PLaMo Translate: 翻訳特化大規模言語モデルの開発

今城 健太郎[1,a)]　平野 正徳[1,b)]　野沢 健人[1,c)]　中鉢 魁三郎[1,d)]

概要：大規模言語モデル (LLM) の発展により、自然言語処理タスクの性能は飛躍的に向上したが、翻訳タスクに特化したモデルの最適化は依然として課題である。本研究では、日本語と英語の翻訳に特化した大規模言語モデル「plamo-2-translate」を提案する。提案モデルは、専用のフォーマットを活用した入出力制御、対訳コーパスと合成データを用いたファインチューニング、Iterative DPO による最適化を組み合わせ、流暢で文脈に即した翻訳を実現する。評価実験では、BLEU、chrF、BERTScore、COMET、GEMBA-MQM などの複数の指標において、ベースモデルや他の LLM と比較して同等以上の性能を達成し、特に人間の評価に近い GEMBA-MQM で顕著な改善を確認した。さらに、文体指定やコンテキスト保持などの機能を備え、多様な翻訳ニーズに対応する。本研究で構築したモデルは、Huggingface を通じて公開しており、そのほかも様々な形態での公開を進めている。

キーワード：LLM, 翻訳, MT, SFT, Iterative DPO

1. はじめに

大規模言語モデル (LLM) の台頭により、様々な自然言語処理のタスクにおいて、画期的な性能が実現している。特に、ChatGPT [6] や GPT-4 [7] をはじめとした最新の言語モデルは、性能向上と汎化が著しく、様々な人間のタスクをも代替するようになってきている。LLM の活用の幅は、テキスト分析や要約、レポート生成、翻訳などのタスク方向と、金融や医療といったドメイン方向の軸で広がりを見せている。

特に、LLM の発展により、自然言語処理タスクにおいて当然のように LLM が活用されるようになった分野として、翻訳が存在する。他のタスクと同様に機械翻訳分野でも、LLM を活用したアプローチが従来の系列変換モデルに代わる手法として大きな注目を集めている。例えば、LLM にプロンプトで「以下の文を英語に翻訳してください」などと指示をすれば、簡単に翻訳を行うことができるわけである。

しかしながら、汎用 LLM は多様なタスクに対応する必要がある一方、単一タスク性能、とりわけ翻訳タスクにおける最適化は十分に進んでいない。従来のニューラル機械翻訳 (NMT) は、Seq-to-Seq ベースのモデル [13] や Attention(Transformer) ベースのモデル [16] といわれるような Encoder-Decoder モデルで構成されるモデルが主流であった。例えば、Transformer ベースの Encoder-Decoder モデルとして有名な T5[11] のモデルは、大きいモデルであっても、11B パラメータ程度[*1]である。一方で、ChatGPT の前身の GPT-3[1] のモデルパラメータサイズは 175B であり、さらに後続のモデルは、そのパラメータ数も増えているとされている。例えば、DeepSeek-R1[2] は、685B パラメータを持つ。一般に、大きいモデルほど高い性能を持つとされているものの、翻訳というタスクだけに特化して考えた場合に、これらのモデルは計算効率が悪い。

LLM と従来の NMT を比較して考えた場合、翻訳タスクにおいては、長所短所がある。従来の NMT は、主に文単位の対訳コーパスで学習されているため、文書レベルの文脈処理や、専門分野に適した文体の選択には限界がある。学術論文や法律文書などの専門分野文書においては、単語レベルの翻訳精度だけでなく、文書全体での文体の統一性や専門用語の正確性が重要な要素となる。一方で、LLM は、様々な常識や多言語間の知識転移があるため、比較的流暢な翻訳が実現可能な一方で、プロンプトによる制御が完全ではないために余計なリード文がでてきたり、大規模な演算器を要するなどの課題がある。

これらを踏まえると、LLM と従来の NMT の長所を掛け合わせるようなモデルの開発が求められる。つまり、LLM のような様々な常識や多言語間の知識のリンケージを持ち

[1] 株式会社 Preferred Networks
[a)] imos@preferred.jp; Corresponding Author
[b)] research@mhirano.jp
[c)] nzw0301@preferred.jp
[d)] kaizaburo@preferred.jp

[*1] https://huggingface.co/google-t5/t5-v1_1-xxl

つつも、プロンプトによる制御が不要でかつ、余分な部分のない対訳を出力し、パラメータ数も現在の超大規模なモデルよりも控えめなモデルに意義があると考えられる。

この要求に合致するモデルとして、本研究では、翻訳特化大規模言語モデルを提案する。より具体的には、PLaMo 2 という、日本語と英語に対応した大規模言語モデルに対して翻訳に特化する学習を行った「plamo-2-translate」を構築した。このモデルの特徴は以下である。

- モデル：LLM をベースのモデルとして採用することで、様々な常識や多言語間の知識のリンケージを理解した翻訳を実現可能である。
- データ：従来の NMT と同様の対訳コーパスを用いることで、翻訳したい文章に対応する文章のみを出力する学習を実現可能である。
- フォーマット：LLM で課題となる、入出力の制御に関して、PLaMo 2 で採用されている PFML2(Preferred Markup Language 2) 形式を採用することで、無駄のない入出力を実現可能である。
- 学習：LLM の事後学習に使用される技術である、Supervised Fine-Tuning(SFT) や Direct Preference Optimization(DPO) などを用いることで、モデルを翻訳に特化させつつ、好ましい翻訳について学習をさせる。

これらの技術の組み合わせの結果、従来の手法では困難だった専門分野文書の翻訳をはじめとして、高精度な翻訳を実現することに成功した。

本研究で構築されたモデルは、一般に公開しており、Hugging Face (`https://huggingface.co/pfnet/plamo-2-translate`) からダウンロード可能である。また、デモサイト `https://translate-demo.plamo.preferredai.jp/` を通じて利用できる[*2]。

2. 提案手法: 翻訳特化モデルの構築

本章では、PLaMo Translate の構築に使用した手法について述べる。

2.1 ベースモデル

翻訳特化モデルの構築におけるベースモデルの選定に当たっては、モデルの言語性能および、翻訳文章の扱いが楽になるようなフォーマットに対応していることを重視した。

具体的には、PLaMo Translate の構築に当たっては、pfnet/plamo-2.1-8b-cpt[*3]をベースモデルとして用いた。このモデルは、State Space Model（SSM）アーキテクチャである Mamba を採用した LLM であり、元々は 31B パラメータのモデルを 2 兆トークンで事前学習した後、プルーニングによりモデルパラメータ数を削減したモデルとなっている。

言語性能面では、日本語の言語能力が高いとされるシリーズであり[*4]、8B のモデルでも日英・英日翻訳において高い性能が期待できると考えた。

また、このモデルは、Preferred Markup Language 2 (PFML2) と呼ばれる独自フォーマットをサポートしている。PFML2 フォーマットは、`<|plamo:op|>`トークンを用いた特別なマークアップ言語である。これは、OpenAI 社が採用している、`<|im_start|>`などと類似のものである。基本的な構造は `input` ブロックと `output` ブロックで構成され、言語指定には `lang` 属性を使用できる。この PFML2 により、LLM の入出力を構造化できるため、従来の単純なプロンプトベース翻訳と比較して、より細かな翻訳制御が可能となると考えられる。

以下に PFML2 フォーマットの使用例を図 1 に示す。

今回は、日本語を軸に翻訳モデルを実装したため、pfnet/plamo-2.1-8b-cpt を採用したものの、手法という観点では、PFML2 マークアップに相当する機能を持つモデルであれば、利用可能である。

```
<|plamo:op|>dataset
translation
<|plamo:op|>input lang=English
Preferred Networks (PFN) rapidly realizes practical
↪  applications of deep learning and other emerging
↪  technologies.
<|plamo:op|>output lang=Japanese
Preferred Networks（PFN）は、既存技術では解決が困難な現実
↪  世界の課題解決に向けて、深層学習をはじめとする先端技術
↪  の実用的な応用を迅速に実現している。
<|plamo:op|>
```

図 1 PFML2 フォーマットの使用例

2.2 学習データ

ベースモデルを翻訳用に学習させるにあたっては、ファインチューニング用の対訳データセットが必要となる。対訳データセットの多くはすでにベースモデルの学習に使用されていることが多く、これらを再度学習させるだけでは翻訳の性能を向上させられないと考えた。そこで、本研究では、指示追従性の高い DeepSeek モデルを用いたデータの合成を実施した。

まず、対訳データを作成するためのシードとなる被翻訳データは、青空文庫などのオープンなデータを収集した。

その上で、DeepSeek-V3-0324[*5]を用いて、翻訳データを構築した。ここで、田中コーパスのように、既存の対訳がある場合は、DeepSeek にヒントを与えながら対訳が適切にアラインされるように処理した。

[*2] 2025/7 時点。デモサイトは閉鎖される可能性もあるため、その場合は Hugging Face のモデルを利用されたい。
[*3] `https://huggingface.co/pfnet/plamo-2.1-8b-cpt`
[*4] `https://www.preferred.jp/ja/news/pr20250522/`
[*5] `https://huggingface.co/deepseek-ai/DeepSeek-V3-0324`

さらに、翻訳データを同様の手順で再翻訳することで、追加のデータを作成した。

これらの手順により作成された

- 元データ (被翻訳データ) → 翻訳データ
- 翻訳データ → 再翻訳データ

の 2 種類を学習データとして採用した。なお、ここで、これらのデータの逆向き (例えば、翻訳データ → 元データ) はデータとして採用しないことに注意されたい。

2.3 ファインチューニング手法

ベースモデルを翻訳向けにファインチューニングするにあたっては、LLM での事後学習のプロセスを参考にして、SFT と DPO[10] を採用した。以下にその詳細について述べる。

2.3.1 Supervised Fine-Tuning (SFT)

最初に、学習データを用いて SFT を実施した。一般に、LLM を翻訳する場合においては、「以下の文章を英語に翻訳してください」などの指示を行うプロンプトが必要であるが、本研究で構築するモデルはこのような指示を不要にして翻訳をしたい。一方で、プロンプトを使用しないことにより、フォーマットの指示や、詳しい文脈などの情報が欠落してしまっても適切な翻訳ができなくなってしまうため、これらの情報を含めることができるような SFT を行うこととした。

そこで、SFT を実施するにあたって、PFML2 フォーマットを拡張して、学習を行った。具体的には、文体指定には texttttstyle 属性、日本語の敬体・常体指定には `form` 属性を追加した。また、`context` ブロックで文脈情報、`vocabulary` ブロックで特定用語の翻訳を制御できるようにした。

これらの学習を SFT にて行うことにより、これまでベースモデルでは対応できなかった、翻訳に特化したフォーマットを学習させた。

2.3.2 Iterative DPO

SFT の後、翻訳報酬モデルを用いた Iterative DPO を実施した。

報酬モデルの構築に当たっては、出力の好ましさの優劣関係を MetricX[4] および DeepSeek を用いて作成した。まず、DeepSeek-V3-0324 を用いた LLM-as-a-judge を優劣関係を判定する 1 つの手法として採用した。さらに、MetricX[4] をもう一つの優劣関係を判定する手法として採用した。その上で、1 つの入力に対して、学習の対象とするモデルを用いて生成した 100 個の回答に対して、これらの 2 つの手法の順位を足し合わせたアンサンブルの順位を用いて、その中で良い方のサンプルと悪い方のサンプルをそれぞれ正例と負例として採用し、報酬モデルを学習させた。

その上で、3 段階からなる Iterative DPO[3], [17] を実施した。第 1 段階では基本的な選好データを用いた DPO 学習、第 2 段階では第 1 段階モデルから生成された新しい候補を含む選好データでの学習、第 3 段階では累積的な選好データを用いた最終的な品質向上を行う。

各段階で生成されるデータの品質を報酬モデルで評価し、より高品質な選好データを用いて次の段階の学習を行った。この手法により、単一段階の DPO では達成困難な性能向上を実現している。

なお、DPO の各段階では、機械的に翻訳が失敗したケース (行数の相違など) を除外しながら、報酬モデルの情報に基づいて選好データを作成した。

2.3.3 モデルマージ

3 段階の Iterative DPO の後、モデルマージを行うことで最終的な性能向上を達成した。各段階で得られたモデルの長所を組み合わせることで、単一段階の学習では得られない高性能を実現している。

モデルマージには重み付き平均手法を採用し、各 DPO 段階で得られたモデルの重みを組み合わせた。この重み付けは、各段階モデルの翻訳品質評価結果に基づいて決定され、後の段階ほど高い性能を示したため、より大きな重みを配分している。

3. 評価実験

評価に当たっては、現時点で利用可能な一般的な翻訳ベンチマークを採用して、評価を行うこととした。ここでは、まず、ベースモデルと比べて、チューニングを行ったモデルがどのように精度が上がっているのかを確認することを目的としつつ、GPT-4o のようなグローバルでより大規模な LLM による翻訳との性能差についても追加で確認を行う。

翻訳ベンチマークの計測に当たっては、先行の取り組み [19][*6]を参考にし、WMT 2024 General Translation Task (English-to-Japanese) のテストセットの翻訳結果に関して、BLEU[8]、chrF[9]、BERTScore[18]、COMET[12]、GEMBA-MQM[5] を計測した。

なお、BERTScore の計算においては、日本語の計測には izumi-lab/deberta-v2-base-japanese[14], [15][*7]を、英語の計測には microsoft/deberta-xlarge-mnli[*8]を使用した。COMET の計測に当たっては、標準の Unbabel/wmt22-comet-da[*9]を使用した。また、GEMBA-MQM の計測に当たっては、先行の取り組み [19] と同様に GPT-4 ではなく、GPT-4o を使用した。

なお、GEMBA-MQM のみは、評価値が思考によってぶ

[*6] https://sites.google.com/view/nlp2025ws-langeval/task/translation

[*7] https://huggingface.co/izumi-lab/deberta-v2-base-japanese. 先行研究とはことなるが、juman++の依存関係の解決が環境によっては難しいために、使用するモデルを差し替えた。

[*8] https://huggingface.co/microsoft/deberta-xlarge-mnli

[*9] https://huggingface.co/Unbabel/wmt22-comet-da

れるため、10 回計測した平均値を採用する。

今回、評価の対象にしたモデルは、OpenAI の gpt-4o-2024-11-20 と gpt-4o-mini-2024-07-18 および、plamo-2-translate とそのベースモデルである plamo-2.1-8b-cpt を採用した。

加えて、評価対象の LLM で翻訳を生成する際には、GPT-4o シリーズには、「You are a professional translator. Translate the following text from en to ja. Only return the translated text without any additional explanation.」とシステムプロンプトを与えており、plamo-2-translate には一切のプロンプトを与えていない。さらに、plamo-2-8b に関しては、ベースモデルであるために指示追従性が低いことを考慮し、田中コーパスから抽出した 4 つのサンプルを few-shots example として与えつつ、かつ、`<pre></pre>`タグでそれぞれの example を囲んで与えることで、出力の安定性を担保した。

4. 結果と考察

表 1 Scores: BLEU (↑)

	ja→en	en→ja
gpt-4o-2024-11-20	23.647	27.158
gpt-4o-mini-2024-07-18	22.864	25.223
plamo-2.1-8b-cpt	19.400	22.423
plamo-2-translate	21.179	21.312

表 2 Scores: chrF (↑)

	ja→en	en→ja
gpt-4o-2024-11-20	51.134	36.649
gpt-4o-mini-2024-07-18	50.065	34.587
plamo-2.1-8b-cpt	40.486	31.198
plamo-2-translate	49.061	31.277

表 3 Scores: BERTScore (↑)

	ja→en	en→ja
gpt-4o-2024-11-20	0.779	0.770
gpt-4o-mini-2024-07-18	0.770	0.760
plamo-2.1-8b-cpt	0.686	0.735
plamo-2-translate	0.765	0.747

表 4 Scores: COMET (↑)

	ja→en	en→ja
gpt-4o-2024-11-20	0.832	0.878
gpt-4o-mini-2024-07-18	0.827	0.865
plamo-2.1-8b-cpt	0.789	0.829
plamo-2-translate	0.832	0.876

表 5 Scores: GEMBA-MQM (↑)

	ja→en	en→ja
gpt-4o-2024-11-20	81.237 ± 0.159	77.974 ± 0.179
gpt-4o-mini-2024-07-18	78.864 ± 0.169	73.032 ± 0.270
plamo-2.1-8b-cpt	57.102 ± 0.215	63.350 ± 0.246
plamo-2-translate	77.126 ± 0.307	73.770 ± 0.213

```
[原文] "People Swimming in the Swimming Pool" from 2022 is one Vicente Siso artwork that will display at Tierra del Sol Gallery beginning Jan. 13. (photo courtesy of Vicente Siso)
------------------------------------------------------
[文体指定なし] 2022 年制作の『プールで泳ぐ人々』は、1 月 13 日からティエラ・デル・ソル・ギャラリーで展示されるビセンテ・シソの作品の一つです。(写真提供：ビセンテ・シソ)
------------------------------------------------------
[だ・である調] 2022 年制作の『プールで泳ぐ人々』は、1 月 13 日からティエラ・デル・ソル・ギャラリーで展示されるビセンテ・シソの作品の一つである。(写真提供：ビセンテ・シソ)
------------------------------------------------------
[です・ます調] 2022 年制作の『プールで泳ぐ人々』は、1 月 13 日からティエラ・デル・ソル・ギャラリーで展示されるビセンテ・シソの作品の一つです。(写真提供：ビセンテ・シソ)
```

図 2 「だ・である調」、「です・ます調」の指定による翻訳結果の違い

表 1–5 にベンチマークスコアを示した。

この結果に従うと、翻訳に特化したモデルの効果は、ベースモデルである plamo-2.1-8b-cpt と比べてほぼすべてのスコアにおいて改善が確認できる。特に、BLEU や chrF という表層的な評価よりも LLM-as-a-judge で人間に近い評価をできるとされる GEMBA-MQM においてのスコアの改善は顕著であり、流暢な翻訳が実現できている可能性が示唆される。

GEMBA-MQM における GPT-4o シリーズのスコアは、GPT-4o 自身で評価していることもあり、参考値程度でしかない点には注意されたいが、plamo-2-translate は、翻訳に特化した結果、8B という小さいモデルであっても、gpt-4o や gpt-4o-mini に並ぶ水準の翻訳が実現できている可能性が示唆される結果が得られた。

plamo-2.1-8b-cpt の翻訳結果を確認すると、翻訳先の言語を明示的に指定しているにもかかわらず、元の言語で応答しているケースなども存在している一方で、plamo-2-translate では、言語の間違いは発生しておらず、翻訳における指示追従性の向上も確認できた。

今回のベンチマークでは、plamo-2-translate ならではの style 指示機能などの指示追従性については確認することができていないものの、「だ・である調」、「です・ます調」を日本語の文体として指定できるようになっており、実際に、適切な翻訳がされていることが確認できている (図 2)。

本研究では、定量的な検証が困難であるため、詳細には

踏み込まないものの、コンテキストを与える機能や単語帳機能、テキストフォーマット指定機能 (マークダウンや html、rst などの指定)、パラグラフ構造を維持する機能なども plamo-2-translate には含まれている。これらの定量評価は今後の課題である。

本研究では、主にモデルの学習と評価に着目したが、エンジニアリング面として vLLM と MLX の 2 つの主要な推論エンジンでの実行をサポートを通して、ローカル環境での推論を可能にしている[*10]。4 ビット量子化モデルは 6GB 以上の空きメモリがあれば十分に高い性能で翻訳が行える。

今後の課題としては、日英以外の言語ペアに対する Iterative DPO 最適化の適用、量子化技術とモデル圧縮による推論速度向上、医療や金融等の専門分野へのより良い対応、より包括的な翻訳品質評価指標の開発などがあげられる。

5. まとめ

本研究では、LLM をベースにした翻訳モデルの構築として、plamo-2-translate という、翻訳特化大規模言語モデルの構築手法について提案した。モデルの構築に当たっては、ベースモデルとなる LLM に対して、翻訳に特化したコーパスを構築し、SFT および Iterative DPO を適用することで、好ましい翻訳を学習させた。その結果、各種翻訳ベンチマークにおいて、ベースモデルを上回るスコアを達成し、グローバルモデルである GPT-4o や GPT-4o-mini に匹敵する翻訳性能を 8B のモデルで実現した。本研究で構築したモデルは、huggingface で公開しているほか、様々な形態での公開を進めている。

謝辞 本研究は、経済産業省及び国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）が実施する、国内の生成 AI の開発力を強化するためのプロジェクト「GENIAC（Generative AI Accelerator Challenge）」の支援を受けて実施されました。また、本研究の社会実装として PLaMo 翻訳の一般公開版・ブラウザ拡張等のリリースにおいて、奥井寛樹氏にご尽力いただいたきました。

参考文献

[1] Brown, T. B.: Language models are few-shot learners, *arXiv preprint arXiv:2005.14165* (2020).

[2] DeepSeek-AI: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (2025).

[3] Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., Jiang, N., Sahoo, D., Xiong, C. and Zhang, T.: RLHF Workflow: From Reward Modeling to Online RLHF, *Transactions on Machine Learning Research*, (online), available from ⟨https://openreview.net/forum?id=a13aYUU9eU⟩ (2024).

[4] Juraska, J., Deutsch, D., Finkelstein, M. and Freitag, M.: MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task, *Proceedings of the Ninth Conference on Machine Translation* (Haddow, B., Kocmi, T., Koehn, P. and Monz, C., eds.), Miami, Florida, USA, Association for Computational Linguistics, pp. 492–504 (online), available from ⟨https://aclanthology.org/2024.wmt-1.35⟩ (2024).

[5] Kocmi, T. and Federmann, C.: GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4, *Proceedings of the Eighth Conference on Machine Translation*, Singapore, Association for Computational Linguistics, pp. 768–775 (online), DOI: 10.18653/v1/2023.wmt-1.64 (2023).

[6] OpenAI: ChatGPT (2023).

[7] OpenAI: GPT-4 Technical Report (2023).

[8] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318 (2002).

[9] Popović, M.: chrF: character n-gram F-score for automatic MT evaluation, *Proceedings of the tenth workshop on statistical machine translation*, pp. 392–395 (2015).

[10] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S. and Finn, C.: Direct preference optimization: Your language model is secretly a reward model, *Advances in neural information processing systems*, Vol. 36, pp. 53728–53741 (2023).

[11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67 (online), available from ⟨http://jmlr.org/papers/v21/20-074.html⟩ (2020).

[12] Rei, R., Stewart, C., Farinha, A. C. and Lavie, A.: COMET: A Neural Framework for MT Evaluation, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 2685–2702 (online), DOI: 10.18653/v1/2020.emnlp-main.213 (2020).

[13] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to sequence learning with neural networks, *Advances in neural information processing systems*, Vol. 27 (2014).

[14] Suzuki, M., Sakaji, H., Hirano, M. and Izumi, K.: Constructing and analyzing domain-specific language model for financial text mining, *Information Processing & Management*, Vol. 60, No. 2, p. 103194 (online), DOI: 10.1016/j.ipm.2022.103194 (2023).

[15] Suzuki, M., Sakaji, H., Hirano, M. and Izumi, K.: FinDeBERTaV2: Word-Segmentation-Free Pre-trained Language Model for Finance, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 39, No. 4, pp. FIN23–G_1–14 (online), DOI: 10.1527/tjsai.39-4_FIN23-G (2024).

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention Is All You Need, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5999–6009 (2017).

[17] Xu, J., Lee, A., Sukhbaatar, S. and Weston, J.: Some things are more cringe than others: Iterative preference optimization with the pairwise cringe loss, *arXiv preprint arXiv:2312.16682* (2023).

[18] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, *International Confer-*

[*10] `https://huggingface.co/mlx-community/plamo-2-translate`

ence on Learning Representations, (online), available from ⟨https://openreview.net/forum?id=SkeHuCVFDr⟩ (2020).

[19] 須藤克仁, 小町　守, 梶原智之, 三田雅人: NLP2025 ワークショップ：LLM 時代のことばの評価の現在と未来, 自然言語処理, Vol. 32, No. 2, pp. 738–745（オンライン）, DOI: 10.5715/jnlp.32.738 (2025).