diverse and hiGh-quality instruction finetuning data for machine translation. We use influence function (Koh and Liang, 2017), a gradient-based method that quantifies the impact made by individual training samples. Concretely, we (1) measure the response quality of each training sample with the influence score of the training sample on test instances and (2) enhance the diversity of the training data by maximizing the variety of influences they have on the model.

Specifically, we hypothesize that high-quality data should have positive influences on high-quality test samples. Hence, we first manually create a small set of high-quality seed data and then automatically select high-quality data that have positive influences on seed data. Moreover, we utilize K-means clustering algorithms to cluster training data with similar influences, using gradients representing their influences on the model. Unlike existing methods that introduce an external model to decide quality and diversity, our methods directly use model gradients, which capture the model behavior through learning algorithms and back to the training data.

[†]The work was done when the author was an intern at ByteDance.

[‡]Corresponding author.

[1]Code is available at https://github.com/xypan0/G-DIG

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15395–15406
August 11-16, 2024

We conduct experiments on Zh ⇒ En and De ⇒ En translation tasks. Specially, We collect large candidate pools and manually construct two small sets of seed data. We then finetune different LLM backbones on various sizes of selected subsets and compare their performances with different selective methods and existing SOTA LLMs. Under a thorough comparison in a range from 1k to 64k selected samples, our proposed method not only surpasses the baseline selective methods but also achieves competitive performance against SOTA models. Extensive experiments and in-depth analysis emphasize the need for data selection and demonstrate the effectiveness and generalization of our proposed methods.

2 Related Work

LLM for Machine Translation. Due to their strong in-context learning and instruction-following abilities, powerful LLMs like GPT-4 have achieved remarkable progress in machine translation, with comparable performance to the top systems on the WMT translation task (Zhu et al., 2023; He et al., 2023; Raunak et al., 2023). To fully leverage LLMs' translation ability, various methods have been proposed, including in-context translation exemplar selection (Garcia et al., 2023; Lin et al., 2022; Zhang et al., 2023a; Agrawal et al., 2022), prompt optimization (Jiao et al., 2023) and decoding strategies (Zeng et al., 2023a).

The aforementioned studies all focus on the inference stage optimization, while another line of work focuses on instruction tuning the LLMs for better translation performance. Xu et al. (2023) proposes to first finetune the model on monolingual data and then on high-quality parallel data. Li et al. (2024b) trains the model to elicit translation ability by multilingual instruction tuning. Li et al. (2024a) proposes to create high-quality instruction-tuning data from larger models by a patching mechanism. Chen et al. (2023b) improves the model instruction understanding by adding a global instruction representation and improves model faithfulness by comparing over-translation and misstranslation results with the correct translation. Zeng et al. (2023b) proposes a novel framework using examples in comparison to teach LLMs to learn translation. However, all these methods neglect the importance of instruction finetuning data quality and diversity in machine translation. And in this paper, we propose a novel approach to enhance the quality and diversity of translation data.

Traning Data Quality and Diversity. Various studies evident that the quality and diversity of instruction finetuning data predominate the performance of LLMs (Zhou et al., 2023; Touvron et al., 2023; Li et al., 2023). For example, Zhou et al. (2023) manually curated a small, high-quality instruction set to elevate the model's instruction following power. Although the methods in (Zhou et al., 2023) rely heavily on human effort, they motivate research aiming to automatically obtain high quality instructions. Cao et al. (2023) propose to score the quality of each instruction by combining several language indicators using a linear model. However, all these indicators rely on external models. Besides, Du et al. (2023) present a comprehensive approach for selecting high quality and diverse instruction based on reward model score and semantic diversity. Still, their methods rely heavily on external models and overlook the direct impact that finetuning data has on the model. Remarkably, Li et al. (2023) propose a self-guided approach to select difficult instructions with the guidance of the LLM itself. Admittedly, their methods are free of any external models. However, they select training examples more associated with necessity and complexity than quality, and they overlook the importance of finetuning data diversity. Despite the efforts these methods have made to automate instruction selection, they either overlook the importance of the quality and diversity of training data or rely on external models for judging. Significantly different from them, our methods take both quality and diversity into consideration and are free of external models.

Gradient-based Data Selection. Data selection with influence function has been widely studied in NLP. Lam et al. (2022) propose to identify erroneous training data by using synthetic noisy data, showing that vanilla influence functions are not sufficient for good retrieval performance. On the contrary, we select high-quality and diverse finetuning data with the aid of gradient information. Also, we show that our use of influence function is capable of selecting high-quality data. Akyurek et al. (2022) demonstrate the potential of using gradient information to trace factual knowledge in language models back to the training data. Nevertheless, their practical applications remain understudied. Methods for scaling up influence function (Schioppa et al., 2022) and explaining black

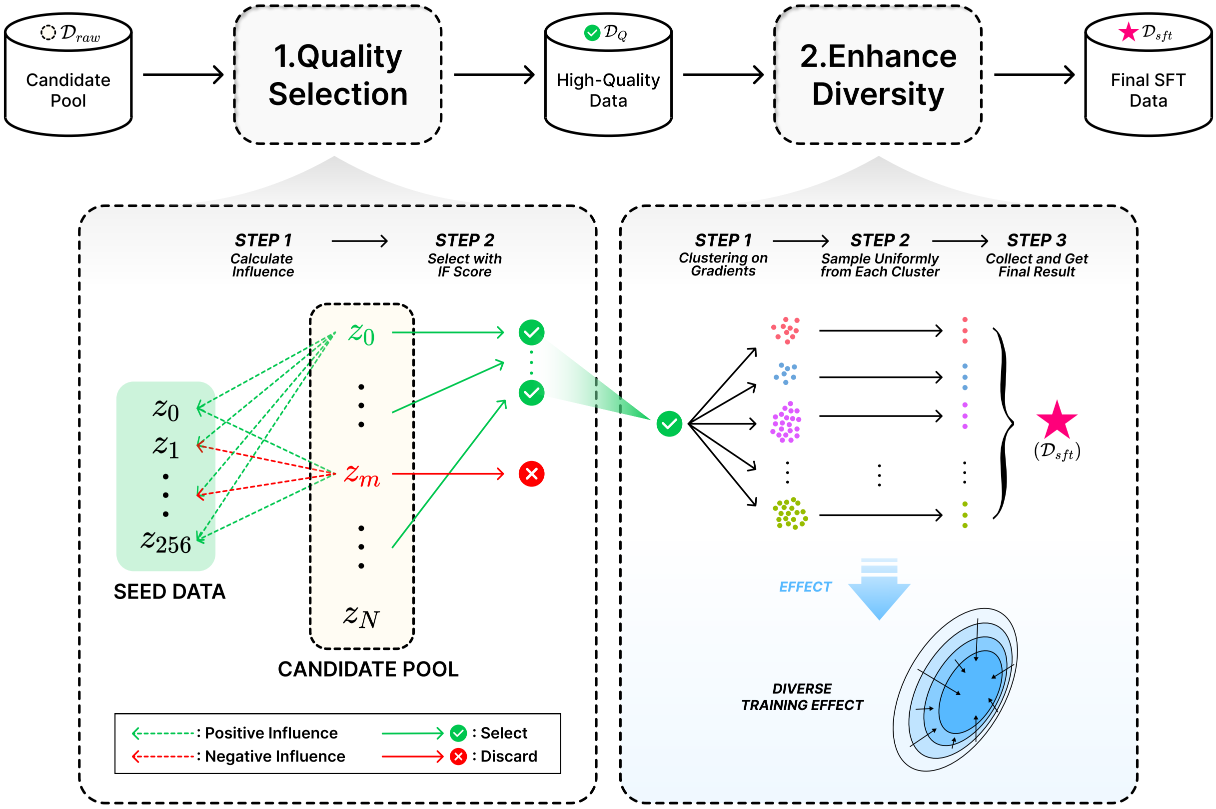

Figure 1: Overview of our proposed method. Our overall method consists of two components: (1) high-quality data selection and (2) enhancing their diversity. In high-quality data selection, we calculate the pair-wise influence (dashed arrows) of the candidates on seed data. Then we select those with all positive influences (as marked green). Afterwards, we cluster on the selected high-quality data to distinguish dissimilar influences (as marked in dots with different colors) and resample to further obtain high-quality and diverse finetuning data.

box predictions of NLP models have been proposed (Han et al., 2020). Remarkably, the concurrent work (Xia et al., 2024) proposes a similar method to select task-specific LLM finetuning data by estimating influences for training data. Different from them, we select high-quality finetuning data for machine translation. And we show that our method is capable of capturing a higher-level concept (i.e., the quality of training data).

3 Methods

In this section, we describe our gradient-based method to select high-quality and diverse training data for instruction finetuning, as displayed in Figure 1. Our methods consist of two parts: 1) high-quality data selection with the influence function (§ 3.1) and 2) diverse data selection with gradient-based clustering (§ 3.2). For high-quality data selection, importantly, we utilize the influence function to quantify the impact of individual training on the test sample. For diverse data selection, we use gradient distance to assess the overall diversity of the instruction training data.

3.1 High-quality Data Selection

In this section, we detail our approach for selecting high-quality training data for machine translation. Intuitively, if a training example significantly benefits the model to generate high-quality outputs, it is likely to be of high quality itself. Consequently, we first manually curate a small set of high-quality translation data that we refer to as the *seed data* to form a criteria for evaluating training data quality. Then we select training data that aids the model in generating high-quality seed data.

Concretely, we employ Influence Function (IF) (Koh and Liang, 2017) to quantify how a training example z_m influences the model's behavior on a test example z_t. In our influence function setting, we start with the following finetuning objective:

$$\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L(z_i|\boldsymbol{\theta}), \qquad (1)$$

where $\boldsymbol{\theta}$ is the model parameter, $z_i = (x_i, y_i)$ is the i-th prompt-response pair and the loss L is simply the language modeling loss of the response solely:

$$L(x, y) = -\log P(y|x) = -\log \prod_{j=1}^{T} p(y^j | x, y^{<j}), \tag{2}$$

where x^j denotes the j-th token of x. Then influence function calculates the influence of z_m on z_t by:

$$\mathcal{I}(z_m, z_t) = -\nabla_{\boldsymbol{\theta}|\boldsymbol{\theta}^*} L(z_t)^\top \mathbf{H}_{\boldsymbol{\theta}^*}^{-1} \nabla_{\boldsymbol{\theta}|\boldsymbol{\theta}^*} L(z_m), \tag{3}$$

where $\boldsymbol{\theta}^*$ is the optimal model parameter of finetuning in (1) and $\mathbf{H}_{\boldsymbol{\theta}^*} = \nabla^2_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L(z_i | \boldsymbol{\theta}^*)$ is the Hessian of the training objective at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Though the calculation of the influence function is complex, the most significant aspect for readers is that *if the scalar $\mathcal{I}(z_m, z_t)$ is negative, training the model on z_m reduces the model's loss on z_t. In this case, z_m is considered to be helpful for the model to generate z_t.* In our implementation of IF, we use the gradients of model's Multilayer Perceptron (MLP) parameters in $\{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$-th layers to speed up calculation. And we average the gradients of each token to form a vector. Also, we use Kronecker-Factored Approximate Curvature (KFAC) (Martens and Grosse, 2015) to approximate Hessian for reducing memory consumption. We detail the derivation of IF and our modification for fitting it into LLM fining tuning in Appendix A.

Correspondingly, as depicted in Figure 1, our proposed high-quality data selection method consists of two steps: (1) calculating influences and (2) selection. We select training examples in our candidate pool $\mathcal{D}_{raw}$ that exert beneficial impact on all samples in the seed data, i.e., example $z_m \in \mathcal{D}_{raw}$ meets:

$$\forall z_t \in \mathcal{D}_{seed}, \ \mathcal{I}(z_m, z_t) < 0. \tag{4}$$

Thus, we select training data that contributes to the model's high-quality generation.

In practice, we find that the seed instruction dataset $\mathcal{D}_{seed}$ of size 256 suffices for selecting high quality instructions. Hence, we set the size of $\mathcal{D}_{seed}$ to 256 in our implementation. As the focus of this paper is to select high-quality finetuning data for translation task, we randomly select 256 parallel texts from the validation set and have them revised by human translators. We select high-quality data from our candidate pool, for which we collect publicly available WMT22's parallel texts. Throughout our implementations, we use the prompt template: *Translate the following text into {trg_lang}.\n\nText:\n"{src_text}"*, where *{trg_lang}* are target languages such as "English" and *{src_text}* are the source text. And the response is simply the target text. We detail our data preparation in Section 4.

3.2 Diverse Data Selection

After obtaining high-quality translation pairs, we further ensure the diversity of the training data. To ensure coverage of different translation patterns, we propose to use gradient similarity to assess the diversity. Specifically, we consider the gradient of the response loss in equation (2) with respect to the final MLP layer and average them out on all the tokens. We utilize the Euclidean distance as the similarity measure.

To maximize the diversity of training data influences, we cluster on the gradients of training examples to obtain different patterns. Then we sample uniformly from the clustering result to ensure the diversity of training data. Moreover, we employ *K-means* as the clustering algorithm due to its ability to process large-scale datasets. Furthermore, to speed up and reduce memory, we use *Random Projection* as the dimensionality reduction method to reduce the dimension of the gradients to 400 (Bingham and Mannila, 2001). In practice, we cluster the training data into 512 clusters.

4 Experiment Setup

Datasets. We conduct experiments on Zh ⇒ En and De ⇒ En tasks. We collect separate candidate pools for different translation directions. Specifically, our Zh ⇒ En pool is composed of 1.9 million examples sampled from WMT22's ParaCrawl v9, News Commentary v16, UN Parallel Corpus V1.0, WikiMatrix and Back-translated news and our De ⇒ En pool contains 256k examples sampled from WMT22's Europarl v10, ParaCrawl v9, News Commentary v16 and WikiMatrix. We split 512 examples to validation sets for evaluation. We test our methods on the latest WMT22 test sets from the news translation track of WMT22 competition[2] and Flores-101 dev-test split (Goyal et al., 2022). The WMT22 Zh ⇒ En and De ⇒ En test sets contain 1875 and 1984 samples, respectively.

[2]https://github.com/wmt-conference/wmt22-news-systems

Figure 2: The comparison results of model trained on various amounts of data selected by our G-DIG, G-DIG w/o Diversity and Random selection on Zh ⇒ En and De ⇒ En translations. We plot the results on Zh ⇒ En and De ⇒ En translations in the left and right two columns respectively.

And the Flores-101 dev-test split contains 1012 samples for each Zh ⇒ En and De ⇒ En.

Implementation Details. We use Baichuan2-7B (Baichuan, 2023) for Zh ⇒ En and Llama2-7B for De ⇒ En. For all finetuning experiments, we adopt the same setting. The finetuning process lasts for 3 epochs with an initial learning rate of $1e-5$ and a global batch size of 64. The instruction template we use is *Translate the following text into {trg_lang}.*\n\n*Text:*\n*"{src_text}"*, where *{trg_lang}* are target languages such as "English" and *{src_text}* are the source text. The models are evaluated every 10 steps. We use the checkpoint with the smallest loss on the valid set for the final test. During the inference, we use beam search as the decoding strategy with a beam size of 4. The training and inference of 7B size models are conducted on 16 NVIDIA A100 80GB GPUs with DeepSpeed ZeRO-3 Offload.

Evaluation. For automatic evaluation, we use the widely used metrics BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020). We use ScareBLEU[3], BLEURT-20 (Pu et al., 2021) and Unbabel/wmt22-comet-da[4] in the evaluation implementations.

[3]https://github.com/mjpost/sacrebleu
[4]https://huggingface.co/Unbabel/wmt22-comet-da

Baselines and Comparisons. In order to demonstrate the superiority and effectiveness of our methods, we compare our G-DIG to validate the performance of our overall approach. Also, we assess its variant G-DIG w/o Diversity (G-DIG without enhancing the data diversity) to solely investigate our high-quality data selection module. For comparisons, we consider the following baselines:

- Bayling-13B (Zhang et al., 2023b), an English / Chinese LLM based on Llama with superior translation capabilities.
- BigTranslate-13B (Yang et al., 2023), a multilingual LLM based on Llama with the capability of translating over 100 natural languages.
- TIM (Zeng et al., 2023b). We present the results of BLOOMZ-7b-mt and Llama2-7b trained with TIM-Full-LM-based Bad Output for WMT22 test sets and FLORES respectively.
- Model trained on the random subset. To emphasize the need for instruction selection, we choose k random instructions to form a finetuning subset. We finetune the LLM on the selected subsets and report their performances.
- Model trained on the reward subset. We also compare our method with the existing

Model	Zh ⇒ En						De ⇒ En					
	WMT22			FLORES			WMT22			FLORES		
	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT
SOTA Models												
Bayling-13B	77.72	20.12	-	-	-	-	**83.02**	**27.34**	-	-	-	-
BigTranslate-13B	74.26	14.16	-	-	-	-	80.68	23.35	-	-	-	-
TIM-7B	**79.33**	**23.81**	-	**85.81**	**26.25**		78.19	25.43	-	**88.82**	**41.96**	-
NLLB-54B	70.70	16.56	-	-	-	-	78.94	26.89	-	-	-	-
Baseline Models												
Random	75.55	14.31	61.35	85.63	24.25	73.84	82.58	27.60	70.57	87.80	36.89	77.52
Reward	77.90	16.29	64.64	86.16	24.48	74.73	**83.28**	27.52	**71.51**	88.37	38.69	78.62
Ours												
G-DIG	78.29	**20.63**	64.86	86.18	25.04	74.75	**83.28**	**28.99**	71.25	88.59	**39.97**	78.79
G-DIG w/o Diversity	**78.43**	20.35	**65.08**	**86.45**	**25.81**	**75.19**	82.93	28.71	70.96	**88.68**	**39.97**	**78.83**

Table 1: In this table, we present the comparison results of our methods with various baselines in accordance with Section 4. We directly adopt the results from the original paper and omit the missing metrics. We report the results of our G-DIG and G-DIG w/o Diversity. The Best result in each group is in **bold**. The Best result in each column is in red.

selection method. We use the commonly used reward model-based method for selecting high-quality training data (Du et al., 2023; Cao et al., 2023). Specifically, we follow (Du et al., 2023) to use the reward-model-debertav3-large-v2[5] to score each instruction and select the top k instructions with the highest score to form the finetuning subset.

5 Experimental Results

5.1 Main Results

In this section, we present our main experimental findings. We start with comparing our methods with baselines on various amounts of training data in Figure 2. Then we compare our best results with SOTA models and baselines in Table 1. Furthermore, we conduct human evaluation and present the results in Table 2. We show that our approach is superior in terms of effectiveness and robustness.

Our Methods Improve the Translation Performance Across Various Amount of Training Data. In order to demonstrate the scalability of our method in terms of the amount of selected training data, we present the results as a function of the amount of training data from 1000 to 64,000 in Figure 2 for Zh ⇒ En and De ⇒ En translation. Notably, our G-DIG model consistently surpasses the random model across *all* metrics and dataset sizes for Zh ⇒ En translation. Also, for De ⇒ En translation, our G-DIG outperforms the random baseline in almost all cases. Statistical analysis results in Appendix C further suggest our methods excel the random baseline. These results demonstrate the efficacy and robustness of our methods in terms of the amount of selected data. Meanwhile, we also observe that the quality and diversity of finetuning data dominate the performance of LLM. Specifically, for both Zh ⇒ En and De ⇒ En translations we can see that models trained on 1000 examples selected by our G-DIG outperform the models trained with 64k randomly selected examples.

Model	Score	Win ↑	Tie	Lose ↓
Zh ⇒ En				
w/ Random	3.59	—	—	—
w/ Ours	3.92	35.0%	53.0%	12.0%
De ⇒ En				
w/ Random	3.92	—	—	—
w/ Ours	4.21	34.0%	53.0%	13.0%

Table 2: Human evaluation results on randomly sampled sets. “Win”/“Tie”/“Lose” stands for the percentage of translations where ours is better than, tied with, or worse than the random subset.

Our Methods Surpass Baselines and Achieve Comparable Results with SOTA Models. To see how our method performs compared with baselines and SOTA models, in Table 1 we present the

[5] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2

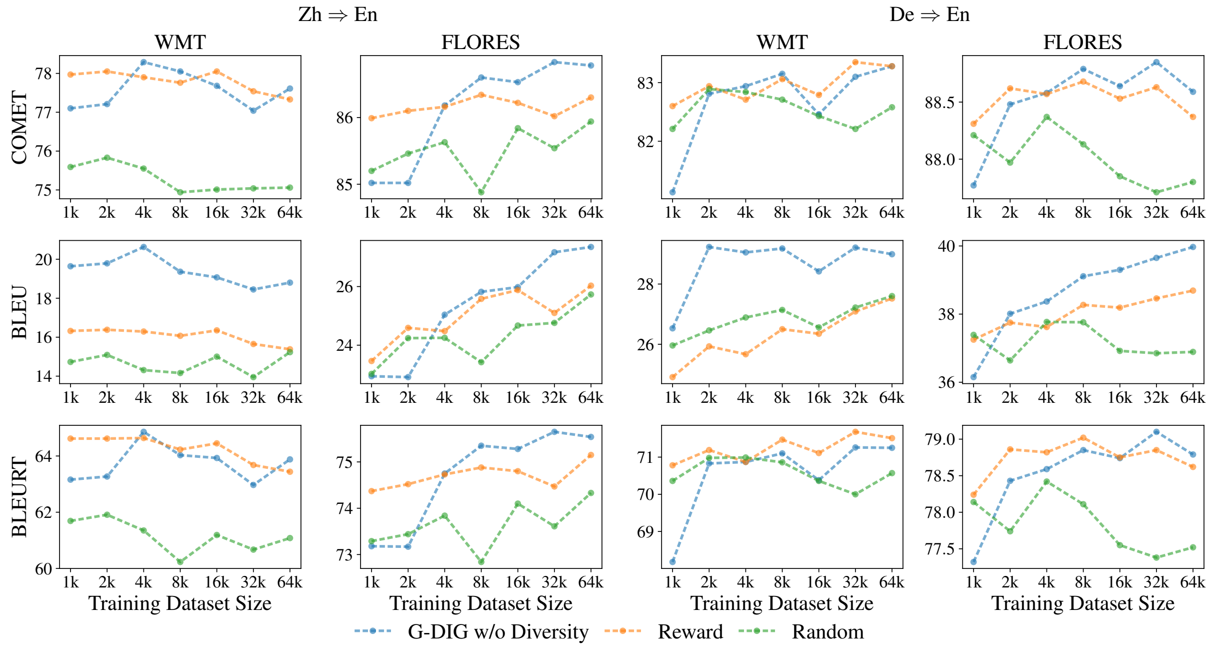

Figure 3: The comparison results of model trained on various amounts of data selected by our G-DIG w/o Diversity, Reward model selection and Random selection on Zh ⇒ En and De ⇒ En translations. We plot the results on Zh ⇒ En and De ⇒ En translations in the left and right two columns respectively.

results of SOTA models, our best and corresponding baselines. Specifically, our G-DIG achieves its best results at 4000 training examples for Zh ⇒ En translation and 64k for De ⇒ En translation. For baselines comparisons, we observe that our methods surpass baselines in almost all cases, demonstrating the effectiveness of our approach. For SOTA models comparisons, our 7B models achieve comparable results with TIM-7B and even better results compared with Bayling-13B, BigTranslate-13B and NLLB-54B.

Our Methods Align the Model Better Compared with Random Baseline. We further conduct human evaluation to analyze the translation quality. We respectively randomly pick 100 sentences from Zh ⇒ En and De ⇒ En test sets, and recruit three human judges for evaluation. For each sentence, the judges read the source sentence and two candidate translations, which are from the random subset model and G-DIG subset model. The judges are required to rate each candidate on a scale of 1 to 5 and pick the better one.

From Table 2, we can see our methods enable the model to translate better with a higher average score in both Zh ⇒ En and De ⇒ En translations. Also, our G-DIG subset model is more frequently rated as better translation than the random subset model on both Zh ⇒ En and De ⇒ En translations, indicating our methods better align the model than the random selection method.

The Fewer the Instructions, the Greater the Significance of Diversity. To see the role that diversity plays during finetuning, we compare the results of G-DIG and G-DIG w/o Diversity on Zh ⇒ En and De ⇒ En translation in Figure 2. Remarkably, our diversity enhancement makes significant advancements in enhancing the translation performance when there are only few training data provided ($k \leq 2000$). For Zh ⇒ En translation with 1000 training examples, our G-DIG further improves G-DIG w/o Diversity by 1.7 in terms of COMET on WMT and by 2.17 in terms of BLEU on FLORES. In addition, for De ⇒ En translation with 1000 training examples, our G-DIG improves the BLEU by 2.11 and 1.24 on WMT and FLORES respectively compared with G-DIG w/o Diversity. However, as the amount of instructions increases, this effect fades away, since large instruction sets are already coupled with high diversity. As shown in Figure 2, the G-DIG curves almost coincide with the G-DIG w/o Diversity curves in all metrics when the amount of instructions goes beyond 4000 and 8000 for WMT and FLORES, respectively.

Zh ⇒ En		De ⇒ En	
Source Text	Target Text	Source Text	Target Text
Ours			
例如，如果在大部分时间里价格在1,200美元到1,800美元之间，这就是常见值域。	For example, if the price is between $1,200 and $1,800 most of the time, this is the common range.	Der ganze MDF Italia Tense Tisch wird mit einer 3mm dünnen Acrylharzfolie und Steinmineralien in Weiß verkleidet.	The entire MDF Italia Tense Table is coated with a 3 mm thin acrylic resin film and stone minerals in white.
根据武器法第6条，若无司法部长颁发许可证，禁止出口(任何种类)武器和武器设备。	According to the Weapons Act, Section 6, it is prohibited without a license from the Minister of Justice to export weapons (of any kind) and war equipment.	X850 Infrarot-Induktions-Funk-Türklingel w / blaue Anzeige - Weiß (3 x AAA + 3 x AA)	X850 Infrared Induction Wireless Doorbell w/ Blue Indicator - White (3 x AAA + 3 x AA)
Random			
在对她的行为进行调查之前，西拉德福德议员被剥夺了议会的鞭打并被禁止参加政党活动。	The West Bradford MP was stripped of the parliamentary whip and barred from party activity pending an investigation of her behaviour - which David Cameron branded racist.	Ebenso wichtig ist das Engagement in interdisziplinären und internationalen Forschungszweigen wie der Critical Psychology, den Cultural und Postcolonial Studies, den Gender wie auch den Religious Studies.	In this regard, sociology, anthropology, history, education, and philosophy are considered as important as work in research fields like critical psychology, cultural and postcolonial studies, gender studies, and religious studies.
我们提供2134 个酒店在西安	Hotels in Xi'an (Shaanxi)	Hin- und Rückflüge von Lagos nach Madrid	Return flights from Lagos to Madrid

Table 3: Examples of ours and randomly selected training data in the form of parallel texts.

5.2 Analysis

In this section, we conduct ablation studies on our high-quality data selection module and our diversity enhancement module. Specifically, to see the superiority of our high-quality data selection module, we compare G-DIG w/o Diversity with the reward model-based baseline in Figure 3. In order to demonstrate the advantage of our gradient-based diverse data selection module, we compare our G-DIG with its embedding-based counterparts in Figure 4.

Our High-quality Data Selection Module Surpasses the Reward Model-based Method In Figure 3, we show that our high-quality data selection module G-DIG w/o Diversity achieves superior results compared with the existing reward model-based method in Figure 3. Specifically, for Zh ⇒ En translation our G-DIG achieves overall better results compared with the reward model-based method, falling behind in only few cases. For De ⇒ En translation, it is hard to distinguish between ours and reward models-based baselines in COMET and BLEURT. However, our G-DIG w/o Diversity significantly outperforms the reward model in terms of BLEU on both WMT22 and FLORES.

Our Diversity Enhancement Module Improves the Training Data Diversity. In this part, we substantiate the superiority of our gradient-based diversity enhancement module. We compare our G-DIG

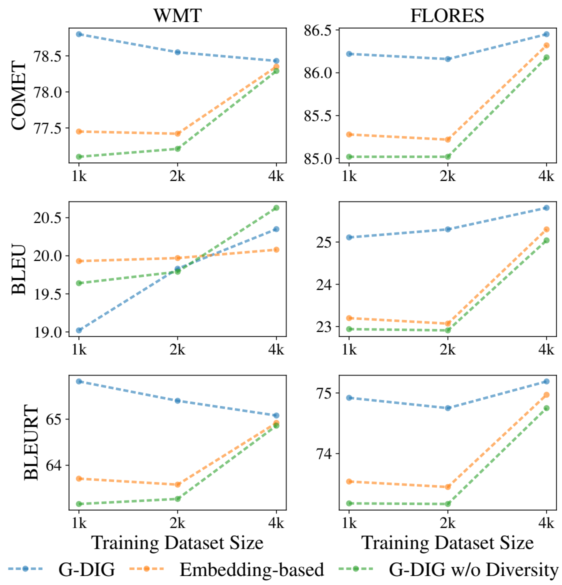

Figure 4: The comparisons between our G-DIG,G-DIG w/o Diversity and embedding-based methods on various amounts of training data on Zh ⇒ En translation.

with its embedding-based counterpart. Specifically, embedding-based methods measure the similarity of training examples based on embeddings. Their objective is to maximize the semantic richness of the training data. We follow the model used in (Du et al., 2023) for diversifying the training data. We use BERT (Devlin et al., 2018) to extract embeddings. Then we run *Kmeans* to cluster these embeddings and use our sampling procedure to enhance the training data diversity. As illustrated in Figure 4, our proposed method outperforms its embedding-based counterpart in almost all cases, which validates the advantage of our gradient-based method. For instance, with 1000 instructions provided, our methods surpass the embedding-based method by 2.11 of BLEURT and 1.91 of BLEU on WMT and FLORES respectively. Moreover, in almost all scenarios, the curves of embedding-based method coincide with G-DIG w/o Diversity curves with trivial improvements, indicating that the existing methods that enhance semantic coverage contribute little to the finetuning data diversity.

Our Methods Select Highly Parallel Texts. In this part, we showcase our selected high-quality data and randomly selected data in the form of parallel texts in Table 3. Remarkably, not only is our selected data accurate and coherent in the target text space, but it is also natural and of the correct format and grammar in terms of the source text.

5.3 Discussion on Hyperparameters

There are two primary hyperparameters in our methods: the number of seed data $|\mathcal{D}_{seed}|$ in high-quality data selection and the number of clusters for K-means in diverse data selection. Intuitively, the more seed data we use, the more strict the criteria in (4) would be. Therefore, increasing the number of seed data improves the quality of selected data. And using less seed data can be regarded as a relaxation to the criteria in (4). We experiment with seed dataset sizes of 128, 256, and 512 and note that further enlarging the seed dataset contributes little to improving the quality of selected seed data. Considering the high cost of obtaining human-annotated high-quality data, we ultimately use 256 seed data after balancing annotation costs and model performance benefits.

The reasonable choice of the number of clusters in diverse data selection is crucial for our methods. Since our candidate pool is quite large, we cluster the data into 512 clusters to ensure the variety of training data.

6 Conclusion

In this paper, we propose G-DIG, a novel gradient-based method for selecting high-quality and diverse LLM finetuning data for machine translation. Specifically, using Influence Function and a seed dataset we select high-quality training data that have beneficial influence on the model. Furthermore, we enhance the training data diversity by clustering on their gradients and resampling. Extensive experiments prove that our methods improve the LLM in terms of translation ability. Also, human evaluation results demonstrate that our methods better align the LLM compared with the baseline model. We hope this work facilitates the research on LLM finetuning data selection.

Limitations

Computational Cost for Influence Function. In this paper, we utilize Influence Function to measure the influence of a training sample on a test sample's prediction. However, the computational cost for calculating this influence can be large for LLMs. The asymptotic computational complexity for calculating Hessian is $\mathcal{O}(P^2)$ where P is the number of model parameters. And the computational complexity for calculating pair-wise influence between the seed data and candidate pool is $\mathcal{O}(MNP)$ where M and N denote the number of seed data and candidates, respectively. How to reduce the computational cost of influence function still remains a challenge. We leave this for future work.

Ethical Considerations

All the data sources are publicly available and do not involve privacy issues. For all human evaluations mentioned in the paper, we hire full-time professional translators and pay them with market wage. All of our translators are graduates.

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*.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Ella Bingham and Heikki Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: When data mining meets large language model finetuning. *arXiv preprint arXiv:2311.15653*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023a. Alpagasus: Training a better alpaca with fewer data.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023b. Improving translation faithfulness of large language models via augmenting instructions. *arXiv preprint arXiv:2308.12674*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

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 *International Conference on Machine Learning*, pages 10867–10878. PMLR.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Steven George Krantz and Harold R Parks. 2002. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media.

Tsz Kin Lam, Eva Hasler, and Felix Hieber. 2022. Analyzing the use of influence functions for instance-specific data filtering in neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 295–309. Association for Computational Linguistics.

Jiahuan Li, Shanbo Cheng, Shujian Huang, and Jiajun Chen. 2024a. Mt-patcher: Selective and extendable knowledge distillation from large language models for machine translation. *arXiv preprint arXiv:2403.09522*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024b. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

James Martens and Roger Grosse. 2015. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR.

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *arXiv preprint arXiv:2305.14878*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. 2022. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8179–8186.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Stefano Teso, Andrea Bontempelli, Fausto Giunchiglia, and Andrea Passerini. 2021. Interactive label cleaning with example-based explanations. *Advances in Neural Information Processing Systems*, 34:12966–12977.

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*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

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. *arXiv preprint arXiv:2309.11674*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023a. Improving machine translation with large language models: A preliminary study with cooperative decoding. *arXiv preprint arXiv:2311.02851*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023b. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023b. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

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 Influence Function.

Influence Function (IF) was introduced to deep learning by (Koh and Liang, 2017). In the classical influence function setting, we are given the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where x_i and y_i are the input and label of the i-th training example. And the model parameter θ^* is obtained through empirical risk minimization:

$$\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) \tag{5}$$

$$= \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L(z_i|\boldsymbol{\theta}), \tag{6}$$

where $z_i = (x_i, y_i)$ denotes the i-th input-label pair and L is the loss function, e.g., cross entropy loss. Influence function measures the impact of a training example z_m to the response function by up-weighting z_m by ε in the training objective:

$$\boldsymbol{\theta}^*(\varepsilon) = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} L(z_i|\boldsymbol{\theta}) + \varepsilon L(z_m). \tag{7}$$

The influence of the training example z_m on the response function $\boldsymbol{\theta}^*$ is the derivative of $\boldsymbol{\theta}^*$ with respect to ε at $\varepsilon = 0$:

$$\mathcal{I}_{\boldsymbol{\theta}^*}(z_m) = \left.\frac{\mathrm{d}\boldsymbol{\theta}^*(\varepsilon)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0}. \tag{8}$$

Metric	Zh ⇒ En						De ⇒ En					
	WMT22			FLORES			WMT22			FLORES		
	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT
p value	1.6e-7	1.6e-8	2.0e-6	7.1e-5	3.4e-3	1.6e-5	3.5e-2	1.6e-6	2.1e-1	1.7e-4	1.7e-3	1.8e-3

Table 4: In this table, we present the statical analysis outcome of our experimental results in Figure 2. We conduct t-test on our G-DIG compared with random baseline and present the p-value.

Using the Implicit Function Theorem (Krantz and Parks, 2002) and first-order Taylor approximation, we define the influence

$$\mathcal{I}_{\boldsymbol{\theta}^*}(z_m) := -\mathbf{H}_{\boldsymbol{\theta}^*}^{-1} \nabla_{\boldsymbol{\theta}|\boldsymbol{\theta}^*} L(z_m), \tag{9}$$

where $\mathbf{H}_{\boldsymbol{\theta}^*} = \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}^*)$ is the Hessian of the training objective at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. In this paper, we are interested in the influence of the training example z_m on the model behavior at the test example z_t, i.e., $L(z_t|\boldsymbol{\theta}^*)$. Using chain rule, we define the influence of z_m on the model loss at z_t:

$$\mathcal{I}(z_m, z_t) := \left.\frac{\mathrm{d}L(z_t|\boldsymbol{\theta}^*(\varepsilon))}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} \tag{10}$$

$$= -\nabla_{\boldsymbol{\theta}|\boldsymbol{\theta}^*} L(z_t)^\top \mathbf{H}_{\boldsymbol{\theta}^*}^{-1} \nabla_{\boldsymbol{\theta}|\boldsymbol{\theta}^*} L(z_m). \tag{11}$$

Since in fine tuning the model is trained with small learning rate and few steps, the model parameter at the end of training is not the minimizer of the objective in (5) and the Hessian might not be positive definite thus is not invertible. The following equivalence between non-converged model parameter and regularized empirical risk minimizer provides a method for modifying IF:

Lemma 1 *(Goodfellow et al., 2016) Assuming that the model is trained via T steps of Stochastic Gradient Descent (SGD) and the learning rate η is fixed during training. Then the model parameter at the end of training $\boldsymbol{\theta}_T$ is equal to the model parameter obtained through regularized empirical risk minimization $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ when the following two conditions are satisfied:*

$$|\boldsymbol{I} - \eta\boldsymbol{\Lambda}| \preceq \boldsymbol{I}, \tag{12}$$

$$(\boldsymbol{I} - \eta\boldsymbol{\Lambda})^T = (\boldsymbol{\Lambda} + \lambda\boldsymbol{I})^{-1}\lambda, \tag{13}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix in the eigendecomposition of $\mathbf{H}_{\boldsymbol{\theta}^*} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top$.

Hence we approximate of the Hessian of fine tuning as $\mathbf{H}_{\boldsymbol{\theta}_T} + \lambda\boldsymbol{I}$. For fine tuning, we have:

$$\mathcal{I}(z_m, z_t) = -\nabla_{\boldsymbol{\theta}} L(z_t)^\top (\mathbf{H}_{\boldsymbol{\theta}_T} + \lambda\boldsymbol{I})^{-1} \nabla_{\boldsymbol{\theta}} L(z_m). \tag{14}$$

Note that calculating the Hessian $\mathbf{H}_{\boldsymbol{\theta}_T}$ accurately is time-consuming. Therefore, we follow (Teso et al., 2021) to approximate the Hessian using empirical Fisher Information Matrix (eFIM). Also, we further use the Kronecker-Factored Approximate Curvature (KFAC) (Martens and Grosse, 2015) to reduce the memory usage. We implement the IF calculation based on *nngeometry*[6].

B Annotation Guidelines

We hire full-time translators who are fluent in both Chinese and English and translators who are fluent in both German and English. They are recruited to conduct human evaluations. The translators are shown the source sentence and two candidate translations, which are from the random subset model and G-DIG subset model. Then the translators are required to rate on the translation quality from 1 to 5 with 1 the worst and 5 the best and pick the better one. All our translators are Chinese.

C Statistical Analysis of Experimental Results

We conduct statistical analysis on our experimental results as shown in Table 4. Specifically, we use t-test to analyze the results of our G-DIG across various dataset sizes from Figure 2 compared with random baseline. Our experimental results have demonstrated statistical significance, as evidenced by the p-values less than 0.05 in almost all cases.

[6]https://github.com/tfjgeorge/nngeometry