Self-Evolution Knowledge Distillation for LLM-based Machine Translation

Yuncheng Song[♡], **Liang Ding**[ℜ], **Changtong Zan**[♠], **Shujian Huang**[♡*]
[♡]Nanjing University [ℜ]The University of Sydney
[♠]China University of Petroleum (East China)
songyuncheng@smail.nju.edu.cn, liangding.liam@gmail.com
zanct@s.upc.edu.cn, huangsj@nju.edu.cn

Abstract

Knowledge distillation (KD) has shown great promise in transferring knowledge from larger teacher models to smaller student models. However, existing KD strategies for large language models often minimize output distributions between student and teacher models indiscriminately for each token. This overlooks the imbalanced nature of tokens and their varying transfer difficulties. In response, we propose a distillation strategy called Self-Evolution KD. The core of this approach involves dynamically integrating teacher distribution and one-hot distribution of ground truth into the student distribution as prior knowledge, which promotes the distillation process. It adjusts the ratio of prior knowledge based on token learning difficulty, fully leveraging the teacher model's potential. Experimental results show our method brings an average improvement of approximately 1.4 SacreBLEU points across four translation directions in the WMT22 test sets. Further analysis indicates that the improvement comes from better knowledge transfer from teachers, confirming our hypothesis.

1 Introduction

Large language models (Achiam et al., 2023; Touvron et al., 2023, LLMs) have achieved remarkable success in generating high-quality translations (Hendy et al., 2023; Peng et al., 2023c; Jiao et al., 2023b) and other tasks (Kocoń et al., 2023; Zhong et al., 2023a; Lu et al., 2024). However, previous research indicates that LLM-based translation models (with billion-level parameters) must be several orders of magnitude larger than traditional neural machine translation systems (which typically have millions of parameters) to achieve comparable performance (Garcia et al., 2023). This high computational and deployment cost severely hinders the widespread application of LLMs in translation.

In general, a simple and effective technique to reduce high computational footprints is knowledge distillation (KD) (Hinton et al., 2015; Kim and Rush, 2016), which trains a smaller model (aka. student) under the supervision of a larger model (aka. teacher). Recent research on KD for large language models (LLMs) has shown promising results (Gu et al., 2024; Ko et al., 2024; Agarwal et al., 2024; Zhong et al., 2024; Rao et al., 2024), driven by one key factor: exploration of various divergence losses.

However, several problems are still underexplored. Their training objectives are achieved by indiscriminately minimizing the output distributions between the student and teacher model for each token. In fact, due to the token imbalance nature (Piantadosi, 2014) and the truth that different tokens contribute differently to the sentence meaning (Chen et al., 2020), adaptively reweighting the token-level loss would promote the model training, as evidenced by its effectiveness in sequence-to-sequence training (Zhang et al., 2022; Peng et al., 2023b). It motivates us to speculate that indiscriminately adopting the same distillation mode to each token might be sub-optimal. Besides, in human learning patterns, human teachers often provide human students with personal insights (aka prior knowledge) to facilitate student learning. Excellent teachers could adjust the amount of prior knowledge to fully stimulate students' potential. This pattern further supports our hypothesis that the distillation mode should be differentiated based on the student's learning status rather than adopting a uniform strategy. It also hints at the necessity of providing prior knowledge to optimize the distillation strategy and enhance student outcomes.

Therefore, a natural question arises: ***how to effectively transfer the teacher knowledge based on the student's mastering with the help of prior knowledge?*** It should be a dynamic strategy, which controls the integration of prior knowledge based

*Corresponding author

on the student's training state.

To address it, we propose a simple but effective strategy – self-evolution knowledge distillation (Self-Evolution KD) for LLMs. It mainly includes two stages: ① **Self-Question** and ② **Self-Evolution**. In **Stage 1**, we utilize the Kullback-Leibler (KL) divergence between the student distribution and the target distribution (averaged by the teacher distribution and the one-hot distribution of ground-truth) to quantify the learning difficulty. By comparing this measure with a preset threshold, we assess the student model's learning status at the token level, thereby identifying hard-to-learn and easy-to-learn tokens. This enables us to provide tailored prior knowledge for different tokens in the next stage to enhance the student model's learning. It should be noted that, although Qiu et al. (2022) has demonstrated the potential of prior knowledge in the field of computer vision, treating prior knowledge as input would cause the risk of dimensionality mismatch and significantly increase training costs. To introduce the prior knowledge in a lightweight way, we design a simple strategy — distribution adjustment. Specifically, in the **Stage 2**, if the token expresses hard-to-learn property, builds proxy distribution by smoothing the student distribution and target distribution, then used to learn target distribution, thus leading to faster convergence and superior performance. Otherwise, the proxy distribution is the student distribution. By emulating the human teaching mode, assigning prior knowledge to hard-to-learn tokens to fully leverage the target information, while omitting its integration to easy-to-learn tokens, maximizes the potential of the student model.

Empirically, we apply Self-Evolution KD to Llama series models (Touvron et al., 2023), with parameter sizes ranging from 7 to 13 billion, and evaluate our proposed method on the WMT22 test sets (En↔De and En↔Cs). The results show that Self-Evolution KD significantly achieves satisfactory gains over four competitive baselines. Further analysis suggests that Self-Evolution KD is more effective in transferring knowledge from the teacher model to the student model.

2 Related Work

2.1 Language Models for Translation

Before the era of large-scale language models, researchers had already begun leveraging language models to enhance machine translation tasks. This included using discriminative language models, such as BERT (Kenton and Toutanova, 2019), to improve representational capabilities (Zhu et al., 2020; Guo et al., 2021), designing Encoder-Decoder models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) to enhance translation quality (Liu et al., 2020), as well as various subsequent follow-up works to facilitate knowledge transferring (Liu et al., 2021; Zan et al., 2022b,a; Pan et al., 2024). With the increasing capacity of LLMs, they have already become new standards for various NLP tasks, including machine translation (Jiao et al., 2023b; Wang et al., 2023). One line of work focuses on a comprehensive evaluation of LLMs across various translation scenarios. For example, Zhu et al. (2023a); Jiao et al. (2023b); Hendy et al. (2023) assess the multilingual translation capabilities of LLMs. Hendy et al. (2023); Wang et al. (2023); Karpinska and Iyyer (2023) evaluates their performance in document-level translation, while Guerreiro et al. (2023) explores the phenomenon of hallucination. Another line is instruction tuning, such as Zhu et al. (2023b) boosted the translation capability of LLMs by translation data alongside cross-lingual general task data. Xu et al. (2023) proposed a new translation paradigm: initial fine-tuning on monolingual data, followed by a small set of high-quality parallel data. Jiao et al. (2023a) enhance the translation abilities by leveraging open-source LLMs, human-written translation and feedback data. Zan et al. (2024) enhanced the ability to follow instructions related to translation direction during the instruction tuning process. Different from these approaches, we focus on transferring the translation capabilities from stronger LLM models to weaker LLM models under instruction tuning.

2.2 Knowledge Distillation for Large Language Models

The application of KD in LLM falls into two categories: black-box KD which accesses only teacher-generated texts (Chen et al., 2024; Hsieh et al., 2023; Taori et al., 2023; Peng et al., 2023a), and white-box KD which can employ the teacher parameters. Recently, with the increasing accessibility of open-source models, white-box distillation has gained more attention, particularly concerning the role of KL divergence (Zhong et al., 2024; Gu et al., 2024; Wu et al., 2024; Ko et al., 2024; Agarwal et al., 2024). Concurrently, several works aim to mitigate the training-inference mismatch problem by leveraging generated text of the student

model (Agarwal et al., 2024; Gu et al., 2024; Ko et al., 2024). Nevertheless, they ignore the efficacy of prior knowledge and indiscriminately handle tokens without differentiation. In this paper, we draw from human learning patterns, dynamically providing prior knowledge based on the learning state of the student model to enhance the distillation. Notably, a concurrent work – SKEW KLD loss (Gu et al., 2024) is also a modified distillation function that integrates the prior knowledge into the student model. However, it still maintains the traditional uniform distillation strategy and limits prior knowledge to the teacher's knowledge. Besides, although Zhong et al. (2024); Wang et al. (2021) assign varied distillation modes by token category, they fix the classification ratio and ignore the effectiveness of prior knowledge.

2.3 Self-Evolution Learning

Self-evolution learning is a novel and effective method to exploit the knowledge from data, it is designed to regularize the model training by dynamically learning under-explored tokens. For example, Zhong et al. (2023b) dynamically selected hard-to-learn tokens, then encourages the model to learn smoothed distribution which considers precise reference labels and easily digestible distribution generated by the model itself, thereby improving the training efficiency and scalability (up to 6 billion in their follow-up technical report (Zhong et al., 2022)). Peng et al. (2023b) employed a similar strategy and verified the effectiveness of this learning on typical sequence-to-sequence learning tasks, e.g., machine translation, summarization and grammatical error correction tasks. Moreover, Zheng et al. (2023) introduced self-evolution learning to construct more adaptive and model-friendly pseudo samples to strengthen the mix-up-based text classification model. In this work, we focus on applying this learning strategy to distil the translation-tailored LLMs.

3 Self-Evolution Knowledge Distillation

3.1 Preliminary

We provide some preliminary information about KD for LLMs. It typically employs a pre-trained and fixed teacher model to transfer knowledge into the parameterized student model by providing soft labels from the teacher's output. Given source and ground-truth sequence pair (s, t) from a fixed dataset (S, T), KD could be formulated as an optimization problem aimed at minimizing the Kullback-Leibler (KL) divergence between the token-level distributions of the student $\mathbf{q}$ and teacher $\mathbf{p}$ models:

$$\mathcal{L}_{kl}(\mathbf{p}||\mathbf{q}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}(t_i|t_{<i},s)log\frac{\mathbf{p}(t_i|t_{<i},s)}{\mathbf{q}(t_i|t_{<i},s)}, \quad (1)$$

where N is the length of ground-truth sequence t=$\{t_1, ..., t_n\}$.

Furthermore, during the KD process, the student model also requires training under the ground-truth sequence t. The corresponding training objective could be calculated:

$$\mathcal{L}_{sft} = \frac{1}{N}\sum_{i=1}^{N}(-log\mathbf{q}(t_i|t_{<i},s)). \quad (2)$$

Finally, the overall loss function of KD is a linear interpolation between the supervised fine-tuning (SFT) loss and the KL loss:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{sft} + \lambda\mathcal{L}_{kl}(\mathbf{p}||\mathbf{q}), \quad (3)$$

where the parameter λ serves as a weight to control the influence of each loss.

3.2 Self-Evolution Knowledge Distillation

Here we introduce the proposed self-evolution knowledge distillation (Self-Evolution KD) in detail. As illustrated in Figure 1, it primarily emulates the human teaching mode and comprises two stages: evaluates the student's learning status to identify its hard and easy parts (Stage *1*), offers varied proportions of teacher knowledge for different parts to assist the student learning (Stage *2*).

Stage *1* Self-Question Stage Due to the imbalanced nature of token properties (Piantadosi, 2014), we recommend evaluating the student's learning status at the token level. Therefore, the goal of this stage is to classify tokens as either hard-to-learn or easy-to-learn. However, *how to categorize these tokens?* Inspired by previous findings that reveal the dynamic training difficulty of tokens during the training process (Peng et al., 2023b), we leverage the model itself to divide tokens wisely.

Specifically, we first calculate the learning difficulty for each ground-truth token t_i, denoted as $\{d_1, ..., d_T\}$:

$$\tilde{\mathbf{y}}_i = (1-\lambda)\mathbf{y}_i + \lambda\mathbf{p}_i \quad (4)$$

$$\mathbf{d}_i = \mathcal{L}_{kl}(\tilde{\mathbf{y}}_i||\mathbf{q}_i), \quad (5)$$

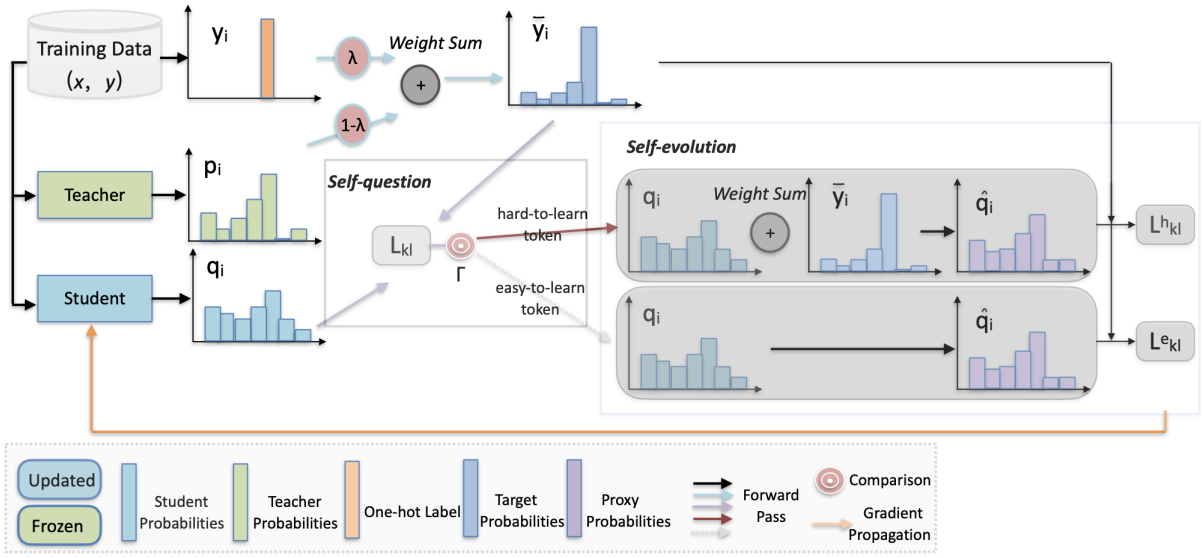


Figure 1: **Overall framework of our Self-Evolution KD**. It mainly contains two stages: ① *self-question*: calculating the learning difficulty by the KL divergence between the student distribution and target distribution, and dividing tokens into different categories. *comparison* means comparing the learning difficulty with the preset threshold Γ; ② *self-evolution*: building proxy distribution for different tokens by smoothing the target and student distributions. ***Updated*** represents that the parameter needs to be updated, while ***Frozen*** means not.

where $\mathbf{y}_i$ is the one-hot distribution of ground-truth token t_i at position i. Notably, previous works primarily assesses the student's learning status by measuring the discrepancy between the student and teacher distributions (Zhong et al., 2024). However, as indicated in Eq. 3, the student's learning status should be determined by both the teacher model and ground-truth sequence, since knowledge distillation inherently involves multiple objectives. Therefore, we combine the distributions of the ground-truth sequence and teacher model linearly to derive the target distribution $\tilde{\mathbf{y}}_i$, and calculate the divergence gap between it and the student distribution to represent the learning difficulty d_i of the student model.

Then, we preset a threshold Γ, and select tokens which corresponding KL divergence exceed Γ as hard-to-learn tokens, *i.e.*, $\mathcal{T}_h = \{t_i | d_i > \Gamma\}$ where $i \in \{1, ..., N\}$, and the others are easy-to-learn tokens.

Stage *2* Self-Evolution Stage After identifying different types of tokens, our primary focus shifts to leveraging the teacher information to enhance the learning of the student model. Although previous work (Qiu et al., 2022) employs the teacher's hidden states as input improve the distillation, there exists a dimensionality mismatch problem due to the model size discrepancy between the student and teacher, while significantly increases training time. Therefore, *how to integrate teacher information in a lightweight way to enhance the distillation?* It is important to note that the "teacher" information here not only comprises the knowledge from the teacher model but also contains the ground-truth information, as discussed in Stage *1*. To address this problem, we propose to promote token distillation by a simple method – distribution adjustment.

Specifically, we introduce a parameter β to mix the student distribution $\mathbf{q}_i$ and target distribution $\tilde{\mathbf{y}}_i$ to obtain proxy distribution $\hat{\mathbf{q}}$. Then, using proxy distribution to match target distribution:

$$\hat{\mathbf{q}}_i = \beta \mathbf{q}_i + (1 - \beta)\tilde{\mathbf{y}}_i \tag{6}$$

$$\mathcal{L}_{kl}^{h_i} = \mathcal{L}_{kl}(\tilde{\mathbf{y}}_i || \hat{\mathbf{q}}_i). \tag{7}$$

By adjusting the information on the distribution, we avoid the dimensionality mismatch problem caused by integrating hidden states, while incurring almost no extra training cost. Furthermore, since the student distribution owns a partial target distribution, it empirically leads to faster convergence and superior performance (Ko et al., 2024).

As for the easy-to-learn tokens, given their ability to effectively capture the "teacher" information, no additional modifications are necessary. Their corresponding optimization objective is as follows:

$$\mathcal{L}_{kl}^{e_i} = \mathcal{L}_{kl}(\tilde{\mathbf{y}}_i || \mathbf{q}_i). \tag{8}$$

Overall Optimization Finally, we combine the losses of the hard-to-learn tokens and the other tokens. The overall optimization objective is formulated as:

$$\mathcal{L} = \frac{1}{N}(\sum_{i \in \mathcal{T}_e} \mathcal{L}_{kl}^{e_i} + \sum_{j \in \mathcal{T}_h} \mathcal{L}_{kl}^{h_j}), \tag{9}$$

where $\mathcal{T}_e$ and $\mathcal{T}_h$ represents the easy-to-learn token set and hard-to-learn token set, respectively.

4 Experimental Setup

4.1 Training Data

For training data, we use a small yet high-quality parallel dataset[1] following Xu et al. (2023). It contains 14k and 12k parallel sentence pairs on English-German (EN-DE) and English-Czech (EN-CS) tasks, respectively. Then, we formatted them into translation instructions in four language directions: En→De, De→En, En→Cs and Cs→En, according to the translation prompt, which resulted in 52K multilingual training sets. Following Jiao et al. (2023a), our translation instructions include a fixed preface for all tasks, followed by "### Instruction:" to describe the translation task, "### Input:" for presenting the source sentence, and a "### Response:" with the target sentence to be generated.

4.2 Model Training

Our experiments are conducted based on the *Llama-factory*[2] codebase with the Llama family models (Touvron et al., 2023). We use supervised fine-tuned (SFT) Llama1-13B trained on the train data as the teacher model, and regard the Llama1-7B as the student model. We fine-tuned all models for 3 epochs with a batch size of 128, while keeping a maximum text length of 512. Besides, we set the learning rate and warmup_ratio as $2e-5$ and 0.03, respectively. The experimental parameters and train data remain consistent in both the SFT and KD process. However, while evaluating the performance on the final checkpoints in knowledge distillation, we employ a validation dataset to select the best checkpoint during SFT, as the model easily overfits in this process. The validation dataset is composed of WMT21 En→De and Cs→En test data (Akhbardeh et al., 2021). All experiments are conducted on NVIDIA 8*A800 (80GB) GPUs and utilize DeepSpeed ZeRO[3] Stage 3 for efficient model parallelism.

4.3 Evaluation

Test Data We evaluated the translation performance on the widely used WMT22 test datasets. It is the test sets from the WMT 2022 competition (Kocmi et al., 2022)[4], which consists of diverse domains such as news, social, e-commerce, and conversational. The number of sentence pairs for De→En, En→De, En→Cs and Cs→En is 1984, 2037, 2037 and 1448, respectively.

Metrics For automatic evaluations, we use SacreBLEU (Post, 2018)[5] and the COMET score (Rei et al., 2020)[6] with *Unbabel/wmt22-comet-da*. Specifically, SacreBLEU primarily calculates n-gram similarity to measure the surface lexical matching, while COMET relies on cross-lingual pre-trained models to obtain human-like semantic matching.

Baselines We consider two traditional knowledge distillation (KD) baselines in our main experiment:

- **Forward KD** (Hinton et al., 2015): is defined in Eq. 3;
- **Reverse KD** (Agarwal et al., 2024): $(1-\lambda)\mathcal{L}_{sft} + \lambda\mathcal{L}_{kl}(\mathbf{q}||\mathbf{p})$, swaps the roles of the teacher and student distributions compared to Forward KD.

In addition, we dynamically divide tokens into two groups: *easy-to-learn* and *hard-to-learn*, and our method assumes that different tokens require different distillation modes. To better observe the impact of these dynamic changes, we introduce two extreme comparison baselines:

- **NoEvo KD**: $\mathcal{L}_{kl}(\tilde{\mathbf{y}}||\mathbf{q})$, treat all tokens as easy-to-learn;
- **SKEW KD**: $\mathcal{L}_{kl}(\tilde{\mathbf{y}}||\beta\mathbf{q} + (1-\beta)\tilde{\mathbf{y}})$, regard all tokens as hard-to-learn tokens, which is similar to SKEW KLD (Ko et al., 2024). The default value of β is 0.5, which is consistent with our **Self-Evolution KD** approach.

For reference, we also report the performances of Llama-13B and Llama-7B models after SFT as the upper and lower bounds. Additionally, for two traditional baselines, we closely follow Gu et al. (2024), integrating the SFT loss and KL loss with a mixture ratio λ=0.5. For our proposed baselines and our method, we combine the teacher distribution and the one-hot distribution of ground-truth

[1] https://github.com/fe1ixxu/ALMA
[2] https://github.com/hiyouga/LLaMA-Factory
[3] https://github.com/microsoft/DeepSpeed
[4] https://www.statmt.org/wmt22/translation-task.html
[5] https://github.com/mjpost/SacreBleu
[6] https://github.com/Unbabel/COMET

at a 0.5 ratio to form the target distribution, omitting the SFT loss since the target distribution already encapsulates the ground-truth information. We compute the loss exclusively on the ground-truth sequence. All models use the beam search strategy (Vaswani, 2017; Freitag and Al-Onaizan, 2017) during inference. Due to the high computational cost and potential out-of-memory (OOM) issues associated with beam search, we set the beam size to 1 as default.

5 Experimental Results

5.1 Main Results

We report the comparison of our **Self-Evolution KD** and other competitive baseline distillation methods in Table 1, in terms of translation performance (SacreBLEU and COMET scores). We have the following observations:

Larger models produce better translations. We observe that the translation ability of the Llama model improves with increased model capacity. These gaps between teachers and students exist across all language pairs, which means there is much room for our KD methods.

Forward KD is effective Table 1 shows that Forward KD could achieve an average gain of 0.2 SacreBLEU points and 0.9 COMET score over SFT, highlighting its effectiveness. However, when compared to the performance gap between the teacher and student models (about 1.3 SacreBLEU points and 1.5 COMET score on average), the gains are relatively modest. This limitation suggests that Forward KD does not fully leverage the potential of the teacher model.

Furthermore, recent methods (Gu et al., 2024; Kim et al., 2024) argue that Reverse KD is more suitable for large language models than Forward KD. Nevertheless, our findings indicate that pure Forward or Reverse KD yields similar performance without significant differences. Consequently, in this paper, we emphasize the substantial benefits derived from dynamically integrating prior knowledge for each token.

Self-Evolution KD improves distillation Table 1 shows that:

- Compared to baseline Forward KD, Self-Evolution KD achieves significant improvements (average: +1.44 SacreBLEU points / +0.28 COMET scores). It shows a maximum improvement of approximately 2.33 SacreBLEU points and about 0.47 COMET scores in the Cs→En test set. Besides, Self-Evolution KD achieves a performance comparable to that of the teacher model and even surpasses it on the SacreBLEU metric. These substantial gains underscore the efficiency of our approach.

- As an adaptive strategy, Self-Evolution KD dynamically allocates different strategies based on the learning status of each token. It surpasses both static strategies, as shown in Table 1, corroborating our hypothesis that indiscriminately distilling each token is a sub-optimal strategy. Adopting varied distillation modes on the basis of the student's learning status would better align the distillation curve of the student model, thereby fully unleashing its potential. Moreover, with the integration of prior knowledge, SKEW KD also shows significant improvement over Forward KD, with average gains of 0.9 SacreBLEU points.

5.2 Ablation Study

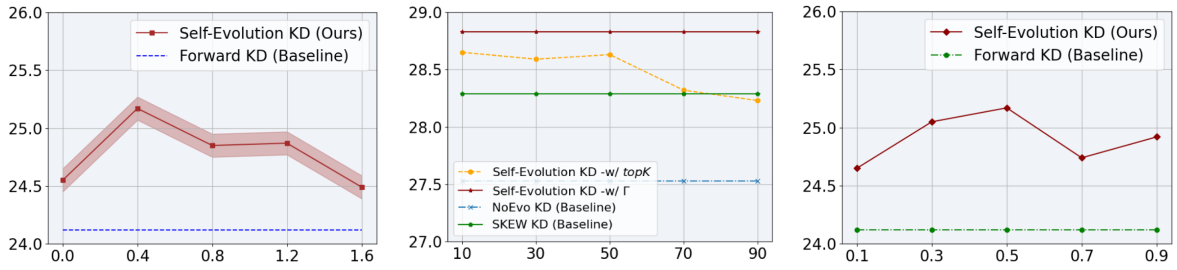
5.2.1 Effect of Token Selection

As a key point in this paper, how to divide tokens after evaluating the student's learning state is worth considering. Traditional methods typically select top K percent of all tokens as hard-to-learn tokens (Zhong et al., 2024). However, we suspect that this approach is sub-optimal, since it forces the model to choose a fixed proportion of hard-to-learn tokens even in the later stages of distillation. In contrast, dynamically selecting hard-to-learn tokens based on a preset threshold Γ would avoid choosing hard-to-learn tokens after all tokens have been sufficiently learned, thereby fully harnessing the potential of the student model. In this section, we delve into the impact of the two strategies. The former is named **Self-Evolution KD -w/ top**K, while the latter is **Self-Evolution KD -w/** Γ.

First, we examine the influence of the hyper-parameter Γ within our dynamic strategy. As shown in Fig 2(a), we observe that a larger or smaller threshold detrimentally affects the performance of Self-Evolution KD. The former leads to an overabundance of easy-to-learn tokens and fails to adequately focus on tokens that need assistance, while the latter is the opposite. Self-Evolution KD performs best with an optimal value of $\Gamma = 0.4$,

	En → De		De → En		En → Cs		Cs → En		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
					Test: *WMT22 Test sets*					
Teacher	26.69	82.77	28.47	83.43	20.84	81.37	37.9	83.67	28.48	82.81
Student	25.14	81.08	27.53	82.92	19.18	78.11	36.92	83.10	27.19	81.30
Forward KD	25.29	81.84	27.77	83.09	20.39	80.60	36.11	83.27	27.39	82.20
Reverse KD	25.37	81.77	27.41	83.03	20.5	79.83	35.34	83.22	27.16	81.96
NoEvo KD	25.51	81.70	27.85	83.02	20.56	80.78	36.21	83.30	27.53	82.20
SKEW KD	26.1	81.74	28.09	83.20	21.22	80.45	37.77	83.53	28.29	82.23
Ours										
Self-Evolution KD	**26.73**†	**82.05**	**28.62**†	**83.42**	**21.54**†	**80.71**	**38.44**†	**83.74**	**28.83**	**82.48**
Δ	*+1.44*	*+0.21*	*+0.85*	*+0.33*	*+1.15*	*+0.11*	*+2.33*	*+0.47*	*+1.44*	*+0.28*

Table 1: **Comparison results of our Self-Evolution KD against baselines** on different translation tasks, where "Δ" indicates the improvement against **Forward KD**, and "†"indicates statistically significant difference ($p<0.05$). **Student** and **Teacher** represent the Llama-7b and Llama-13b after SFT.



(a) Values of Γ (b) Values of $K(\%)$ (c) Values of β

Figure 2: 2(a) and 2(b): **Effect of Γ and percent (K)** for selecting hard-to-learn tokens. 2(c): **Effect of β** to determine the mixsture proportion of prior knowledge. We report their average SacreBLEU points on the above-mentioned validation dataset in 2(a) and 2(c). As for 2(b), the average SacreBLEU points on WMT22 test sets are reported since we compare different distillation strategies.

thus we retain it as our default setting.

Second, from Fig 2(b), the Self-Evolution *-w/ topK* outperforms the NoEvo KD and SKEW KD, it further demonstrates the effectiveness of assigning distinct distillation strategies to different tokens. However, its performance is inferior to the dynamic selection strategy (*-w/ Γ*), which is consistent with our aforementioned hypothesis. Takeaway: ***These observations suggest that selecting the appropriate number of hard-to-learn tokens is crucial, since a larger or smaller number would lead to negative impacts. Besides, a dynamic token selection strategy aligns better with the learning patterns of the student model, thus fully unlocking its potential.***

5.2.2 Influence of β

The factor β in Eq. 6, which serves to control the proportion of target distribution integrated to the student distribution, also requires to be investigated. Figure 2(c) shows the results of varied β ranging from 0.1 to 0.9. As observed, the model performs optimally with $\beta = 0.5$, thus we adopt it as default setting in our experiments.

Besides, we also employ the progressive strategy outlined in Qian et al. (2020), dynamically adjusting the integration ratio of the target distribution. Specifically, we set an initial ratio (β_b) and linearly decrease it to a predefined final value (β_e) throughout the training process. As seen in Figure 3, the dynamic strategy fails to bring substantial gain, despite its potential advantages. Takeaway: ***it proves that indiscriminately distill each token, even with dynamic adjustments to prior knowledge, still constrains the model performance, leading to sub-optimal outcomes.***

5.3 Further Analysis

5.3.1 Does Self-Evolution KD Transfer The Teacher's Knowledge Better?

The core of Knowledge Distillation (KD) is to transfer the distilled knowledge from a well-performing but cumbersome teacher model to a compact and lightweight student model, thus we analyze the ef-

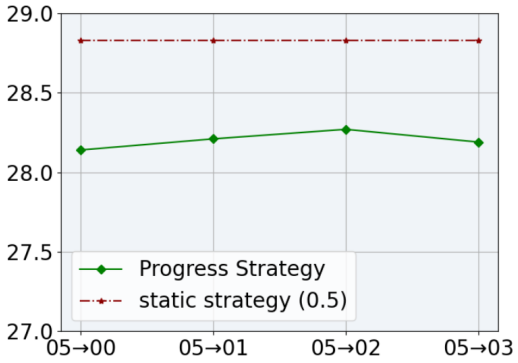


Figure 3: **Comparison of static strategy and progressive strategy for factor** β. "0.5→0.0" means the β_b is 0.5 and the β_e is 0.0. *static strategy (0.5)* indicates the results of Self-Evolution KD ($\beta = 0.5$).

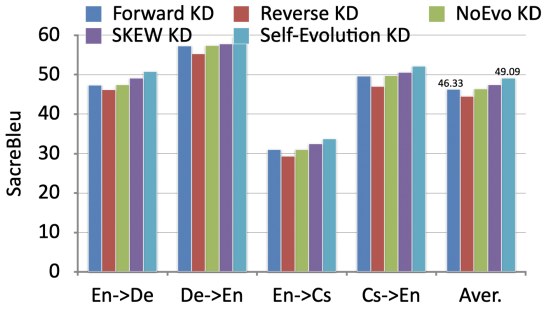


Figure 4: **Comparison of teacher's knowledge transfer across different distillation strategies.**

fectiveness of knowledge transfer across various distillation strategies in this part. We regard the generation text of teacher model on the WMT22 En↔De, En↔Cs test sets as the "reference", calculate the SacreBLEU scores between the "reference" and generated text of various distillation strategies. As illustrated in Figure 4, SKEW KD is superior to Forward KD in terms of similarity to the teacher model's generated text. Notably, Self-Evolution KD achieves the best performance, with an average improvement of 2.8 gains across the four language pairs. Takeaway: ***These findings indicate that the introduction of prior knowledge enables the student model to better capture the target information and enhances the efficiency of the teacher's knowledge transfer. Additionally, a dynamic strategy that integrates prior knowledge based on the student's learning status further enhances its potential and deepens its understanding of the teacher's knowledge.***

5.3.2 Comparison of Prior Knowledge

Given that knowledge distillation inherently involves multi-objective learning, we suspect that

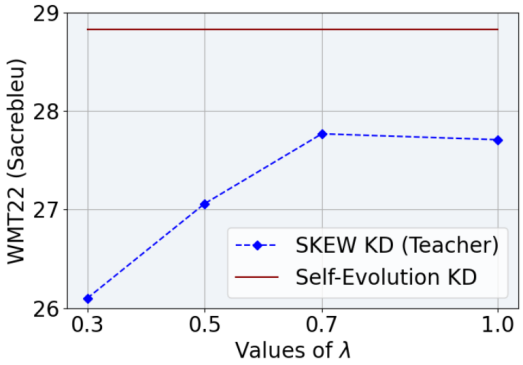


Figure 5: **Effect of the loss weight of the SKEW KD (Teacher).** We only report the Self-Evolution KD for reference.

only considering the teacher knowledge as prior knowledge would significantly reduce the KL divergence loss and mislead the student model to emphasize the SFT loss, thus potentially curtailing the benefits derivable from teacher knowledge. Following Ko et al. (2024), we redefine the distillation objective as (SKEW KD (teacher)):

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{sft} + \lambda\mathcal{L}_{kl}(p||\beta q_\theta + (1-\beta)p), \quad (10)$$

where $\beta = 0.9$(Ko et al., 2024).

As shown in Figure 5, we observe the results when setting λ to 0.3, 0.5, 0.7 and 1.0, respectively. While the parameter λ is set to 1.0, as Ko et al. (2024), the result proves that ignoring the ground-truth information leads to a significant decline in performance. Besides, despite adjusting the loss weights to balance the SFT loss and the KL divergence loss, the performance of SKEW KD (teacher) still significantly lags behind that of Self-Evolution KD. Takeaway: ***This finding is consistent with our conjecture that modifying the distillation optimization function by only integrating the teacher's knowledge disrupts the original cooperative relationship between multiple objectives, thus weakening the student's performance.***

5.3.3 Larger Teacher Model

To assess the effectiveness of knowledge distillation in models with substantial size disparities, we employ Llama-30b models after SFT as teachers, distilling knowledge into student models with 7 billion parameters. Table 2 provides a comprehensive comparison against several strong baselines. As observed, Self-Evolution KD significantly outperforms the baseline (Forward KD), with an average improvement of approximately 1.7 SacreBLEU points, and surpasses other distillation strategies.

WMT22	En→De	De→En	En→Cs	Cs→En	Avg.
Teacher	26.81	27.53	20.64	38.17	28.29
Forward KD	25.36	27.17	19.91	35.42	26.96
Reverse KD	24.92	26.62	19.81	34.40	26.44
NoEvo KD	25.32	27.13	19.77	35.92	27.04
SKEW KD	25.97	**28.14**	20.91	37.24	28.07
Self-Evolution KD	**26.73**	**28.14**	**21.65**	**38.07**	**28.65**

Table 2: **Comparison results for different distillation strategies** between the teacher model Llama-30B after SFT and the student model Llama-7B(**SacreBLEU**).

Notably, the distillation gains from the teacher model with 30 billion parameters surpass those from the teacher model with 13 billion parameters (e.g., 1.7 SacreBleu point V.S. 1.5 SacreBleu point). Takeaway: ***It clearly indicates that our method consistently achieves superior performance, even in large model size gaps. It also shows the potential to mitigate the adverse effects of distillation caused by large teacher models.***

6 Conclusion

This paper explores the potential of utilizing prior knowledge in LLM knowledge distillation and highlights the limitations of equally distilling each token without differentiation. In particular, we propose a dynamic teaching mode that enhances student distillation by integrating prior knowledge which contains teacher knowledge and ground-truth information. Besides, we adjust the integration ratio based on the student's token-level learning status. Experimental results demonstrate that our approach consistently enhances distillation performance across multiple translation tasks. Furthermore, in-depth analysis indicates that our method effectively transfers the teacher's knowledge.

Limitation

Our work has several potential limitations. First, given the limited computational budget, our method has not been validated across extensive model size gaps, such as 65B. Scaling up to larger model sizes will be more convincing. Second, one key factor Γ is empirical and preset value, we fix it throughout the training, following previous works (Peng et al., 2023b). It would be sensible and elegant to dynamically determine the threshold according to training difficulty. For example, you can employ an additional network to predict the integration ratio of prior knowledge for each token.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120).

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*.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the sixth conference on machine translation*.

Hongzhan Chen, Xiaojun Quan, Hehong Chen, Ming Yan, and Ji Zhang. 2024. Knowledge distillation for closed-source language models. *arXiv preprint arXiv:2401.07013*.

Kehai Chen, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020. Content word aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*.

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.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

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

Junliang Guo, Zhirui Zhang, Linli Xu, Boxing Chen, and Enhong Chen. 2021. Adaptive adapters: An

efficient way to incorporate bert into neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. Minneapolis, Minnesota.

Gyeongman Kim, Doohyuk Jang, and Eunho Yang. 2024. Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning. *arXiv preprint arXiv:2402.12842*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, et al. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*.

Mike Lewis, Yinhan Liu, Naman Goyal, et al. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xuebo Liu, Longyue Wang, Derek F Wong, Liang Ding, Lidia S Chao, Shuming Shi, and Zhaopeng Tu. 2021. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

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*.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*.

Shilong Pan, Zhiliang Tian, Liang Ding, Haoqi Zheng, Zhen Huang, Zhihua Wen, and Dongsheng Li. 2024. POMP: Probability-driven meta-graph prompter for LLMs in low-resource unsupervised neural machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023a. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong, and Dacheng Tao. 2023b. Token-level self-evolution training for sequence-to-sequence learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023c. Towards making the most of chatgpt for machine translation. In *Findings of EMNLP*.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT 2018*.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2020. Glancing transformer for non-autoregressive

neural machine translation. *arXiv preprint arXiv:2008.07905*.

Zengyu Qiu, Xinzhu Ma, Kunlin Yang, Chunya Liu, Jun Hou, Shuai Yi, and Wanli Ouyang. 2022. Better teacher better student: Dynamic prior knowledge for knowledge distillation. *arXiv preprint arXiv:2206.06067*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*.

Jun Rao, Xuebo Liu, Zepeng Lin, Liang Ding, Jing Li, Dacheng Tao, and Min Zhang. 2024. Exploring and enhancing the transfer of distribution in knowledge distillation for autoregressive language models. *arXiv preprint arXiv:2409.12512*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

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*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Zhe Zhao, and Ngai Wong. 2024. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*.

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*.

Changtong Zan, Liang Ding, Li Shen, Yu Cao, Weifeng Liu, and Dacheng Tao. 2022a. Bridging cross-lingual gaps during leveraging the multilingual sequence-to-sequence pretraining for text generation and understanding. *arXiv preprint arXiv:2204.07834*.

Changtong Zan, Liang Ding, Li Shen, Yu Cao, Weifeng Liu, and Dacheng Tao. 2022b. On the complementarity between pre-training and random-initialization for resource-rich machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building accurate translation-tailored llms with language aware instruction tuning. *arXiv preprint arXiv:2403.14399*.

Songming Zhang, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, Jian Liu, and Jie Zhou. 2022. Conditional bilingual mutual information based adaptive training for neural machine translation. *arXiv preprint arXiv:2203.02951*.

Haoqi Zheng, Qihuang Zhong, Liang Ding, Zhiliang Tian, Xin Niu, Dongsheng Li, and Dacheng Tao. 2023. Self-evolution learning for mixup: Enhance data augmentation on few-shot text classification tasks. *arXiv preprint arXiv:2305.13547*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023b. Self-evolution learning for discriminative language model pretraining. *arXiv preprint arXiv:2305.15275*.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Revisiting knowledge distillation for autoregressive language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue. *arXiv preprint arXiv:2212.01853*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

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

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023b. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.