複数の LLM を活用した機械翻訳のための協力デコーディング

白井尚登　衣川和尭　伊藤均　美野秀弥　河合吉彦
NHK 放送技術研究所
{shirai-n.hk,kinugawa.k-jg,itou.h-ce,mino.h-gq,kawai.y-lk}@nhk.or.jp

概要

大規模言語モデル（LLM）は対話などの生成タスクで成功する一方，翻訳性能の向上や計算コストの高さに課題がある．そこで本論文では，これらの課題を解決するために機械翻訳に適した新たな協力デコーディング手法を提案する．本手法は各言語の単言語データの継続事前学習をしたのち，対訳データによる追加学習をした小型 LLM と，より大型な LLM を組み合わせて翻訳を行う．日英・独英翻訳タスクについて双方向で実験した結果，提案手法はパラメータサイズの小さなモデルを学習することで計算コストを抑え, 既存の協力デコーディング手法を上回る翻訳性能を示した．

1　はじめに

GPT [1] や Llama [2] シリーズを筆頭に，大規模言語モデル (Large Language Model, LLM) は大量のテキストを事前学習することで高い言語能力を有している．さらに，LLM への継続的な学習によって特定の言語や知識などに適応する能力 [3] や，インストラクションデータを用いた学習によって対話，翻訳，要約などの生成能力 [4] を獲得できることが報告されている．

機械翻訳分野においても LLM を活用する動きがあり，継続事前学習と追加学習を組み合わせた ALMA [5] や翻訳のワークフローを LLM に再現させた Tower [6] が提案されている．

こうした LLM を活用した機械翻訳の課題として，特にパラメータサイズの小さい LLM (小型 LLM) は従来のニューラル機械翻訳に比べると精度が低いことが指摘されている [5]．また，パラメータ数の大きな LLM (大型 LLM）は継続事前学習，追加学習，推論にかかる計算コストが膨大となり，効率的な学習と推論方法の確立が求められている．

本研究では，大型 LLM を再学習せず，小型 LLM を活用するプロキシチューニング [7] に着目し，継続事前学習と追加学習の計算コストを抑えた上で，汎用的な大型 LLM の推論の補正によって翻訳性能の向上を目指す．

提案手法（**Collab-MT**）では，機械翻訳タスクにおいて，対象となる言語で再学習した 2 つの小型 LLM の推論結果を用いて大型 LLM の生成確率を補正する．2 つの小型 LLM の再学習には，原言語と目的言語それぞれの単言語データで継続事前学習をしたのち，対訳データで追加学習を行う．そして，デコーディングの各ステップで，2 つの小型 LLM のソフトマックス関数を用いる前の出力値を大型 LLM の出力値に加算することでトークンの生成確率を補正する．日英・独英翻訳について双方向で実験を行った結果，小型 LLM の再学習によって計算コストを抑え，既存の協力デコーディング手法 [7] に比べて，どの翻訳タスクも BLEU スコアが最低 1.93 ポイント向上した．

2　関連研究

LLM を活用した機械翻訳の研究． LLM が対話などの生成タスクで成功している一方, 機械翻訳分野では従来のエンコーダ・デコーダ型の NLLB [8] が小型 LLM に対して最先端の精度を示している．こうした LLM の機械翻訳性能の課題に対し，Xu ら [5] は Llama-2 にモノリンガルコーパスによる継続事前学習と少量の対訳データで追加学習する ALMA を提案した．ALMA は Llama-2 のゼロショットや NLLB-54B に対して精度を上回った．

複数 LLM を活用した研究． 単一の汎用 LLM はパラメータサイズを大きくしたとしても特定のタスクに特化したモデルの性能を上回ることは難しい [9]．また，LLM のパラメータサイズに応じて計算コストが高くなる．これらの解決策として複数モデルを組み合わせる協力デコーディングと呼ばれる手法が提案されている．協力コーディングの一例として，Liu ら [7] は小型 LLM から大型 LLM への知識の転移を目的に，タスクに特化した小型 LLM を利用し

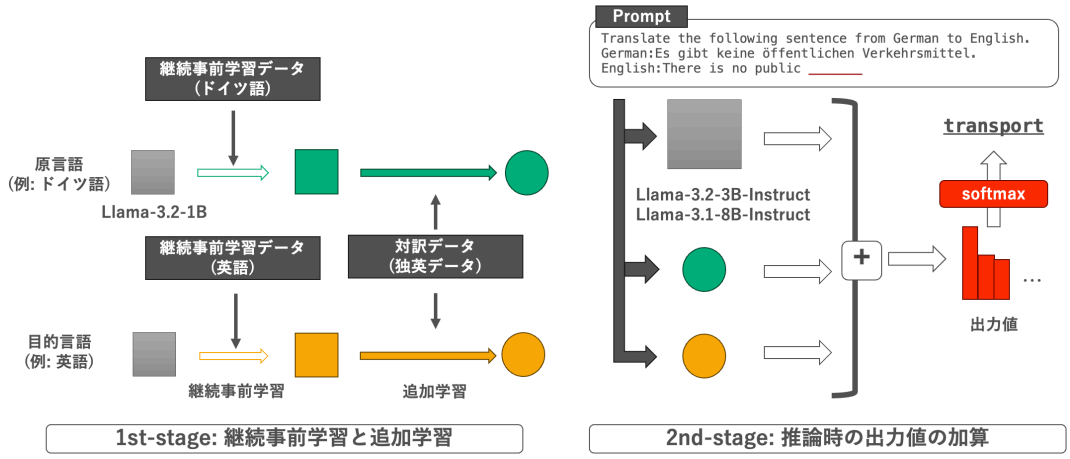

図 1　提案手法 Collab-MT の概略図．複数の小型 LLM に対して継続事前学習と追加学習を行い，デコーディング時に大型 LLM との出力値の加算をし，推論の補正をする．

て大型 LLM のデコーディング時の予測を補正するプロキシチューニングを提案した．

3　提案手法

3.1　プロキシチューニング

Liu らの提案したプロキシチューニング [7] は，デコーディング時のソフトマックス関数を用いる前の出力値に注目し，大型 LLM の出力を再学習した小型 LLM によって補正する手法である．本手法は，はじめに，小型 LLM $S_{\mathcal{M}}$ を下流タスクでインストラクションチューニングしたモデル $S_{\mathcal{M}^+}$ を構築する．そして，デコーディング時に小型 LLM の追加学習前後の出力値の差分を大型 LLM $L_{\mathcal{M}}$ に加算することで，直接追加学習をした大型 LLM $L_{\mathcal{M}^+}$ の出力値に近づける．以上から，デコーディングの各ステップ t のプロキシチューニングの確率分布 $p_{L_{\tilde{\mathcal{M}}^+}}(X_t \mid x_{<t})$ は，式 1 となる．

$$p_{L_{\tilde{\mathcal{M}}^+}}(X_t \mid x_{<t}) = \mathrm{softmax}[L_{\mathcal{M}}(X_t \mid x_{<t}) + \underbrace{S_{\mathcal{M}^+}(X_t \mid x_{<t}) - S_{\mathcal{M}}(X_t \mid x_{<t})}_{\text{小型 LLM の出力値の差分}}] \quad (1)$$

Liu らの実験によると，インストラクションチューニング前後の小型 LLM の出力値の差分を大型 LLM に加算することで，再学習をしていない大型 LLM の汎用的な能力とともに対話形式の生成を可能とした．そして，質問応答やコードのタスクで直接大型 LLM を追加学習したモデルに匹敵する性能を示した．

3.2　提案手法: Collab-MT

本論文では，大型 LLM と，継続事前学習と追加学習を行なった 2 つの小型 LLM による機械翻訳のための協力デコーディング手法 **Collabo**rative Decoding for **M**achine **T**ranslation (**Collab-MT**) を提案する (図 1)．Collab-MT は小型 LLM の学習前後の差分を大型 LLM に加算するプロキシチューニングと異なり，継続事前学習と追加学習をした小型 LLM の出力値を大型 LLM に加算するのみで出力を補正する．既存手法に対して，複数モデルの加算のみをする提案手法によって，各言語のドメインに特化したモデルの補正による翻訳性能の改善（§ 6）や，機械翻訳タスクにおける差分の必要性（§ 7）を検証する．

Collab-MT は以下の二段階のアプローチである．

1st-stage: 継続事前学習と追加学習． 1st-stage では，翻訳タスクの少量の単言語データで継続事前学習をし，対訳データで追加学習をすることで機械翻訳に適応した LLM を構築する．本実験では，原言語側に特化した機械翻訳モデル $S_{\mathcal{M}^{\text{SRC-MT}}}$ と目的言語側に特化した機械翻訳モデル $S_{\mathcal{M}^{\text{TGT-MT}}}$ を構築する．

2nd-stage: 推論時の出力値の加算． 2nd-stage では，追加学習した機械翻訳モデルの出力によって，再学習をしていない大型モデル $L_{\mathcal{M}}$ の予測を補正し，直接追加学習した大型モデル $L_{\mathcal{M}^{\text{MT}}}$ に近づける協力デコーディングを行う．以上から Collab-MT の確率分布 $p_{L_{\tilde{\mathcal{M}}^{\text{MT}}}}(X_t \mid x_{<t})$ は式 2 となる．

$$p_{L_{\tilde{\mathcal{M}}^{\text{MT}}}}(X_t \mid x_{<t}) = \mathrm{softmax}[L_{\mathcal{M}}(X_t \mid x_{<t}) + S_{\mathcal{M}^{\text{SRC-MT}}}(X_t \mid x_{<t}) + S_{\mathcal{M}^{\text{TGT-MT}}}(X_t \mid x_{<t})] \quad (2)$$

表 1　本実験で使用したデータセット．

データセット	継続事前学習データ	対訳データ	テストデータ
ALT (en-ja)	1000	18083	1017
WMT19 (en-de)	2000	18000	2998

4　実験設定

4.1　実験概要

本実験では，日英・独英について双方向で翻訳実験を行うことで提案手法と既存の協力デコーディング手法の翻訳性能の差を調査する．

4.2　使用モデル

本実験では，小型 LLM としてパラメータサイズが 1B である Llama-3.2-1B を使用し，より大型な LLM にはパラメータサイズが 3B である Llama-3.2-3B-Instruct とパラメータサイズが 8B である Llama-3.1-8B-Instruct を使用する [1]．既存研究におけるプロキシチューニングの設定とは異なり，大型 LLM には Instruct モデルを使用する．

4.3　モデルの学習

モデルの学習には torchtune [2] を利用し，プロキシチューニングのコード [3] を参考に実験を行なった．モデルの学習率は 2e-5 に設定し，継続事前学習は 5 エポック，対訳データによる学習は 3 エポックとした．

なお，学習および推論時のプロンプトは Llama3 のフォーマット [4] に従った (Appendix 表 5 参照)．

5　評価方法

5.1　データセット

提案手法の翻訳性能を調べるため，日英・独英双方向の翻訳タスクで評価する．日英・英日の翻訳タスクには多言語 Wikinews データを対訳処理した ALT [10] を使用する [5]．また，独英・英独の翻訳タスクには WMT19 [11] を使用した [6]．

本実験では既存の対訳データのみを使用し，少量の単言語データの継続事前学習と対訳データの追加学習を試みた．そのため，ALT データの検証用の 1000 文の各言語を継続事前学習データ，学習データの 18083 文を対訳データとして扱った．同様に WMT19 においても学習時間の削減と少量データによる翻訳性能の向上を目的に，学習データからランダムシードで 20000 文を抽出し，2000 文を継続事前学習時のデータ，残りを対訳データとして活用した．そして，検証データをテストデータとした．

5.2　評価方法

翻訳の評価方法として，n グラム単位で生成文と参照文の表層的な一致を評価する BLEU，および，文脈埋め込みを用いてトークンの類似度で評価する BERTScore [12] を使用する．BLEU スコアには SacreBLEU [13] を使用し，BERTScore は各文章の F1 スコアの平均をとる macro-F_1 スコアを算出した．

5.3　推論モデルの比較

本論文では下記の 4 つの推論モデルを比較する．

1.Original HuggingFace で提供されているモデル．

2.Fine-tuning 対訳データで追加学習したモデル．

3.Proxy-tuning プロキシチューニングを用いたモデル．対訳データによる追加学習前後の Llama-3.2-1B の小型 LLM の差分を大型 LLM に加算し，推論を補正する．

4.Collab-MT 提案手法を用いたモデル．原言語あるいは目的言語による継続事前学習と対訳データによる追加学習を行なった Llama-3.2-1B の小型 LLM の出力値を大型 LLM に加算することで推論を補正する．

なお，8B の Proxy-tuning, Collab-MT とは，大型 LLM が Llama-3.1-8B のモデルであることを指す．

6　結果

日英・独英翻訳実験の結果を表 2 に示す．Collab-MT はどの翻訳タスクにおいても既存手法のプロキシチューニング (Proxy-tuning) の BLEU スコアを最低 1.93 ポイント上回り，英日翻訳タスクを除いて 6.4 ポイント以上の向上をみせた．また，BERTScore においても提案手法が既存手法を上回る結果となった．

しかし，再学習をしていない大型 LLM (Original) との翻訳性能を比較すると，提案手法は日英・英日翻訳タスクでは翻訳性能が同等であるのに対し，独英・英独翻訳タスクでは最大 6.12 ポイント下回る結

1) https://huggingface.co/meta-llama
2) https://github.com/pytorch/torchtune
3) https://github.com/alisawuffles/proxy-tuning
4) https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3/
5) https://huggingface.co/datasets/mutiyama/alt
6) https://huggingface.co/datasets/wmt/wmt19

表 2　翻訳実験の結果．Llama-3.2-3B の Collab-MT とは，継続事前学習と追加学習を行なった Llama-3.2-1B の 2 つの小型 LLM の出力値を Llama-3.2-3B の大型 LLM に加算して推論するモデルのことを指す．

サイズ	推論モデル	ja-en	en-ja	de-en	en-de
		BLEU / BERTScore	BLEU / BERTScore	BLEU / BERTScore	BLEU / BERTScore
1B	Original	15.80 / 0.92	3.39 / 0.78	31.91 / 0.94	23.19 / 0.84
	Fine-tuning	20.54 / 0.93	7.95 / 0.85	31.37 / 0.94	23.67 / 0.85
3B	Original	24.40 / 0.94	4.67 /0.83	39.64 / 0.95	31.69 / 0.87
	Fine-tuning	30.08 / 0.95	8.76 / 0.86	39.53 / 0.96	31.17 / 0.87
	Proxy-tuning	20.26 / 0.93	5.35 / 0.81	20.15 / 0.93	17.34 / 0.80
	Collab-MT	26.73 / 0.94	7.41 / 0.85	38.02 / 0.95	30.30 / 0.87
8B	Original	28.87 / 0.95	7.76 / 0.84	44.75 / 0.96	37.90 / 0.89
	Fine-tuning	33.07 / 0.95	9.15 / 0.87	43.18 / 0.96	35.52 / 0.88
	Proxy-tuning	21.56 / 0.93	6.03 / 0.83	26.45 / 0.94	21.15 / 0.82
	Collab-MT	28.09 / 0.94	7.96 / 0.85	39.18 / 0.95	31.78 / 0.87

表 3　日英翻訳タスクの翻訳結果の一例．

サイズ		テキスト
	原言語文	シドニーのランドウィック競馬場の 8 頭のサラブレッド競走馬が馬インフルエンザに感染していることが確認された。
	目的言語文	It has been confirmed that eight thoroughbred race horses at Randwick Racecourse in Sydney have been infected with equine **influenza**.
1B	Fine-tuning	Eight thoroughbred racing horses at the Sydney Landmark have been confirmed to have contracted the **flu**.
3B	Original	Eight thoroughbred horses at Randwick Racecourse in Sydney have been confirmed to be infected with equine **influenza**.
	Fine-tuning	Eight thoroughbred horses at the Randwick Racecourse in Sydney have been confirmed to have contracted equine **influenza**.
	Proxy-tuning	Eight Sydney racing thoroughbreds have been confirmed to be infected with **E**.
	Collab-MT	Eight horses at the Sydney Royal Randwick racecourse have been confirmed to have contracted equine **influenza**.

表 4　日英・英日翻訳タスクにおける使用モデルの BLEU スコアの比較結果．

サイズ	推論モデル	式	ja-en	en-ja
1B	Original	$S_{\mathcal{M}}$	15.80	3.39
	Fine-tuning	$S_{\mathcal{M}^{\mathrm{MT}}}$	20.54	7.95
3B	Original	$L_{\mathcal{M}}$	24.40	4.67
	Fine-tuning	$L_{\mathcal{M}^{\mathrm{MT}}}$	30.08	8.76
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\mathrm{MT}}} - S_{\mathcal{M}})$	20.26	5.35
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\mathrm{SRC\text{-}MT}}} - S_{\mathcal{M}})$	19.70	6.18
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\mathrm{TGT\text{-}MT}}} - S_{\mathcal{M}})$	20.31	6.27
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\mathrm{SRC\text{-}MT}}} + S_{\mathcal{M}^{\mathrm{TGT\text{-}MT}}}$	26.73	7.41

果となった．また，対訳データの追加学習のみのモデル (Fine-tuning) に対しては，すべて下回った．他方で Fine-tuning の独英・英独翻訳タスクの BLEU スコアの結果が Original を下回る．このことから，本実験の継続事前学習や追加学習の設定では効果的に LLM を学習できていないことが示唆される．

7　分析

7.1　翻訳結果の例

表 3 は日英・英日翻訳タスクの出力例である．他のモデルに対して Llama-3.2-1B の Fine-tuning モデルと Llama-3.2-3B の Proxy-tuning モデルは文末の “influenza” を生成できていない．

7.2　異なるモデルの既存手法との比較

継続事前学習と追加学習を行なった Collab-MT 用の機械翻訳器でプロキシチューニングの比較実験を行なった．その結果，Collab-MT はどのプロキシチューニングよりも翻訳性能が高かった (表 4)．この結果から，本論文で行なった翻訳タスクの実験においては，小型 LLM の出力値の差分を取るよりも加算のみを行う方が翻訳性能は高い傾向にある (Appendix 表 6 参照)．

8　おわりに

本論文では，機械翻訳のための協力デコーディング手法 Collab-MT を提案した．本手法は大型 LLM に継続事前学習と追加学習をした 2 つの小型 LLM の出力値を合算し，推論の補正を行う．日英・独英の双方向で翻訳実験を行った結果，より小型な LLM の再学習によって計算コストを抑え，英日翻訳タスクを除き，既存の協力デコーディング手法よりも 6.4 ポイント以上の BLEU スコアの上昇がみられた．しかしながら，提案手法はより大型な LLM の再学習前後よりも翻訳性能が落ちている．そのため，機械翻訳のための LLM の学習方法と計算効率の良いデコーディング技術の確立が課題である．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究（課題 225）により得られたものです．

参考文献

[1] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[3] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. **arXiv preprint arXiv:2404.17790**, 2024.

[4] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. **arXiv preprint arXiv:2308.10792**, 2023.

[5] 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.

[6] Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. Tower: An open multilingual large language model for translation-related tasks. **arXiv preprint arXiv:2402.17733**, 2024.

[7] Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy, 2024.

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

[9] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. **arXiv preprint arXiv:2407.06089**, 2024.

[10] Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. Introduction of the asian language treebank. In **2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)**, pp. 1–6. IEEE, 2016.

[11] Wikimedia Foundation. Acl 2019 fourth conference on machine translation (wmt19), shared task: Machine translation of news.

[12] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. **CoRR**, Vol. abs/1904.09675, , 2019.

[13] Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

A　プロンプトの例

表 5　日英翻訳タスクにおけるプロンプトの一例.

<|begin_of_text|><|start_header_id|>system <|end_header_id |>
Translate the following sentence from Japanese to English. <|eot_id |><|start_header_id|>user<|end_header_id|>
Japanese: シドニーのランドウィック競馬場の 8 頭のサラブレッド競走馬が馬インフルエンザに感染していることが確認された。
English:<|eot_id |>
<|start_header_id|>assistant <|end_header_id|>

B　プロキシチューニングと Collab-MT の比較実験

表 6　異なる小型 LLM で加算と差分を実験した際の BLEU スコアの比較結果．太字はプロキシチューニングと加算のみの Collab-MT を比較した中で，一番スコアが高いのものである．モデルの変更や組み合わせによってプロキシチューニングの性能は改善したが，加算のみの Collab-MT の方が翻訳性能は高い傾向にある．

サイズ	推論モデル	式	ja-en	en-ja	de-en	en-de
1B	Original	$S_{\mathcal{M}}$	15.80	3.39	31.91	23.19
	Fine-tuning	$S_{\mathcal{M}^{\text{MT}}}$	20.54	7.95	31.37	23.67
3B	Original	$L_{\mathcal{M}}$	24.40	4.67	39.64	31.69
	Fine-tuning	$L_{\mathcal{M}^{\text{MT}}}$	30.08	8.76	39.53	31.17
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\text{MT}}} - S_{\mathcal{M}})$	20.26	5.35	20.15	17.34
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\text{SRC-MT}}} - S_{\mathcal{M}})$	19.70	6.18	28.82	27.87
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\text{TGT-MT}}} - S_{\mathcal{M}})$	20.31	6.27	29.74	27.44
	Proxy-tuning	$L_{\mathcal{M}} + (0.5 * S_{\mathcal{M}^{\text{SRC-MT}}} + 0.5 * S_{\mathcal{M}^{\text{TGT-MT}}} - S_{\mathcal{M}})$	20.73	6.01	30.54	28.20
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{MT}}}$	26.33	8.81	38.96	32.20
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{SRC-MT}}}$	27.20	7.84	39.24	32.37
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{TGT-MT}}}$	26.96	7.63	39.49	32.24
	Collab-MT	$L_{\mathcal{M}} + 0.5 * S_{\mathcal{M}^{\text{SRC-MT}}} + 0.5 * S_{\mathcal{M}^{\text{TGT-MT}}}$	27.00	7.48	39.50	32.45
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{SRC-MT}}} + S_{\mathcal{M}^{\text{TGT-MT}}}$	26.73	7.41	38.02	30.30
8B	Original	$L_{\mathcal{M}}$	28.87	7.76	44.75	37.90
	Fine-tuning	$L_{\mathcal{M}^{\text{MT}}}$	33.07	9.15	43.18	35.52
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\text{MT}}} - S_{\mathcal{M}})$	21.56	6.03	26.45	21.15
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\text{SRC-MT}}} - S_{\mathcal{M}})$	22.38	6.79	33.14	30.69
	Proxy-tuning	$L_{\mathcal{M}} + (S_{\mathcal{M}^{\text{TGT-MT}}} - S_{\mathcal{M}})$	22.78	6.47	32.24	30.06
	Proxy-tuning	$L_{\mathcal{M}} + (0.5 * S_{\mathcal{M}^{\text{SRC-MT}}} + 0.5 * S_{\mathcal{M}^{\text{TGT-MT}}} - S_{\mathcal{M}})$	22.67	6.39	33.79	30.87
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{MT}}}$	27.50	**9.46**	41.39	34.93
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{SRC-MT}}}$	**29.54**	8.27	41.57	**35.03**
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{TGT-MT}}}$	29.47	8.43	41.71	34.80
	Collab-MT	$L_{\mathcal{M}} + 0.5 * S_{\mathcal{M}^{\text{SRC-MT}}} + 0.5 * S_{\mathcal{M}^{\text{TGT-MT}}}$	29.46	8.44	**41.87**	**35.03**
	Collab-MT	$L_{\mathcal{M}} + S_{\mathcal{M}^{\text{SRC-MT}}} + S_{\mathcal{M}^{\text{TGT-MT}}}$	28.09	7.96	39.18	31.78