BEYOND TEACHERS: STUDENT-ASCENDANT KNOWLEDGE DISTILLATION FOR LLM-BASED MACHINE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) can effectively compress a large machine translation (MT) model into a smaller variant while preserving performance, supporting deployment in constrained environments, especially under the context of large language model (LLM)-based MT. However, the smaller student model is intrinsically bounded by the performance ceiling of the teacher model in recent studies of KD. Yet for humans, students routinely outstrip their teachers through progressive, deliberate learning. In this paper, we rethink the process of KD and propose SaKD, **S**tudent-**A**scentant **K**nowledge **D**istillation, to break the performance ceiling of teacher model in KD for LLM-based MT via mimicking the human learning behaviors. SaKD first distills knowledge from multiple teachers, and then introduces a human-like three-stage learning process, *Imitation*, *Reflection*, and *Transcendence*—to guide the student model toward surpassing its teachers. Experiments on extensive MT benchmarks show the effectiveness of SaKD, significantly outperforming existing KD methods. And importantly, the student model with SaKD transcends the performance bound of the teacher model, achieving a complete performance surpass.[1]

1 INTRODUCTION

Large language models (LLMs) BigScience (2022); Google (2022); MetaAI (2023); OpenAI (2023); DeepSeekAI (2024); Yang et al. (2025); KimiTeam (2025) have demonstrated strong performance on machine translation (MT). However, the ever-increasing size of pretrained LLMs creates substantial challenges for deploying these systems in low-latency settings or on resource-constrained hardware. Knowledge distillation (KD) Li et al. (2024); Yang et al. has proven effective for compressing large models, such as BERT Devlin et al. (2018), with limited performance degradation. Nevertheless, in conventional KD, the performance of student models is often bounded by that of their teachers, which restricts further improvement. In contrast, human learning suggests that students can progressively refine acquired knowledge and eventually surpass their teachers.

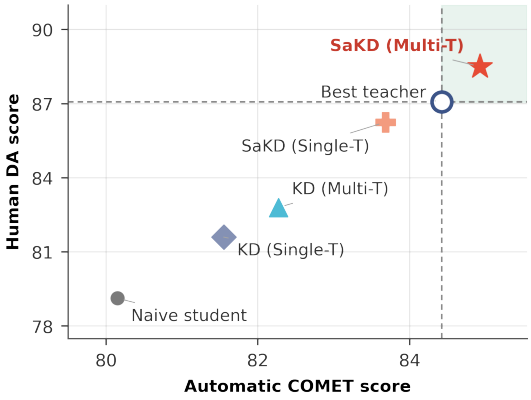


Figure 1: Performance comparison of the Qwen3-8B student model under automatic and human evaluation. Dashed lines mark the best-teacher scores, and the shaded upper-right region indicates simultaneous improvement over the best teacher. **Single-T** and **Multi-T** denote knowledge generated from one and three teacher models, respectively.

Motivated by this observation, we rethink the KD process and introduce SaKD (**S**tudent-**A**scendant **K**nowledge **D**istillation), a novel KD framework for MT. By emulating the progressive nature of human learning, SaKD enables student

[1]We will release our code on Github.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

models not only to match but also to exceed the translation performance of their teachers. Inspired by Kim & Rush (2016), SaKD adopts a pseudo-labeling KD approach Shleifer & Rush (2020); Kasai et al. (2021). Specifically, SaKD first leverages teacher models to generate translation knowledge, i.e., pseudo-labeled data, and then organizes student learning into three stages analogous to human learning: *Imitation*, *Reflection*, and *Transcendence*.

Unlike prior approaches that rely on a single teacher model, we generate knowledge collaboratively from multiple teacher models. This design is inspired by human pedagogy: exposure to diverse viewpoints from multiple teachers encourages critical comparison and reflection, enabling students to move beyond the limitations of any single mentor and develop a deeper and more comprehensive understanding. Similarly, we argue that heterogeneous translation knowledge distilled from an ensemble of teacher models provides the student with richer translational signals, mitigates biases from any single source, and lays a more robust foundation for subsequent improvement. Figure 1 provides a comparison from both automatic evaluation and human evaluation, showing that multi-teacher knowledge generation consistently improves the student model over single-teacher distillation and helps SaKD move beyond the best teacher under both evaluation protocols.

Building on this multi-teacher knowledge, SaKD guides the student through three stages that mirror the progression of human learning: acquiring knowledge, evaluating alternatives, and exploring beyond learned examples. During *Imitation*, the student learns from teacher-generated translations through supervised fine-tuning (SFT), establishing a foundation for translation. *Reflection* builds on this foundation through Contrastive Preference Optimization (CPO) Xu et al. (2024a), which trains the student to distinguish higher-quality translations from lower-quality alternatives produced by different teachers. Finally, *Transcendence* shifts the focus from learning from teacher outputs to exploring translations beyond them. To this end, we employ Group Relative Policy Optimization (GRPO)-based reinforcement learning Shao et al. (2024) with a transcendence-aware reward that encourages the student to generate translations of higher quality than those of its teachers.

Overall, the main contributions of this paper are summarized threefold:1) We propose a multi-teacher knowledge distillation method that aggregates diverse knowledge from multiple LLMs, enabling the student model to surpass individual teacher models; 2) We introduce a human-like three-stage learning framework, consisting of imitation, reflection, and transcendence, which is task-agnostic and can be readily extended to text generation tasks beyond machine translation; 3) Extensive experiments on translation tasks demonstrate the effectiveness of the proposed approach.

2 Related Work

LLM-based Machine Translation. With the rise of Large Language Models (LLMs), their translation capabilities have attracted extensive research attention. Previous work has evaluated LLMs in multilingual and long-text translation scenarios Zhu et al. (2023a); Jiao et al. (2023b); Hendy et al. (2023); Wang et al. (2023); Karpinska & Iyyer (2023), and has also investigated the phenomenon of translation hallucinations Guerreiro et al. (2023). Another line of research focuses on instruction tuning, demonstrating that translation performance can be improved through carefully designed training data, staged training paradigms, human feedback, reinforcement learning, and self-refinement techniques Zhu et al. (2023b); Xu et al. (2023); Jiao et al. (2023a); Zan et al. (2024); Lyu et al. (2024); Xu et al. (2024a); Feng et al. (2025b;a). In contrast to these studies, our work focuses on transferring translation capabilities from stronger LLMs to weaker ones, employing a human-like learning framework to enable student models to surpass their teachers.

Knowledge Distillation for LLM. Knowledge distillation for LLM can be broadly categorized into black-box and white-box approaches. Black-box KD relies solely on the outputs generated by the teacher model without accessing its internal parameters Chen et al. (2024); Hsieh et al. (2023); Taori et al. (2023); Peng et al. (2023), whereas white-box KD directly leverages the teacher's parameters during training. With the increasing availability of open-source LLMs, recent research has increasingly focused on white-box distillation, particularly exploring the role of KL divergence in aligning the distributions of the student and teacher models Zhong et al. (2024); Gu et al. (2024); Wu et al. (2024); Ko et al. (2024); Agarwal et al. (2024); Velayuthan et al. (2025). Several studies further address the training-inference mismatch by incorporating outputs generated by the student model into the distillation process Agarwal et al. (2024); Gu et al. (2024); Ko et al. (2024); Song et al. (2025). However, existing studies predominantly distill knowledge from a single teacher, which

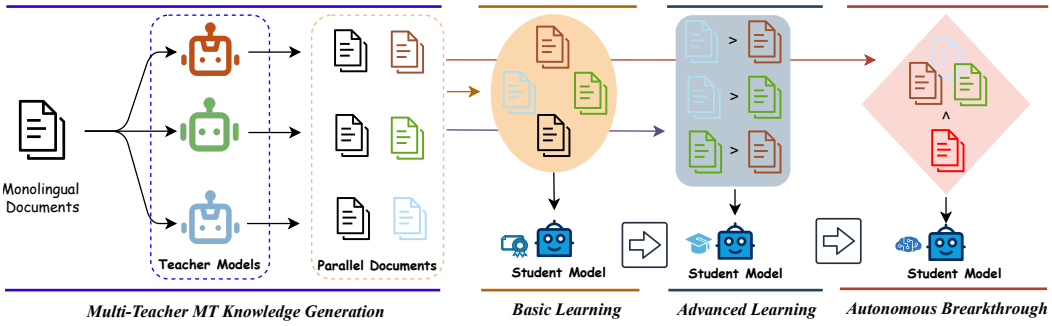


Figure 2: Illustration of the proposed human-like three-stage knowledge distillation framework. Notably, the framework can be easily transferred to other text generation tasks.

inherently constrains the student's performance to the teacher's capacity. Differently, we propose a human-inspired distillation framework that integrates knowledge from multiple teacher models, enabling the student to achieve a higher performance ceiling.

3 METHODOLOGY

In this section, we introduce the proposed SaKD, **S**tudent-**A**scendant **K**nowledge **D**istillation, which distills the MT capabilities of the multiple high-capacity teachers LLM into a lightweight student LLM, yielding student-ascendant performance: the resulting compact model not only matches but also empirically exceeds the translation quality of its substantially larger progenitor. Given a source document set $\mathcal{D} = \{S_1, \ldots, S_a\}$, we detail the procedure of SaKD in the following section, and the overview illustration of SaKD is provided in Figure 2.

3.1 MULTI-TEACHER KNOWLEDGE GENERATION

SaKD adopts a multi-teacher sequence-level distillation approach Kim & Rush (2016); Gandhi et al. (2023), initially leveraging multiple teacher models to generate knowledge with diverse distributions. This design emulates the early learning stage of students, in which they engage in broad knowledge acquisition, thereby laying a solid foundation for subsequent breakthroughs.

Initialized MT Knowledge Generation. Given $\mathcal{D}$ and a teacher models set $\mathcal{M} = \{M_1, \ldots, M_b\}$, we first distill the initialized MT knowledge from all teacher models. Specifically, for i-th document S_i in $\mathcal{D}$, its MT knowledge T_i^j from j-th teacher model in $\mathcal{M}$ can be obatined as follows:

$$T_i^j = \text{argmax}_T \, p_{\theta_{M_j}}(T \mid S, inst), \tag{1}$$

where we directly prompt teacher model M_j to generate the translation of S_i, i.e., T_i^j, as the MT knowledge. $inst$ is MT prompt for the teacher model, as shown in Appendix D. In this way, we can obtain all MT knowledge $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_b\}$ from all teacher models for the source documents $\mathcal{D}$ and where $\mathcal{T}_j = \{T_1^j, \ldots, T_a^j\}$. Finally, we formalize the source-knowledge data as a set of $(b+1)$-tuple $\mathcal{P}$, i.e., $\mathcal{P} = \{(S_1; T_1^1, \ldots, T_1^b), \ldots, (S_a; T_a^1, \ldots, T_a^b)\}$. **Knowledge Selection.** Although the teacher LLM with large-sized parameters demonstrates strong overall translation performance, it remains prone to hallucinations and residual errors Xu et al. (2024b). To mitigate this, we employ a heuristic approach to filter out unreliable MT knowledge, thereby preserving the effectiveness of subsequent distillation. Specifically, given the i-th source–knowledge tuple in $\mathcal{P}$, we first compute its reference-free MQM scores C_i with XCOMET-QE between S_i and all translations generated by different teacher models, i.e., $C_i = \{c_i^1, \ldots, c_i^b\}$. XCOMET-QE does not require reference translations and is used only for knowledge filtering and training-signal construction. Then, we discard the source–knowledge tuple if all its corresponding MQM scores are below the threshold, i.e., $c < \lambda, \forall c \in C_i$. We apply the filting strategy to all source-knowledge tuples and obtain the final set of source-knowledge. For simplicity, we still denote the filtered source–knowledge tuple set as $\mathcal{P}$, and its corresponding score as $\mathcal{C} = \{C_1, \ldots, C_m\}$, $m = |\mathcal{P}|$.

3.2 *Imitation*: BASIC LEARNING VIA SFT

In the basic learning stage of the student model, the primary objective is to endow the student model-initially devoid of any translation capability-with rapid acquisition of fundamental translation proficiency by strictly imitating the pseudo-parallel corpus $\mathcal{P}$ synthesized through a multi-teacher ensemble. Specifically, we accomplish the objective of the basic learning stage by fine-tuning the student model on $\mathcal{P}$ through minimizing the following cross-entropy loss:

$$\mathcal{L}_{\text{sft}} = -\frac{1}{b \cdot |\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \sum_{j=1}^{b} \log \pi_{\text{sft}}(T_i^j \mid S_i), \tag{2}$$

where π_{sft} denotes the student model in the basic learning stage, while S_i and T_i^j represent the source sequence and its corresponding MT sequence generated by the j-th teacher model, respectively.

Basic learning phase seeks to align the student's performance with, or approximate, the aggregate competence of the teacher ensemble, thereby establishing a robust linguistic and structural substrate for the subsequent learning stages.

3.3 *Reflection*: ADVANCED LEARNING VIA CPO

During the basic learning stage, the student model acquires basic translation proficiency by imitating the "standard answers" provided by multiple teachers. The bounds of performance expectation for the student model can be formalized as follows:

$$\mathbb{E}_{S\sim\mathcal{D}}\Big[\mathbb{E}_{T\sim\pi_{\text{sft}}} r(S,T)\Big] = \mathbb{E}_{S\sim\mathcal{D}}\Big[\tfrac{1}{b}\sum_{j=1}^{b} r(S,T^j)\Big], \tag{3}$$

where $r(\cdot)$ is a translation-quality assessment function, given a source sequence and a corresponding target sequence. From this, we observe that the performance of the student model is inherently capped by the average quality of these teacher outputs. To overcome this ceiling, the student must learn to *reflect* among the teacher-generated translations, identifying the defining characteristics that distinguish a good translation from a poor one. To achieve this, we introduce advanced learning via Contrastive Preference Optimization (CPO)Xu et al. (2024a), which simulates the student's *Reflection* process. Specifically, we first construct the CPO triplet dataset $\mathcal{O}$ using $\mathcal{P}$ and $\mathcal{C}$. For the i-th source–knowledge tuple in $\mathcal{P}$, we rank its translations according to their reference-free MQM scores in $\mathcal{C}$, i.e., C_i, and then select the best and worst translations to form a triplet (S_i, T_i^+, T_i^-), which serves as a CPO training instance. Based on $\mathcal{O} = \{(S_1, T_1^+, T_1^-), \ldots, (S_a, T_a^+, T_a^-)\}$, the CPO training objective in the reflection stage is to minimize the following objective:

$$\begin{aligned}
\mathcal{L}_{\text{cpo}} &= \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{nll}}, \\
\mathcal{L}_{\text{cons}} &= -\frac{1}{|\mathcal{O}|}\sum_{i=1}^{|\mathcal{O}|} \log \sigma\big(\beta \log \pi_{\text{cpo}}(T_i^+|S_i) - \beta \log \pi_{\text{cpo}}(T_i^-|S_i)\big), \\
\mathcal{L}_{\text{nll}} &= -\frac{1}{|\mathcal{O}|}\sum_{i=1}^{|\mathcal{O}|} \log \pi_{\text{cpo}}(S_i|T_i^+),
\end{aligned} \tag{4}$$

where π_{cpo} refers to the student model during advanced learning stage and is initialized from π_{sft}. The $\mathcal{L}_{\text{cons}}$ enforces reflection between good and poor translations, while $\mathcal{L}_{\text{nll}}$ aligns the student model with high-quality translation features.

By explicitly distinguishing superior from inferior translations through contrastive training, CPO encourages the model to iteratively refine its outputs, allowing it to move beyond average teacher performance.

3.4 *Transcendence*: BREAKTHROUGH VIA GRPO

While the reflection stage (CPO) enables the student to distinguish between superior and inferior teacher outputs, its improvement is still bounded by the quality of teacher-provided translations. In other words, the student remains constrained by the teacher demonstration space. To achieve true

Table 1: BLEU and COMET scores on four translation directions when using various student LLMs. The column of ***Average*** refers to the averaged performance across all translation directions. Scores exceeding the teacher models' best performance are indicated with **bold** font. * denotes statistically significant improvement over the best teacher model at $p < 0.01$.

System	En ⇒ Zh		Zh ⇒ En		En ⇒ De		De ⇒ En		*Average*	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Teacher Models										
Qwen3-80B	33.78	87.92	32.45	86.66	30.01	85.49	30.78	86.79	31.76	86.72
Llama-3.3-70B	33.45	87.23	32.56	87.01	30.89	87.11	32.01	88.56	32.23	87.48
Deepseek-R1-70B	32.11	86.21	31.34	84.49	29.78	85.01	30.45	85.23	30.92	85.24
Qwen3-8B as Student Model										
KD (Single-T)	31.78	85.62	31.12	85.53	29.34	83.11	30.11	86.21	30.59	85.12
KD (Multi-T)	32.34	86.21	31.78	86.25	30.01	84.25	30.78	86.83	31.23	85.89
KD (Ensem-T)	32.56	86.25	31.99	86.34	30.45	84.82	31.34	87.51	31.59	86.23
SaKD (Single-T)	33.12	87.41	**32.78**	**87.23**	30.78	87.01	**32.12**	**88.58**	32.20	87.56
SaKD (Multi-T)	**34.45***	**89.21***	**33.56***	**87.99***	**32.23***	**88.09***	**33.45***	**89.51***	**33.42**	**88.70**
Qwen3-4B as Student Model										
KD (Single-T)	29.78	83.55	29.45	84.21	27.78	81.88	28.56	84.69	29.02	83.58
KD (Multi-T)	30.56	84.66	30.12	85.67	28.45	83.04	29.23	85.67	29.59	84.76
KD (Ensem-T)	30.67	84.54	30.34	85.47	28.78	83.56	29.45	85.88	29.81	84.86
SaKD (Single-T)	31.78	86.51	31.45	86.11	30.12	85.01	30.89	87.43	31.31	86.27
SaKD (Multi-T)	32.56	87.22	31.78	**87.11**	**31.45***	**87.21**	**32.12**	**88.77**	**32.48**	**87.57**

Transcendence, the student must move beyond passive reflection and actively optimize the pseudo-label space induced by teachers. Inspired by the success of reinforcement learning in mathematical reasoning, we adapt Group Relative Policy Optimization (GRPO) with sequence-level outcome supervision Shao et al. (2024) to enable this capability.

Given source sequence S_i, we first sample a group of translations $\{\hat{T}_i^1, \ldots, \hat{T}_i^c\}$ from the old student model and then optimize the new student model by minimizing the following objective:

$$\mathcal{L}_{\text{grpo}} = -\frac{1}{c}\sum_{j}^{c} \frac{\pi_{\text{grpo}}(\hat{T}_i^j|S_i)}{\pi_{\text{cpo}}(\hat{T}_i^j|S_i)} A_j + \beta \text{KL}[\pi_{\text{cpo}}||\pi_{\text{grpo}}] \tag{5}$$

where π_{grpo} is the optimized student model during *transcendence* training and initialized from π_{cpo}. KL refers to the penalty term and we follow Shao et al. (2024) and use an unbiased estimator to estimate the Kullback-Leibler divergence. Importantly, the advantage A_j is calculated based on the following *transcendence*-aware rewards of translations inside the group:

$$A_j = \frac{R_j - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})}, \tag{6}$$

$$R_j = r(S_i, \hat{T}_i^j) - \max(C_i), \tag{7}$$

where $\mathbf{R} = \{R_1, ..., R_c\}$ denotes the rewards of all translations in the group. $r(\cdot)$ denotes the reference-free MQM score estimated by XCOMET-QE. C_i is the set of MQM scores for the teacher-generated knowledge of S_i, and $\max(C_i)$ returns the MQM score of the best teacher-generated translation.

In this way, GRPO converts the objective of *transcending teachers* into a reward signal, encouraging the student model to explore solutions beyond the teacher ensemble and improve upon teacher performance through *Transcendence*.

4 Experimentation

4.1 Experimental Setup

Dataset. During the knowledge generation stage, we use the OSCAR monolingual corpus, which covers long-text monolingual data in 152 languages.(`https://huggingface.co/datasets/os`

car-corpus/OSCAR-2301) From this corpus, we extract 3 millions documents each in Chinese, German, and English, and employ the teacher model to construct parallel translation knowledge $\mathcal{P}$. After filtering, around 2 million knowledge instances remain, from which 1.2 million are randomly sampled per translation task for the *Imitation* stage. In the *Reflection* and *Transcendence* stages, we sample 50K long-text parallel knowledge instances from $\mathcal{P}$ to build the training set. In addition, to more effectively unleash the potential of the student model, we further introduce high-quality long-text parallel translation data and report the performance results on the test set used in Lyu et al. (2024). The data processing and detailed statistics of the dataset are provided in Appendix A.

Model Selection and Settings. We employ three powerful open-source LLMs, including two instructed models, `Qwen3-Next-80B-Instruct` (`https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct`), `Llama-3.3-70B-Instruct` (`https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`) and a reasoning-oriented model, `DeepSeek-R1-Distill-Llama-70B` (`https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B`), as the teacher models. For student model, we use two lightweight LLMs in Qwen series, including `Qwen3-8B` (`https://huggingface.co/Qwen/Qwen3-8B`) and `Qwen3-4B` (`https://huggingface.co/Qwen/Qwen3-4B`). We use reference-free MQM scores from XCOMET-QE (`https://huggingface.co/Unbabel/XCOMET-XL`) for knowledge selection, applying a threshold of $\lambda = 0.7$. The impact of λ is further analyzed in Appendix E. For *Reflection* and *Transcendence* training, we follow the settings in Xu et al. (2024a) and Guo et al. (2025a), respectively. We implement the proposed method upon *EasyR1* (`https://github.com/hiyouga/EasyR1`). Please refer to Appendix B for more training and inference details.

Baselines. In addition to teacher-model baselines, we compare our method with three conventional KD approaches: **KD (Single-T)** uses a single teacher (Llama3.3-70B-Instruct) following a two-stage pipeline of knowledge generation and SFT; **KD (Multi-T)** directly combines knowledge generated by multiple teachers for SFT; and **KD (Ensem-T)** distills only the best translation outputs selected from multiple teachers. We further evaluate two variants of our method: **SaKD (Single-T)** uses knowledge from a single teacher, and **SaKD (Multi-T)** leverages knowledge from multiple teachers.[2]

Evaluation. We evaluate SaKD on translation tasks with BLEU and COMET. To avoid reward hacking and metric overfitting, we separate the training/filtering signal from the evaluation metrics: reference-free MQM scores from XCOMET-QE are used only for knowledge filtering and reward construction, whereas the main evaluation is conducted with BLEU Papineni et al. (2002)and COMET Rei et al. (2020). Since the original COMET metric has input-length constraints and may inadequately capture quality distortion in long-text translation, we adopt the Align-then-Slide method proposed by Guo et al. (2025b) to better evaluate long-text translation quality.

4.2 Main Results

Table 1 summarizes the experimental results. For the Qwen3-8B student, **SaKD (Multi-T)** achieves 33.42 BLEU and 88.70 COMET, outperforming **KD (Single-T)** by +2.83 BLEU and +3.58 COMET, and the strongest KD baseline, **KD (Ensem-T)**, by +1.83 and +2.47, respectively. Remarkably, it also surpasses the strongest teacher, **Llama-3.3-70B**, by +1.19 BLEU and +1.22 COMET, suggesting that human-like learning enables student models to surpass the performance limits of their teachers. For the smaller **Qwen3-4B** student, **SaKD (Multi-T)** achieves 32.48 BLEU and 87.57 COMET, improving over **KD (Single-T)** by +3.46 BLEU and +3.99 COMET. It also outperforms the reasoning-oriented teacher **DeepSeek-R1-70B** by +1.56 BLEU and +2.33 COMET, while approaching the performance of the strongest teacher, **Llama-3.3-70B**. These results demonstrate the effectiveness of SaKD even under limited model capacity.

Overall, while multi-teacher supervision is beneficial, SaKD exploits multiple teachers more effectively than traditional KD, resulting in substantially larger performance gains. Moreover, instruction-oriented teachers consistently outperform the reasoning-oriented teacher, suggesting that task-aligned supervision is more suitable for translation. Finally, SaKD also demonstrates clear advantages in the single-teacher setting, confirming that its effectiveness arises from the distillation mechanism itself rather than relying solely on multi-teacher signals.

[2]For SaKD (Single-T), we sample three candidate translations to perform *Reflection* and *Transcendence*.

Table 2: Results of human evaluation on the test set. * denotes statistically significant improvement in DA over the best teacher model at $p < 0.01$.

System	En ⇒ Zh			Zh ⇒ En			En ⇒ De			De ⇒ En		
	DA	CE↓	LE↓	DA	CE↓	LE↓	DA	CE↓	LE↓	DA	CE↓	LE↓
Qwen3-80B	87.1	10.3	12.1	86.2	11.1	13.3	84.6	12.3	13.8	85.5	13.1	14.9
Llama-3.3-70B	85.7	11.0	13.7	84.7	12.3	15.5	85.6	12.6	12.6	86.4	12.4	13.8
DeepSeek-R1-70B	83.7	13.0	16.0	83.8	15.3	19.3	82.1	14.8	15.1	82.8	15.9	16.9
SaKD (Multi-T)	88.7*	9.5	11.0	87.1*	10.5	9.3	86.9*	9.5	10.0	88.0*	10.5	12.3

Table 3: Ablation study for *SaKD* on En⇒Zh and En⇒De translation tasks.

Setting	Qwen3-8B				Qwen3-4B			
	En ⇒ Zh		En ⇒ De		En ⇒ Zh		En ⇒ De	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
SaKD (Multi-T)	34.45	89.21	32.23	88.09	32.56	87.22	31.45	87.21
w/o K. S.	33.56	88.31	31.35	87.21	32.23	86.89	30.85	86.61
w/o Reflec.	33.16	87.91	31.14	87.00	31.46	86.12	30.25	86.01
w/o Transc.	32.25	87.21	30.32	86.27	30.95	85.61	29.84	85.67

4.3 Human Evaluation

In this section, we conduct human evaluation to further assess translation quality. Specifically, we adopt Direct Assessment (DA) Graham et al. (2017) to evaluate translations generated by SaKD and baseline systems. We recruited professional translators with at least five years of experience as evaluators. Given a reference translation and system outputs, evaluators assigned a quality score ranging from 0 to 100. Details of the evaluation setup are provided in Appendix C. Additionally, we manually measure coherence errors (CE) and logical errors (LE) in long-document translation. CE refers to inter-sentential consistency issues, such as inconsistent translations of the same entity across sentences, while LE captures inter-sentential logical inconsistencies, including improper use or translation of transition words and conjunctions.

For each translation direction, we randomly sampled 50 documents (2,045 sentences in total). Two professional translators independently evaluated outputs from three teacher models and SaKD. As shown in Table 2, SaKD consistently achieves higher DA scores and fewer CE/LE errors, demonstrating its effectiveness in improving translation quality.

4.4 Ablation Study

To further investigate the contribution of each component in SaKD, we conduct ablation studies on two translation directions, English→Chinese and English→German. Specifically, we examine the effects of knowledge selection (K.S.), Reflection (Reflec.), and Transcendence (Transc.)process absent from SaKD. The results are shown in Table 3.

Overall, knowledge selection has a limited impact: removing it leads to small performance drops of 0.33–0.89 BLEU and 0.33–0.90 COMET across different settings, suggesting that SaKD remains robust to imperfect teacher signals. In contrast, removing the learning components causes substantially larger degradation. Eliminating Reflection reduces BLEU/COMET by 1.09–1.29/1.09–1.30 on Qwen3-8B, while removing Transcendence results in the largest declines of 1.91–2.20/1.82–2.00. This highlight that Transcendence is the key component enabling SaKD to achieve consistent improvements across different translation directions and model scales.

5 Discussion and Analysis

In this section, we provide more detailed analysis to elucidate how the proposed SaKD framework improves the translation quality upon the `Qwen3-8B` student model.

5.1 How Students Surpass Teachers?



Figure 3: Semantic distributions of source texts and model translations in the PCA space. Points represent individual samples, while contours summarize the outer density boundary of each distribution.

To analyze how the student model can surpass its teachers, we visualize whether the translations produced by the SaKD-trained student move beyond the performance boundary formed by the teacher models after the three-stage learning process. We sampled 100 documents from the source test set to compare the semantic characteristics of SaKD with the teacher model. Semantic embeddings were obtained by encoding translations and source documents with mBERT Devlin et al. (2018), averaging the last hidden states to generate fixed-length vector representations.

We project the embeddings into two dimensions using Principal Component Analysis (PCA), retaining the leading directions of variance. Figure 3 shows that SaKD translations align more closely with the source-text distribution while extending beyond teacher translation distributions in the projected space. This suggests that the student learns translation patterns beyond teacher imitation, providing qualitative support for the semantic fidelity improvements observed in our evaluations.

5.2 Behavioral Analysis of Learning Stages

We conduct a fine-grained human evaluation of the ablation variants along three dimensions: semantic completeness (Adequacy), lexical appropriateness (Appropriateness), and overall fluency (Fluency), to assess how translation quality evolves across learning stages.

Figure 4 shows that *Imitation* substantially improves Adequacy, helping the student preserve source semantics. *Reflection* and *Transcendence* yield more pronounced gains in Appropriateness and Fluency, refining word choice and translation naturalness. These findings support the intended progression of SaKD: imitation establishes semantic fidelity, while subsequent stages improve higher-order translation quality rather than merely reinforcing semantic coverage.

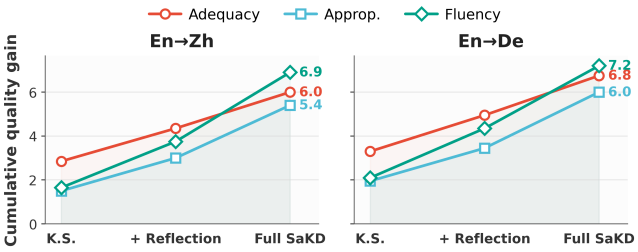


Figure 4: Cumulative human-evaluated quality gains on En⇒Zh and En⇒De across knowledge selection, Reflection, and the complete SaKD framework.

5.3 LLM-based Evaluation

To provide a more comprehensive assessment of the experimental results, we adopt an LLM-based comparative evaluation framework as a complement to traditional metrics such as COMET. While standard automatic metrics often struggle to capture fine-grained quality differences among multiple system outputs, the proposed evaluation approach instead treats `GPT-4o` as a relative judge, comparing competing translations to determine which output better reflects overall translation quality under long-text translation conditions.

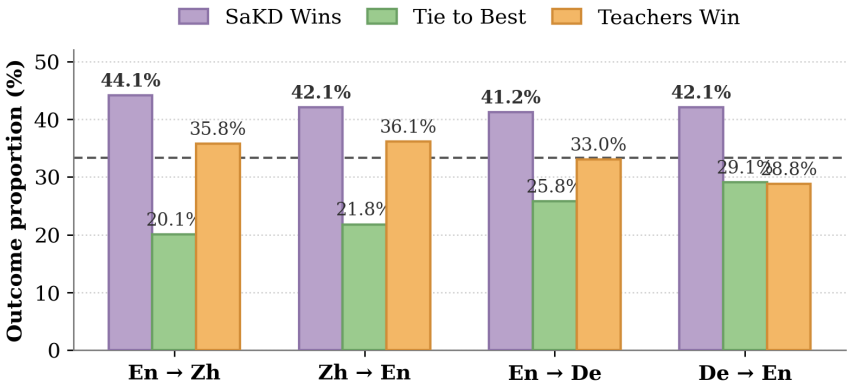


Figure 5: Grouped comparison of LLM-based evaluation outcomes across four translation directions. The dashed line indicates the equal-share baseline of 33.3%.

For each instance, we present the source document along with candidate translations and instruct `GPT-4o` to assign a score ranging from 1 to 10 to each translation. We then compute the win rate, defined as the percentage of instances in which a system receives the highest score, as a quantitative measure of comparative performance. As shown in Figure 5, SaKD achieves a win rate exceeding 44% on En⇒Zh, outperforming all teacher models. The results indicate that the student model trained with SaKD produces translations that are consistently preferred in head-to-head comparisons.

5.4 Analysis of Error Distribution

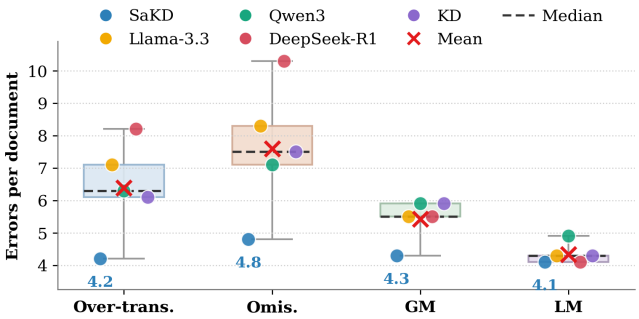


Figure 6: Visualization of translation errors. Colored bubbles denote systems; red crosses and dashed lines indicate means and medians, respectively. Lower values are better, and SaKD values are annotated.

In this section, we further investigate how SaKD enhances translation quality. To this end, we compared the average number of errors of different types across systems. Inspired by the success of LLM-as-a-judge work, we employed GPT-4o to evaluate translation errors, asking the model to classify them into four categories: over-translation (Over-trans.), omission (Omis.), grammar mistranslation (GM), and lexical mistranslation (LM). We then manually calculated the average number of errors per document and report the averages across four translation directions.

The results are visualized in Figure 6. As shown, the improvements achieved by SaKD primarily stem from a reduction in over-translation and omission errors. There are also modest improvements in grammar and lexical mistranslation. These findings demonstrate that SaKD has significant potential in mitigating translation hallucinations and underscore its superiority over traditional KD methods.

6 Conclusion and Future Work

Inspired by the progression of human learning, we introduce SaKD, a knowledge distillation framework that combines knowledge from multiple teachers with three stages of student learning: *Imitation*, *Reflection*, and *Transcendence*. Experiments on machine translation demonstrate that SaKD substantially outperforms conventional distillation methods and enables the student to surpass its teachers in translation quality. These findings highlight the value of moving beyond imitation to help students evaluate and refine teacher-provided knowledge. Future work will investigate whether this framework can deliver similar benefits on other natural language processing tasks.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation and manuscript polishing. The authors independently verified all code, technical claims, derivations, citations, and empirical results and take full responsibility for the final content of this work.

REFERENCES

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

BigScience. Bloom: A 176b-parameter open-access multilingual language model. *Computing Research Repository*, arXiv:2211.05100, 2022.

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

DeepSeekAI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. Mt-r1-zero: Advancing llm-based machine translation via r1-zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*, 2025a.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. Tear: Improving llm-based machine translation with systematic self-refinement. In *Findings of NAACL*, pp. 3922–3938, 2025b.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.

Google. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2022.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. Improving evaluation of document-level machine translation quality estimation. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of EACL*, pp. 356–361, 2017.

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

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025a.

Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Xiaoyu Chen, Zhanglin Wu, Huan Yang, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, et al. Align-then-slide: A complete evaluation framework for ultra-long document-level machine translation. *arXiv preprint arXiv:2509.03809*, 2025b.

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

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

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

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

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

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *Proceedings of ICLR*, 2021.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of EMNLP*, pp. 1317–1327, 2016.

KimiTeam. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

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

Jiahuan Li, Shanbo Cheng, Shujian Huang, and Jiajun Chen. Mt-patcher: Selective and extendable knowledge distillation from large language models for machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6445–6459, 2024.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, Hao Yang, and Min Zhang. DeMPT: Decoding-enhanced multi-phase prompt tuning for making LLMs be better context-aware translators. In *Proceedings of EMNLP*, pp. 20280–20295, 2024.

MetaAI. Llama 2: Open foundation and fine-tuned chat models. *Computing Research Repository*, arXiv:2307.09288, 2023.

OpenAI. Gpt-4 technical report. *Computing Research Repository*, arXiv:2303.08774, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pp. 311–318. Association for Computational Linguistics, 2002.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of EMNLP*, pp. 2685–2702, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sam Shleifer and Alexander M Rush. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*, 2020.

Yuncheng Song, Liang Ding, Changtong Zan, and Shujian Huang. Self-evolution knowledge distillation for llm-based machine translation. In *Proceedings of COLING*, pp. 10298–10308, 2025.

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

Menan Velayuthan, Nisansa De Silva, and Surangika Ranathunga. Encoder-aware sequence-level knowledge distillation for low-resource neural machine translation. In *Proceedings of LoResMT*, pp. 161–170, 2025.

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

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

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

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *Proceedings of ICML*, 2024a.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *Computing Research Repository*, arXiv:2401.11817, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*.

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

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

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

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

A Details of Datasets

During the knowledge generation stage, we utilize the OSCAR monolingual corpus, which contains document-level texts in 152 languages. From this corpus, we extract 3 million monolingual documents each in Chinese, German, and English, and employ the teacher model to generate parallel translation knowledge. After applying a filtering procedure, approximately 2 million high-quality parallel instances are retained. For the Imitation stage, we randomly sample 1.2 million instances for each translation task to construct the training set. In the Reflection and Transcendence stages, we sample 50K document-level parallel instances from the filtered knowledge for training. For the validation and test sets, we use high-quality document-level parallel translation data from Lyu et al. (2024). Detailed statistics of the dataset are provided in Table 4.

System	En ⇒ Zh		Zh ⇒ En		En ⇒ De		De ⇒ En	
	Sent.	*Doc.*	*Sent.*	*Doc.*	*Sent.*	*Doc.*	*Sent.*	*Doc.*
Training-*Imitation*	-	1,200,000	-	1,200,000	-	1,200,000	-	1,200,000
Training-*Reflection*	-	50,000	-	50,000	-	50,000	-	50,000
Training-*Transcendence*	-	50,000	-	50,000	-	50,000	-	50,000
Valid	6,061	150	6,061	150	5,866	150	5,866	150
Test	5,747	150	5,747	150	5,976	150	5,976	150

Table 4: Overview statistics of the training, validation, and test set for all translation tasks. *Sent.* and *Doc.* represent the counts of documents and sentences, respectively.

B Details of Training and Inference

During training, documents are not segmented into smaller units; instead, the model is trained directly on full-length documents. During the *Imitation* learning stage, we set the learning rate to 2e-5 and use a batch size of 128. The AdamW optimizer is employed, and the model is trained for 2 epochs with a linear warmup followed by decay for learning rate scheduling. Mixed precision with BF16 is enabled to reduce memory usage and accelerate training. For the *Reflection* stage, the learning rate is set to 1e-6, with a batch size of 64, and AdamW is again used. The model is trained for 10 epochs. During the *Transcendence* stage, we use a learning rate of 1e-6 and a KL coefficient of 0.04. For each source document, 16 translations are sampled, and the batch size is set to 8. Mixed precision with BF16 is also enabled to reduce memory usage and improve efficiency. All models are trained on NVIDIA 8×A800 GPUs.

During inference, we conduct both document-level and sentence-level translation, segmenting documents according to the golden sentences in the test set to generate predictions at each level.

C Details of Human Evaluation

We incorporate a **human evaluation** component to provide a more comprehensive assessment of translation quality. Specifically, we adopt **Direct Assessment (DA)** Graham et al. (2017) to evaluate translations produced by SaKD and its baseline systems. Evaluators are professional translators with at least five years of experience. For each source sentence, given a reference translation and corresponding outputs from multiple systems, evaluators assign a quality score ranging from 0 to 100.

The detailed scoring criteria are as follows:

- **0–20**: The translation is largely incorrect and unclear, with only a few words or phrases correct. It is essentially unreadable and difficult to comprehend.
- **21–40**: The translation conveys very little of the source semantics, with key information missing or incorrect. It contains numerous unnatural expressions, lacks fluency, and has significant grammatical errors.
- **41–60**: The translation expresses part of the key semantics but contains many non-key semantic errors. Fluency and idiomaticity are limited.
- **61–80**: The translation conveys the key semantics but includes some non-key information errors and noticeable grammatical mistakes. It lacks full idiomaticity.
- **81–100**: The translation accurately conveys the source semantics, with only minor non-key information errors and small grammatical issues. The output is fluent and idiomatic.

D Prompts Template

We present the translation prompt adopted in Figure 7 during translation generation by the teacher models and during the training of the student models.

Translation Prompt Template

You are a professional translation system. Please translate the following source text from <SOURCE LANGUAGE> to <TARGET LANGUAGE>. The translation should be accurate, fluent, and faithful to the original meaning. Preserve the original structure and discourse coherence. Do not add, omit, or explain any content. Output only the translation.

Source text:
<INSERT SOURCE TEXT HERE>

Translation:

Figure 7: Prompt used for translation generation of teacher models and training of student models.

E Impact of the Knowledge Filtering Threshold on Translation Performance

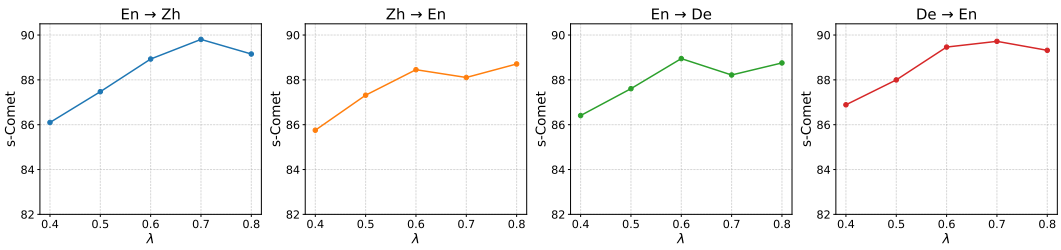


Figure 8: Effect of the hyperparameter λ on s-COMET scores across four translation directions. Each subplot corresponds to one translation direction, reporting the s-COMET score under different values of λ.

To analyze the impact of the hyperparameter λ in SaKD, we conduct a sensitivity study by varying λ from 0.4 to 0.8 and evaluating translation quality using the s-COMET metric. Figure 8 presents the results across four translation directions.

Overall, we observe a consistent performance improvement as λ increases, indicating that appropriately strengthening the distillation signal benefits the student model. In particular, s-COMET scores peak around $\lambda = 0.7$ or $\lambda = 0.8$ for most translation directions, suggesting a favorable balance between learning from teacher guidance and preserving the student model's own generation capacity. These results demonstrate that SaKD is relatively robust to the choice of λ, while still exhibiting clear performance gains under well-chosen hyperparameter settings.