大規模言語モデルのタスク特化ドメイン適応における知識獲得効率に関する初期検討

小林 和馬 [1,2]　相澤 彰子 [1]
[1] 国立情報学研究所　[2] 国立がん研究センター研究所
{kazumkob,aizawa}@nii.ac.jp

概要

医療分野は国ごとの制度や慣習の違いが大きく、事前学習済み大規模言語モデルに対してドメイン固有の知識を効率的に習得させる学習戦略が重要となる。本研究では、英日医学翻訳を対象タスクとして設定し、英語の医学用語から適切な日本語の医学用語を想起する能力を、当該タスクにおけるドメイン知識として定義した。これに基づき、医学専門用語の変換精度を定量的に評価するための知識プロービング・ベンチマークを構築した。続いて、Qwen-2.5 ファミリーのベースモデルに対して、英日医学対訳コーパスを用いた継続事前学習と教師ありファインチューニングを異なる比率で実施し、ドメイン知識獲得の観点から最適な学習戦略を検証した。

1　はじめに

事前学習済み大規模言語モデル（PLM: Pre-trained Large Language Model）の追加学習において、事前学習データセットに十分に含まれていない低資源言語や、それらの言語における特定ドメインの知識を効率的に獲得させることは、産業応用における重要な技術的課題である。例えば医療分野では、専門家間の学術的なコミュニケーションの大部分が英語で行われており、事前学習に利用可能な公開医学文献の大部分は英語である一方、日本語文献は極めて限定的である。また、医療分野は国ごとの制度や慣習の違いが大きく、高度な専門知識の正確な運用が不可欠である。このように、低資源ドメインに対しても、そのドメイン固有の知識を効率的に獲得させるための学習戦略が求められている。

PLM の追加学習においては、様々な手法が提案されているが、特に知識獲得の観点で重要なものは継続事前学習と教師ありファインチューニングである。ドメイン適応を目的とした**継続事前学習**（CDP: Continual Domain-adaptive Pretraining）は、事前学習と同様に次トークン予測による自己教師あり学習を行うことで、対象ドメインの新たな知識獲得を図る手法である [1]。一方、**教師ありファインチューニング**（SFT: Supervised Fine-tuning）は指示チューニングとも呼ばれ、新規タスクの学習過程で新たな知識も同時に習得することが期待される [2]。その他の追加学習手法として、人間の意図に沿った適切なテキスト生成を実現するための選好チューニングがある。また、パラメータ更新を伴わないものの、外部知識ベースを活用して新たな知識の利用を可能とする検索拡張生成も広く利用されている。

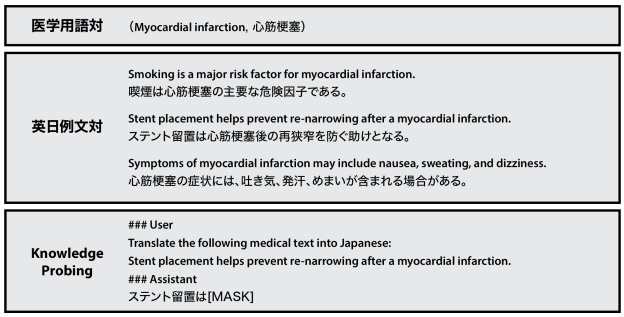

図 1　英日医学翻訳における知識プロービング: 英語医学用語と日本語医学用語からなる医学用語対に対して、それぞれの医学用語が含まれる英日例文対からなる知識プロービング・ベンチマークを構築した。英日例文対の英語テキストを入力とし、対応する日本語テキスト中の日本語医学用語を表す最初のトークンの出力確率を計測することで、英語医学用語から日本語医学用語への対応付け能力を表す指標として用いることができる。

こうした様々な追加学習手法は排他的ではなく、典型的には継続事前学習の後に教師ありファインチューニングを行うという段階的な学習戦略が採用されてきた。この過程において、継続事前学習によって新たな知識を獲得し、教師ありファインチューニングでタスクに応じた入出力形式を学習するという役割分担が期待されてきた。しかし、教師ありファインチューニングではデータ量過多が学習性能に悪影響を与えるという報告がある一方 [3]、

継続事前学習では既存知識の破滅的忘却などの副作用も懸念される。このような状況において、両手法を組み合わせる際の最適な比率に関する知見の蓄積が求められていた。

本研究では、英日医学翻訳を対象タスクとして設定し、英語の医学用語を適切な日本語の医学用語に変換する能力を、当該タスクのドメイン知識として定義した。これに基づき、医学専門用語の変換精度を定量的に評価するための**知識プロービング・ベンチマーク**（Knowledge Probing Benchmark）を構築した（**図 1**）。そして、Qwen-2.5 ファミリーのベースモデルに対して、英日医学対訳コーパスを用いた継続事前学習と教師ありファインチューニングを異なる比率で実施し、ドメイン知識獲得の観点から最適な学習戦略を検証した。その結果、教師ありファインチューニングのみによる学習であっても、ドメイン知識を効率的に獲得できることが示唆された。

2 関連研究

2.1 継続事前学習

ドメイン適応のための継続事前学習の効果は、追加の学習データだけでなく、事前学習データに含まれていた情報量、モデルサイズ、新規タスクと既存タスクの類似度、既存知識の忘却など、様々な要因の影響を受ける [1]。特に、事前学習データに頻出していた知識は獲得されやすい一方で、継続事前学習で初めて登場する希少な知識（Long-tail Knowledge）については、十分な習得が困難であることが報告されている [4]。また、従来のテキストコーパスの代わりに、特定のタスクに応じて指示形式に整形した入力テキストを用いる学習（Instruction Pre-training）によって、性能が向上することも報告されている [5]。

2.2 教師ありファインチューニング

教師ありファインチューニングは次トークン予測に基づく学習を行うが、継続事前学習とは異なり、入力プロンプト部分では損失を計算せず、モデルの出力部分でのみ損失を算出する [2]。従来、教師ありファインチューニングはタスクに応じた入出力形式への適応が主目的とされ、少数の高品質なデータで十分とされてきた [6]。しかし、その目的を新たな知識の獲得とした場合、特定のドメイン知識の新規獲得には不十分であることや [7]、継続事前学習と同様に希少な知識の習得効率が低いことが報告されている [8, 4]。さらに、事前学習データに全く出現しなかった新たな知識を導入する際には、ハルシネーションのリスクが伴うことも知られている [9]。

2.3 知識プロービング

PLM が学習段階で獲得し、内部知識として保持した情報を適切に想起（recall）する能力を評価することは、依然として困難な課題である。このような評価手法として、LAMA ベンチマークにおいて知識プロービングが初めて提案された [10]。この手法では、Wikipedia などの外部知識ベースから抽出したトリプレット（Subject, Relation, Object）を基に、“Italy is located in [MASK].” のような穴埋め問題を PLM に解かせることで、事実に関する知識の有無を検証できる。近年では、多様な大規模プロンプトを生成することで、PLM の内部知識をより正確かつ包括的に評価するためのデータセットも提案されている [11]。

3 実験

本研究では、英日医学翻訳を対象タスクとした。このタスクの正確な遂行には、一般的な英文法の理解に加え、英語の医学用語を適切な日本語の医学用語に変換する能力が必要である。後者の能力は、PLM が「[英語医学用語] の日本語訳は [日本語医学用語] である」という知識を内部に保持しているものと解釈できる。そこで、英日医学用語の単語レベルでの対応付け能力を、当該タスクにおけるドメイン知識として定義した。

3.1 英日医学対訳コーパス

インターネット上の公開医学教科書および医学論文抄録から日本語テキストを収集し、必要に応じて `Qwen2.5-32B-Instruct` を用いて英語訳を付与することで、英日医学対訳コーパスを構築した。このコーパスをランダムに分割し、**追加学習用データセット**（80,000 サンプル、英日 634,493K 文字）と**評価用データセット**（300 サンプル、英日 241K 文字）を作成した。これらのサンプルの大部分は、パラグラフ単位の英日医学テキストの対からなる（追加学習用データの例は **Appendix A** を参照）。

3.2 追加学習と比率設定

継続事前学習と教師ありファインチューニングでは、英語の医学テキストを入力として日本語の医

学テキストを生成する共通のプロンプトを使用した。追加学習用データセット（80,000 サンプル）を複数の比率で両学習に分割した。分割時には元データセットの順序を維持した。メモリ使用量の削減と学習効率の向上のため、QLoRA[12, 13] を用い、両学習で同一のハイパーパラメータを設定した（詳細な学習条件は **Appendix A** を参照）。

3.3 知識プロービング・ベンチマーク

本研究におけるドメイン知識は、英語医学用語から日本語医学用語への対応付け能力と定義する。この知識を評価するための知識プロービング・ベンチマークを、以下のように構築した（**図 1**）。(1) まず英日医学辞典から医学用語の対応表を抽出し、追加学習用データセットにおいて、英語テキストと日本語テキストの両方に出現する医学用語対を特定した。(2) これらの医学用語対を出現頻度に基づき、低頻度（1-9 回）、中頻度（10-99 回）、高頻度（100 回以上）に分類した。(3) 次に、各頻度帯の医学用語対について、`gpt-4-2024-11-20` を用いて英語例文を生成し、対応する日本語医学用語を含む日本語例文を作成した。(4) これにより、低頻度、中頻度、高頻度の各頻度帯から 20 組の医学用語対を選定し、各医学用語対につき 20 組の英日例文対を含む知識プロービング・ベンチマークを構築した（知識プローブの例は **Appendix B** を参照）。

3.4 ドメイン知識の評価方法

知識プロービング・ベンチマークを用いたドメイン知識の評価方法は、以下のように定式化される。(1) 一つの医学用語対 $p_i = (t_{\mathrm{E}_i}, t_{\mathrm{J}_i})$ は、英語医学用語 t_{E_i} と日本語医学用語 t_{J_i} からなる。(2) 各医学用語対 p_i には 20 組の英日例文対 $\mathcal{S}_i = \{(s_{\mathrm{E}_{i1}}, s_{\mathrm{J}_{i1}}), \ldots, (s_{\mathrm{E}_{i20}}, s_{\mathrm{J}_{i20}})\}$ が対応付けられる。(3) 英日例文対の英語テキスト $s_{\mathrm{E}_{ij}}$ をモデルへの入力とした際、対応する日本語テキスト $s_{\mathrm{J}_{ij}}$ 中の日本語医学用語 t_{J_i} を表す最初のトークンの出力確率 p_{ij} を計算する。(4) 具体的には、位置 N に出現する日本語医学用語 t_{J_i} を表す最初のトークンに対して、位置 $N-1$ のロジットから softmax を計算し、当該トークン ID に対応する確率を算出する。(4) 各医学用語対 p_i について、20 組の英日例文対 $\mathcal{S}_i$ から算出された出力確率の平均 $\bar{p}_i$ を、その医学用語対における英語医学用語から日本語医学用語への対応付け能力を表す指標として解釈する。(5) 最終的に、低頻度、中頻度、高頻度の各頻度帯に含まれる 20 組の医学用語対について、個々に算出された出力確率の平均 $\bar{p}_i$ をさらに平均し、これを **KP スコア**（Knowledge Probing Score）と定義する。この KP スコアに基づいて、医学用語対の出現頻度に応じたドメイン知識習得効率を評価する。

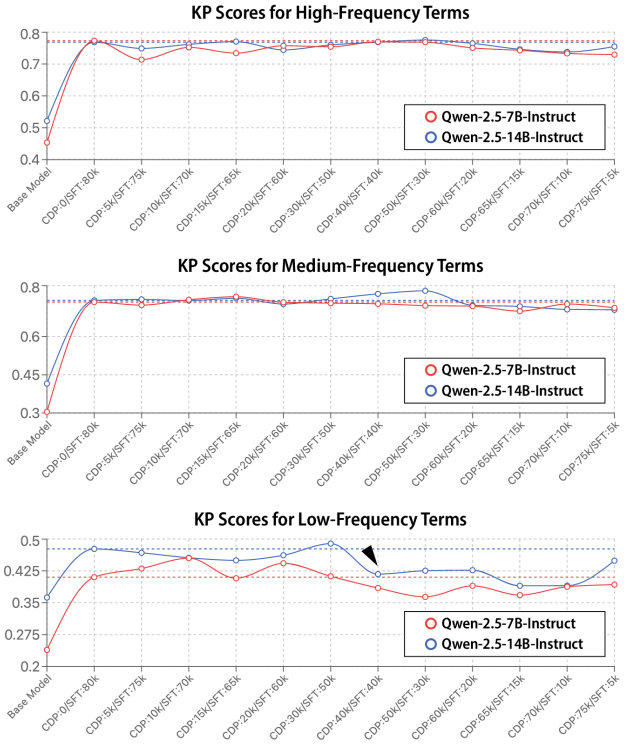

図 2　頻度帯ごとの KP スコア: `Qwen2.5-7B-Instruct` と `Qwen2.5-14B-Instruct` について、追加学習を行っていないベースモデルと、異なる比率で継続事前学習（CDP）および/または教師ありファインチューニング（SFT）を実施したモデルの KP スコアを、高頻度（上段）、中頻度（中段）、低頻度（下段）の各頻度帯で算出した。横軸には、ベースモデルを始点とし、SFT のみを実施したモデル（CDP:0/SFT:80k）から段階的に CDP の比率を高めたモデルを配置している。例えば、CDP:30k/SFT:50k は、全 80,000 サンプルの追加学習用データのうち、前半の 30,000 サンプルを CDP に、後半の 50,000 サンプルを SFT に使用したことを表す。これにより、SFT のみを実施したモデルの KP スコア（点線）と比較して、CDP を実施することによる知識獲得効率を評価することができる。特に低頻度帯では、継続事前学習の比率を 50%以上に設定した場合、教師ありファインチューニングのみの場合と比較して KP スコアが顕著に低下することが観察された（下段矢印）。

4 結果と考察

`Qwen2.5-7B-Instruct` と `Qwen2.5-14B-Instruct` の 2 つのモデルについて [14]、追加学習用データセットを 12 通りの異なる比率で分割し、それぞれ継続事前学習および/または教師ありファインチューニ

ングによる追加学習を実施した。それぞれのモデルについて、低頻度、中頻度、高頻度の各頻度帯で算出した KP スコアを**図 2** に示す。

4.1 ベースモデルにおける KP スコア

追加学習を行っていないベースモデルの性能を評価したところ、`Qwen2.5-7B-Instruct` の KP スコアは、低頻度帯で 0.23、中頻度帯で 0.30、高頻度帯で 0.45 を示した。同様に、`Qwen2.5-14B-Instruct` では、それぞれ 0.36、0.41、0.52 となった。これらの結果から、両モデルともに単語の出現頻度帯と KP スコアの間に正の相関が認められ、低頻度語ほど KP スコアが低く、高頻度語になるほど KP スコアが向上する傾向が確認された。これは、高頻度帯の医学用語対は追加学習用データセットだけでなく、PLM の事前学習データにも高頻度で出現していたため、モデルが既に知識として獲得していることを示している。一方、低頻度帯の医学用語対は PLM にとって新規性の高い知識であることが示唆される。したがって、ベースモデルと比較して、追加学習後のモデルにおける低頻度帯の KP スコアの上昇は新規ドメイン知識の獲得度を、高頻度帯の KP スコアの低下は既存知識の忘却度を反映すると考えられた。

4.2 医学用語の出現頻度と知識獲得効率

ベースモデルに対して教師ありファインチューニングのみを実施したところ、`Qwen2.5-7B-Instruct` をベースにしたモデルの KP スコアは、低頻度帯で 0.41、中頻度帯で 0.73、高頻度帯で 0.77 となった（**図 2 赤点線**）。同様に、`Qwen2.5-14B-Instruct` では、それぞれ 0.47、0.74、0.76 を示した（**図 2 青点線**）。全ての頻度帯において、教師ありファインチューニングによる追加学習の実施により、KP スコアはベースモデルと比較して顕著な上昇を示した。特に中頻度帯と高頻度帯では、いずれも KP スコアが 0.7 を超える高い値となった。これらの結果から、追加学習用データセットにおいて中頻度（10-99 回）以上の出現頻度を持つ医学用語対について、KP スコアの評価範囲内でドメイン知識の獲得が確認されたと考えられる。

4.3 継続事前学習の追加による効果

追加学習における継続事前学習の割合に応じて頻度帯ごとの KP スコアを分析したところ、継続事前学習の追加による顕著な KP スコアの向上は、いずれの頻度帯においても認められなかった。特に低頻度帯では、継続事前学習の比率を 50%以上に設定した場合、教師ありファインチューニングのみの場合と比較して KP スコアが顕著に低下することが観察された（**図 2 下段矢印**）。これらの結果から、新規ドメイン知識の獲得度を反映する低頻度帯の KP スコアは、継続事前学習を導入しても向上せず、教師ありファインチューニングのみの場合と比較して有意な改善効果は認められなかった。

4.4 機械翻訳性能の評価

各モデルについて、評価用データセットから機械翻訳の既存メトリクス（BLEU[15] および COMET[16]）を算出した（詳細な結果は **Appendix C** を参照）。KP スコアと同様に、既存の機械翻訳メトリクスにおいても継続事前学習の導入による顕著な性能向上は見られず、教師ありファインチューニングのみの場合と比較して有意な改善効果は認められなかった。

5 結論

本研究では、英日医学翻訳という特定のタスクと Qwen-2.5 ファミリーという特定のモデルに限定されるものの、新規の知識プロービング・ベンチマークを構築し、追加学習における継続事前学習と教師ありファインチューニングの適切な比率を見出すための初期検討を行った。当初は継続事前学習の知識獲得における有効性が期待されたが、予想に反して、教師ありファインチューニングのみでも十分なドメイン知識の獲得が可能であることが示唆された。

本研究の制約として、英日医学翻訳は既にベースモデルにとって既知のタスクであったと推定されるため、本研究において新規の知識として取り扱った情報が、厳密な意味でのドメイン外知識に該当しない可能性がある。また、今後の課題として、追加学習手法の改善とモデルの網羅性の向上、および知識プロービング・ベンチマークの多角的な検証が必要である。特に注目すべき点として、ベースモデルにおける低頻度帯の KP スコア上昇は新規知識の獲得を、高頻度帯の KP スコア低下は既存知識の忘却を反映すると考えられた。このような観点が確立されることで、PLM が追加学習によって新たな知識を獲得する際の内部状態をより詳細に記述することが可能となり、そのメカニズムの解明につながることが期待される。

謝辞

本研究は、JST ACT-X（JPMJAX23C9）、JSPS 科研費（JP22K07681）、国立がん研究センター研究開発費（2023-A-19）、および内閣府・戦略的イノベーション創造プログラム（SIP）「統合型ヘルスケアシステムの構築における生成 AI の活用」の支援を受けて実施した。

参考文献

[1] Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. Investigating continual pretraining in large language models: Insights and implications. **arXiv preprint arXiv:2402.17400**, 2024.

[2] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models Are Zero-Shot Learners. In **The Tenth International Conference on Learning Representations**, 2022.

[3] Kaito Suzuki. 大規模言語モデルの fine-tuning によるドメイン知識獲得の検討. https://tech.preferred.jp/ja/blog/llm-fine-tuning-for-domain-knowledge/. Preferred Networks Blog, 2023.

[4] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 15696–15707, 2023.

[5] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction Pre-Training: Language Models are Supervised Multitask Learners. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2529–2550, 2024.

[6] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. In **Advances in Neural Information Processing Systems**, pp. 55006–55021, 2023.

[7] Jan Hoffbauer, Sylwester Sawicki, Marc Ulrich, Tolga Buz, Konstantin Dobler, Moritz Schneider, and Gerard De Melo. Knowledge Acquisition through Continued Pretraining is Difficult: A Case Study on r/AskHistorians. In **Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)**, pp. 96–108, 2024.

[8] Eric Wu, Kevin Wu, and James Zou. FineTuneBench: How well do commercial fine-tuning APIs infuse knowledge into LLMs? **arXiv: 2411.05059**, 2024.

[9] Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 7765–7784, 2024.

[10] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2463–2473, 2019.

[11] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. What Matters in Memorizing and Recalling Facts? Multifaceted Benchmarks for Knowledge Probing in Language Models. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 13186–13214, 2024.

[12] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 10088–10115, 2023.

[13] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA Learns Less and Forgets Less. **Transactions on Machine Learning Research**, 2024.

[14] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. **arXiv preprint arXiv:2412.15115**, 2024.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[16] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, 2020.

A　追加学習の設定

4 ビット量子化を用いた QLoRA（rank = 32, alpha = 64, dropout = 0.05）を `Qwen2.5-7B-Instruct` および `Qwen2.5-14B-Instruct` の以下の層に適用した：`gate_proj`、`k_proj`、`q_proj`、`v_proj`、`o_proj`、`up_proj`、`down_proj`。継続事前学習および教師ありファインチューニングのいずれにおいても、オプティマイザーとして AdamW を使用し、Warm up ratio = 0.03 を用いて学習率を初期値の 4.0×10^{-4} まで上昇させた後、コサインスケジューラによって学習率を減衰させた。減衰は各学習段階に割り当てられたサンプル数に応じて調整し、学習終了時に学習率が最小となるようにした。学習は継続事前学習と教師ありファインチューニングの各段階において、それぞれに割り当てられたサンプル数を用いて 1 エポックずつ実施した。例えば、CDP:30k/SFT:50k の場合、継続事前学習では 30,000 サンプルの 1 エポック、続く教師ありファインチューニングでは 50,000 サンプルの 1 エポックを、それぞれ学習した。実験には NVIDIA A100 80GB を用いて学習し、PyTorch を利用して実装した。追加学習用データの例を**表 A.1**、プロンプト・テンプレートを**表 A.2** に示す。

表 A.1　追加学習用データの例

English Sentence	Japanese Sentence
Before beginning a sports or vigorous exercise program, children and adults should undergo screening (ie, a history and physical examination), with emphasis on detecting cardiovascular risks. Testing is done only if disorders are clinically suspected.	スポーツまたは激しい運動プログラムを開始する前に，小児および成人には，心血管リスクの発見に重点を置いたスクリーニング（例，病歴聴取および身体診察）を行うべきである。検査は臨床的に疑われる疾患がある場合にのみ実施する。

表 A.2　プロンプト・テンプレート

Prompt Template
<|im_start|>system You are an expert medical translator with extensive experience in translating complex medical texts from English to Japanese. Your expertise covers a wide range of medical fields including but not limited to clinical medicine, pharmacology, medical devices, and healthcare policy.<|im_end|> <|im_start|>user Translate this into Japanese: {{英語医学テキスト}}<|im_end|> <|im_start|>assistant {{日本語医学テキスト}}

B　知識プローブの詳細

本研究では、低頻度、中頻度、高頻度の各頻度帯において、それぞれ 20 組から構成される医学用語対（英語医学用語と日本語医学用語のペアから構成される）と、各医学用語対に対応する 20 組の英日例文対（それぞれの言語の医学テキストは、対応する医学用語を含む）を生成し、合計 400 の英日例文対からなる知識プロービング・ベンチマークを構築した。**表 B.1** に、各頻度帯における 1 つの医学用語対と、それに対応する 1 つの英日例文対を示す。英日例文対はそれぞれ単一文として生成された。

表 B.1　知識プローブの例

Frequency	Medical Term Pair	Medical Sentence Pair
High	severity	The **severity** of the patient's condition was assessed by the medical team.
	重症度	患者の状態の**重症度**は医療チームによって評価されました。
Medium	parotitis	The patient presented with swelling and pain indicative of **parotitis**.
	耳下腺炎	患者は**耳下腺炎**を示唆する腫れと痛みを訴えました。
Low	sesamoiditis	Walking barefoot on hard surfaces can increase the risk of **sesamoiditis**.
	種子骨炎	硬い表面を裸足で歩くと、**種子骨炎**のリスクが高まる可能性があります。

C　機械翻訳としての性能評価

追加学習を実施した各モデルの評価用データセットにおける評価結果を**表 C.1** および**表 C.2** に示す。COMET-22 は `Unbabel/wmt22-cometkiwi-da`、COMET-23 は `Unbabel/wmt23-cometkiwi-da-xl` により算出した。結果、教師ありファインチューニングのみのモデル（CDP:0/SFT:80k）と比較して、継続事前学習を導入したモデルにおいて顕著なメトリクスの改善は見られなかった。

表 C.1　`Qwen2.5-7B-Instruct` における翻訳性能

Study Name	BLEU	COMET-22	COMET-23
CDP:0/SFT:80k	47.17	83.63	69.94
CDP:5k/SFT:75k	50.94	83.87	70.49
CDP:10k/SFT:70k	50.51	83.81	70.48
CDP:15k/SFT:65k	50.49	83.88	70.54
CDP:20k/SFT:60k	48.77	83.64	70.03
CDP:30k/SFT:50k	49.99	83.76	70.18
CDP:40k/SFT:40k	49.44	83.72	70.10
CDP:50k/SFT:30k	49.56	83.81	70.35
CDP:60k/SFT:20k	48.95	83.63	70.14
CDP:65k/SFT:15k	48.66	83.35	69.62
CDP:70k/SFT:10k	48.11	83.56	69.94
CDP:75k/SFT:5k	47.86	83.75	70.10

表 C.2　`Qwen2.5-14B-Instruct` における翻訳性能

Study Name	BLEU	COMET-22	COMET-23
CDP:0/SFT:80k	52.03	83.97	70.81
CDP:5k/SFT:75k	52.18	84.08	70.85
CDP:10k/SFT:70k	51.94	84.06	70.79
CDP:15k/SFT:65k	51.99	83.95	70.67
CDP:20k/SFT:60k	52.00	83.96	70.77
CDP:30k/SFT:50k	51.46	83.78	70.36
CDP:40k/SFT:40k	51.24	84.06	70.76
CDP:50k/SFT:30k	51.05	83.96	70.62
CDP:60k/SFT:20k	50.51	83.99	70.69
CDP:65k/SFT:15k	50.17	83.89	70.55
CDP:70k/SFT:10k	50.74	83.95	70.65
CDP:75k/SFT:5k	49.58	83.90	70.38