対訳データを用いた大規模言語モデルの継続事前訓練による特許請求項翻訳

浅見遥斗[1] 近藤海夏斗[1] 宇津呂武仁[1] 永田昌明[2]
[1]筑波大学大学院 システム情報工学研究群 [2]NTT コミュニケーション科学基礎研究所
{s2420710,s2320743}_@_u.tsukuba.ac.jp utsuro_@_iit.tsukuba.ac.jp
masaaki.nagata_@_ntt.com

概要

大規模言語モデル (LLM) の技術進歩は著しく，様々な分野でその活用が進んでいる．しかし、特許翻訳の分野では依然として Transformer ベースの翻訳が主流であり，LLM を活用した翻訳能力については十分に検討されていない．そこで本研究では、近藤ら [6] の手法を参考に，対訳データを用いた継続事前訓練と SFT(Supervised Fine-Tuning) を行った LLM を用いて特許請求項の翻訳を行い，Transformer ベースの翻訳と比較した．その結果，BLEU および COMET のスコアがいずれも上回り，訳抜けや繰り返しといった課題が改善されたことを確認した．一方で、従来モデルでは発生しなかったハルシネーションが一部の例で確認され，その影響で翻訳精度が低下したケースも観測された．本研究は，LLM が特許翻訳分野において有望な技術である一方で，課題も存在することを示している．

1 はじめに

大規模言語モデル (LLM) は，その広範な事前訓練により，多様な自然言語処理タスクで有用性が示され，機械翻訳分野においても活躍している．しかし，その中でも特許分野における翻訳能力については十分に検討されていない．特許文書は，専門用語の多用や複雑な文構造を特徴としており，従来の Transformer ベースのモデルでは，訳抜けや繰り返しなどの問題が残る場合がある．本研究では，近藤らの提案する対訳データを用いた継続事前訓練と SFT の手法を参考に，特許請求項翻訳に特化した LLM を構築し，その性能を従来モデルと比較することを目的とする．具体的には，特許文の中でも文の長さや独特の記述形式から翻訳難易度の高い特許請求項の翻訳を行い，BLEU や COMET などの評価指標を用いて翻訳精度を測定することで，それらで従来モデルより統計的に有意に向上し，訳抜けや繰り返しを改善できていることを確認した．しかし、一部の例では、従来モデルでは発生しなかったハルシネーションが観測されるなど，LLM 特有の課題も明らかになった．本研究は，LLM を特許翻訳に適用する際の可能性と課題を検討し，その有効性を評価するものである．

2 関連研究

Xu ら [3] は，Llama-2 [14] に対して 2 段階の fine-tuning を行う ALMA という手法を提案している．ALMA は，まず単言語データを用いた fine-tuning を行った後，少量の高品質な対訳データを用いた fine-tuning を行うことで，13B の LLM を使って GPT-3.5 に匹敵する翻訳精度を達成した．近藤ら [6] は，大規模言語モデル（LLM）の翻訳性能向上を目的として，「継続事前訓練」と「SFT（Supervised Fine-Tuning）」を組み合わせた二段階の訓練手法を提案している．この手法では、Web 上で収集した大規模な並列データである JParaCrawl v3.0 [7] を用いて継続事前訓練を行い，その後 WMT20 [1]，および Flores200 [13] の開発，テストデータ，そして KFTT [9] の訓練データからなる高品質な対訳データを使用して SFT を実施している．このアプローチは，大量の対訳データを活用してモデルの翻訳性能を向上させるだけでなく，高品質なデータを用いたファインチューニングによって，翻訳精度をさらに高めることを目的としている。これらの手法を適用したモデルを，WMT22 [5] のテストデータをはじめとする 12 種のテストデータで評価を行い，対訳データで訓練された Transformer ベースのモデルより統計的に有意に向上することを確認した．

本研究ではこの手法を参考にし，特許分野におけ

る対訳データを活用した継続事前訓練と SFT を組み合わせたモデル訓練を行った．

3　特許データを用いた LLM の訓練

先行研究 [6] にならい，単言語データを用いた継続事前訓練を行った LLM である rinna/llama-3-youko-8b [1](以下 rinna-8b と呼ぶ) に対し，特許対訳データを用いた継続事前訓練および SFT を行った．ここで youko-8b は日本語および英語の単言語データ 220 億トークンを用いて事前訓練されている．

4　実験設定

4.1　データセット

継続事前訓練および SFT には、日英特許対訳コーパスである JParaPat[8] を使用した。具体的には、2016 年から 2021 年までの特許対訳データを利用し、その内訳は表 1 に示す。継続事前訓練では、2016 年から 2020 年までの特許対訳データ全文から，50,000 件抜き取った残りの文を訓練データとした．抽出した 50,000 件は LaBSE [2] [2] で取得した文埋め込みベクトルを用いて計算された類似度でフィルタリングを行い，類似度 0.9 以上 0.95 以下となった 10,984 件を開発データとして用いた．SFT では、2021 年の特許対訳データのうち特許請求項を対象に、LaBSE による類似度（0.8 以上 0.95 以下）を条件としてフィルタリングを行い、訓練データと開発データをそれぞれ選定した。また、2021 年の特許請求項から、重なりのない類似度 0.9 以上 0.95 以下かつ単語数 100 以上の文をテストセットとして使用した。

なお，継続事前訓練では対訳文を

日本語文\n 英語文 という形式で与え，SFT ではプロンプトを伴う以下の形式で

> これを日本語から英語に翻訳してください.
> 日本語: 原言語文
> 英語: 目的言語文

という形式を適用した．なお，SFT は full fine-tuning と LoRA [4] チューニングの両方を行った．

4.2　ハイパーパラメータ

本論文における継続事前訓練および SFT のハイパーパラメータの設定は近藤ら [6] に従った．

1） https://huggingface.co/rinna/llama-3-youko-8b
2） https://huggingface.co/sentence-transformers/LaBSE

表 1　特許対訳データの使用用途とデータ内訳

使用用途	対象期間	データ種別	データ数
継続事前訓練	2016 年 ~ 2020 年	訓練データ	61,364,685
		開発データ	10,984
SFT	2021 年	訓練データ	15,000
		開発データ	1,000
テストセット	2021 年	—	33,923

4.3　比較対象

4.3.1　ベースライン

ベースラインとしては，Transformer [15] を 2016~2020 年までの特許対訳全文で訓練したモデルを用いた．

4.3.2　LLM

比較対象として，近藤らが用いた JParacrawl v3.0 で継続事前訓練を rinna-8b に行ったモデルに対し，WMT20 のテストセット等で SFT をしたモデル，特許請求項で SFT をしたモデルで翻訳を行い評価を行った．

4.3.3　プロンプト

推論に用いるプロンプトとして，4.1 節に示した SFT 訓練データ用の形式を用いたところ，文頭に原言語文に存在しない数字が発生した．具体例は付録 A に示す．これらの明確な理由は不明だが，継続事前訓練，SFT にどのデータを用いた場合でも日英翻訳においてはこの事例が発生するため，rinna-8b の日本語継続事前訓練に要因があると考えられる．

そこで，プロンプトでこれら原言語文に存在しない数字が発生する例を抑えることが出来るかの検討のため，プロンプトを以下の文に変更し推論を行った．

> これを日本語から英語に翻訳してください。
> ただし文頭に関係のない数字を出さないようにしてください。:
> 日本語: `{text}`
> 英語:

4.4　継続事前訓練における必要データ数の調査

今回特許データで継続事前訓練を行う際，6100 万文対のデータを使用したが，実際ここまでデータ数

が必要なのかは不明である．そこで，0.1 エポックごと (610 万文) のチェックポイントに対して SFT を行い性能を比較し，必要データ量を調査した．なお，比較対象として，継続事前訓練のしていない rinna-8b に SFT を適用したモデルでの翻訳精度を 0.0 エポックとして記載した．

4.5 評価指標

評価指標として，BLEU [10] および COMET [12] を使用した．BLEU は sacreBLEU [11] を用いて計測した．COMET のモデルは wmt22-comet-da を使用した．また，ベースラインの翻訳結果と，LLM の中で最もシステムレベルのスコアの高い翻訳結果を，文レベルの BLEU と COMET でも評価を行い，win/lose 例を抽出し傾向の分析を行った．

5 評価結果

表 2 訓練方法ごとの BLEU スコアと COMET スコア．* はベースラインと有意差あり ($p < 0.05$).

訓練方法	BLEU	COMET
ベースラインモデル	50.2	81.92
継続事前訓練 + SFT(手法)		
JParaCrawl + WMT(Full)	38.0	81.42
JParaCrawl + WMT(LoRA)	34.2	80.70
JParaCrawl + 特許請求項 (Full)	43.5	81.36
JParaCrawl + 特許請求項 (LoRA)	43.8	81.37
特許 + 特許請求項 (Full)	50.7*	81.25
特許 ＋ 特許請求項 (LoRA)	51.3*	80.79
特許 + 特許請求項 (Full) +プロンプト改善	52.0*	**82.55***
特許 ＋ 特許請求項 (LoRA) +プロンプト改善	**52.3***	82.52*

5.1 継続事前訓練および SFT をした際の結果

表 2 に，継続事前訓練および SFT をした際の翻訳精度について示す．結果より，JParaCrawl で継続事前訓練を行ったモデルに比べ，特許データで継続事前訓練を行ったモデルが BLEU で大幅に向上していることから，特許データで継続事前訓練を行うことで特許の知識を学習することが出来ていることが確認できた．ただし，COMET においては JParaCrawl での継続事前訓練を行った場合の方が勝っているため，COMET の向上には汎用的な知識が求められると考えられる．

また，特許請求項を用いた SFT を行うことで，BLEU が向上し，ベースラインに対して統計的に有意 ($p<0.05$) に向上した．SFT の手法としては，Full fine-tuning では 50.7，LoRA では 51.3 となり，LoRA が統計的に有意に上回る結果となった．ただし，COMET においては，Full fine-tuning の方が上回る結果となった．

また，推論時のプロンプトを 4.3.3 節に示したように改善した結果，SFT の手法に関係なく，BLEU，COMET ともに向上する結果となった．なお，プロンプト改善前と同様に，BLEU においては LoRA が統計的に有意に上回る結果となったが，COMET においては統計的に有意差があるとは言えない結果となった．スコア向上の要因分析として，個々の例を確認した結果，プロンプトに追加した通り文頭の無関係な数字の削除に成功していたため，プロンプト改善は翻訳において有用であったと考えられる．プロンプト修正前と修正後の出力を付録 A に示す．

5.2 継続事前訓練の必要データ数

図 1 に 0.1 エポックごとに SFT を行った際の翻訳評価結果を示す．0 エポック，すなわちベースモデルである rinna-8b に対して特許請求項データをそのまま SFT した場合，BLEU が 40.8 であった．しかし，0.1 エポック時点で BLEU スコアは 49.8 まで向上しており，少量のデータであっても継続事前訓練を行うことで翻訳精度が大幅に向上することが確認された．BLEU は 0.5 エポックまでは統計的に有意に向上を続けたことから，0.5 エポック (3000 万文) までは訓練に必要であったと考えられる．

一方で，COMET スコアを基準とした評価では，0.7 エポックまでは統計的に有意な向上が見られた．このことから，0.7 エポック（約 4200 万文）までは訓練が必要であったと判断できる．以上の結果を踏まえると，定量的な評価の観点から，翻訳精度向上に必要な訓練量としては、0.7 エポック (約 4200 万文) が適切であると結論づけられる．

5.3 傾向分析の結果

傾向分析のためまず文レベルの COMET，BLEU を計算し，win/lose の文集合の分割を行った．なお，僅差の例では違いの分析が難しいため，COMET 差 0.1 以上 0.2 未満の win/lose，COMET 差 0.2 以上の win/lose の例を 50 例ずつ抽出し分析を行った．これらの差が出た例の総数および，抽出した 50 例の参

表 3　COMET 差 win/lose ごとの文長比平均

COMET 差	件数	ベースライン	LLM
0.1 以上 0.2 未満 LLM 勝ち	613	0.5760	0.9556
0.1 以上 0.2 未満 LLM 負け	355	0.9434	0.9860
0.2 以上 LLM 勝ち	203	0.4209	1.0022
0.2 以上 LLM 負け	380	0.7462	0.3984

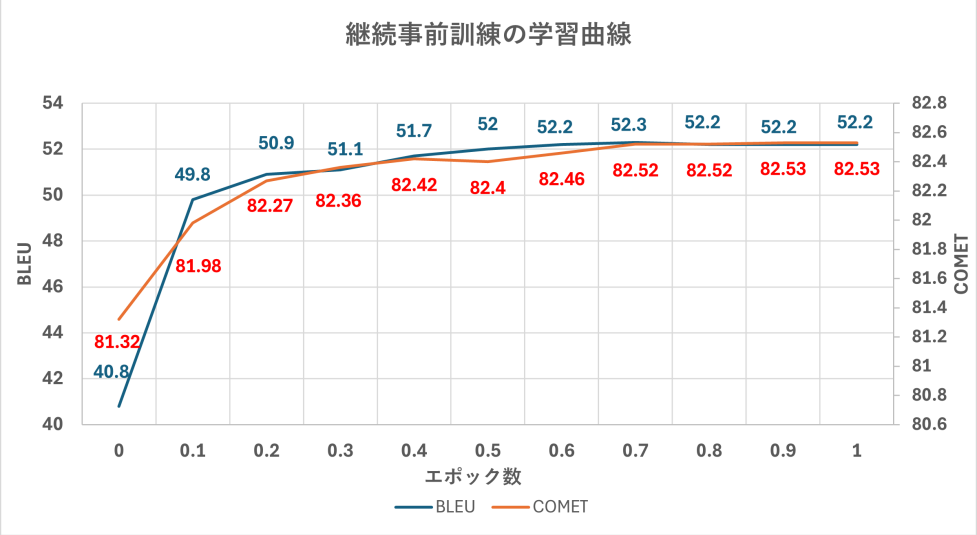

図 1　継続事前訓練の学習曲線

照文と翻訳結果 (ベースライン，LLM) との文長比を表 3 に記載する．ここで参照文と翻訳文の文長比は，抽出した例を観察した結果，lose 側に訳抜けが発生した例が多かったため，その指標として計算した．

結果としては，観察で感じた通り，基本的に lose 側の参照文に対する文長比が短く，訳抜けが多いと考えられる結果になった．0.1 以上 0.2 未満の LLM 負けでは例外のように見えるが，個々の確認の結果外れ値として，非常に文長比の大きい例があるため平均が大きくなっているだけであり，基本的には文長比は小さい結果であった．さらなる観察の結果，ベースラインでスコアの低い例は訳抜けのみではなく繰り返しが発生していることも確認した．それに対し LLM では，ベースラインに比べ繰り返しが少ないが，元々発生していなかったハルシネーションが発生した例が確認できた．具体例とより詳細な分析については，付録 B に記載する．

6　おわりに

本論文では，特許データを使用した継続事前訓練および SFT を行うことで，BLEU，COMET の両方で transformer ベースの翻訳モデルで統計的に有意に向上できることが確認できた．分析の結果，精度向上の主要因としては訳抜け，繰り返しの改善があると考えられるが，逆に新たに訳抜けが発生している例や，ベースラインでは発生していなかったハルシネーションが発生した例が確認できたため，今後の課題が残る結果となった．

参考文献

[1] L. Barrault, et al. Findings of the 2020 conference on machine translation (WMT20). In **Proc. 5th WMT**, pp. 1–55, 2020.

[2] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In **Proc. 60th ACL**, pp. 878–891, 2022.

[3] X. Haoran, K. Young Jin, S. Amr, and A. Hany Hassan. A paradigm shift in machine translation: Boosting translation performance of large language models. In **Proc. 12th ICLR**, 2024.

[4] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In **Proc. 10th ICLR**, 2022.

[5] T. Kocmi, et al. Findings of the 2022 conference on machine translation (WMT22). In **Proc. 7th WMT**, pp. 1–45, 2022.

[6] M. Kondo, T. Utsuro, and M. Nagata. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In **Proc. 21th IWSLT**, pp. 203–220, 2024.

[7] M. Morishita, K. Chousa, J. Suzuki, and M. Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proc. 13th LREC**, pp. 6704–6710, 2022.

[8] 永田昌明, 森下睦, 帖佐克己, 安田宜仁. JaParaPat: 大規模日英特許対訳コーパス. 言語処理学会第 30 回年次大会論文集, pp. 2367–2372, 2024.

[9] G. Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: A method for automatic evaluation of machine translation. In **Proc. 40th ACL**, pp. 311–318, 2002.

[11] M. Post. A call for clarity in reporting BLEU scores. In **Proc. 3rd WMT**, pp. 186–191, 2018.

[12] R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **Proc. 7th WMT**, pp. 578–585, 2022.

[13] NLLB Team, et al. No language left behind: Scaling human-centered machine translation. **arXiv**, Vol. 2207.04672, , 2022.

[14] H. Touvron, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv**, Vol. 2307.09288, , 2023.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In **Proc. 31st NIPS**, pp. 1–11, 2017.

A　プロンプトの出力例

4.1 節のプロンプトでの出力例および，4.3.3 節の改善後プロンプトでの出力例を表 4 に示す．

B　分析

分析の具体例を表 5 に示す．LLM 勝ちの例では，ベースラインでの翻訳が大幅な訳抜けにより，原言語文の最後の式しか残っていない．LLM の方も完璧な翻訳ではないが，ベースラインよりは確実に原言語文の内容を反映している．表 5 のような大幅な訳抜けが多くを占めるわけではないが，基本的に LLM 勝ちの場合ではベースラインに訳抜けが起きていて，LLM ではその訳抜けが軽減されている，といった例が多い．

LLM 負けの例でも同様に訳抜けがベースラインよりも大きく訳抜けしている場合も多いが，表 5 に示すような，同じ種類の単語，表においては癌の種類列挙に対して一つにまとめて表現する，といった例が存在する．これは，LLM の汎用的な知識から発生したものだと考えられるため，LLM 特有の課題となる．

表 4　4.1 節のプロンプトによる翻訳文

原言語文

細胞の単位用量が、規定の数の CD8 + /CCR7 + 細胞、CD4 + /CCR7 + 細胞、CD8 + /CD27 + 細胞、CD4 + /CD27 + 細胞、CD8 + /CCR7 + /CD27 + 細胞、および/または CD4 + /CCR7 + /CD27 + 細胞を含む、請求項 113 記載の方法。

参照訳文

The method of claim 113, wherein the unit dose of cells comprises a defined number of CD8+/CCR7+ cells, CD4+/CCR7+ cells, CD8+/CD27+ cells, CD4+/CD27+ cells, CD8+/CCR7+/CD27+ cells and/or CD4+/CCR7+/CD27+ cells.

LLM による翻訳文

114. The method of claim 113, wherein the unit dose of cells comprises a defined number of CD8+/CCR7+ cells, CD4+/CCR7+ cells, CD8+/CD27+ cells, CD4+/CD27+ cells, CD8+/CCR7+/CD27+ cells, and/or CD4+/CCR7+/CD27+ cells.

プロンプト改善後の翻訳文

The method of claim 113, wherein the unit dose of cells comprises a defined number of CD8+/CCR7+ cells, CD4+/CCR7+ cells, CD8+/CD27+ cells, CD4+/CD27+ cells, CD8+/CCR7+/CD27+ cells, and/or CD4+/CCR7+/CD27+ cells.

表 5　分析具体例

LLM 勝ち例

原言語文

ヘキサメチレンジアミンとテレフタル酸から得られる構成単位、及び 11 －アミノウンデカン酸又はウンデカンラクタムから得られる構成単位を含有し、相対粘度 (RV) が式 (1) の範囲であり、アミノ基末端濃度 (AEG) 、 カルボキシ基末端濃度 (CEG) 及びモノカルボン酸でアミノ基末端を封鎖した末端濃度 (EC) の関係が式 (2) 及び (3) を満たす半芳香族ポリアミド樹脂。1.95 ≦ RV ≦ 3.50 ・ ・ (1) 10eq/t ≦ AEG+CEG ≦ 140eq/t ・ ・ (2) (AEG+CEG) / (AEG+CEG+EC) ≦ 0.50 ・ ・ (3)

参照訳文

wherein the resin contains a constituent unit obtained from hexamethylenediamine and terephthalic acid and a constituent unit obtained from 11-aminoundecanoic acid or undecane lactam,wherein a relative viscosity (RV) of the semi-aromatic polyamide resin satisfies the following formula (I): 1.95RV3.50, andwherein a relationship among a concentration of terminal amino groups (AEG), a concentration of terminal carboxyl groups (CEG) and a concentration of terminal amino groups blocked by a monocarboxylic acid (EC) satisfies the following formula (II): 10 eq/tAEG+CEG140 eq/t, and the following formula (III): (AEG+CEG)/(AEG+CEG+EC)0.50. formulac (II) and (III).

ベースライン

10 eq/tAEG+CEG140 eq/t (2)(AEG+CEG)/(AEG+CEG+EC)0.50 (3)

LLM

A semi-aromatic polyamide resin comprising a structural unit obtained from hexamethylenediamine and terephthalic acid and a structural unit obtained from 11-aminoundecanoic acid or undecane lactam, wherein the semi-aromatic polyamide resin has a relative viscosity (RV) in a range of formula (1), and a relationship between an amino group terminal concentration (AEG), a carboxy group terminal concentration (CEG), and a terminal concentration (EC) obtained by blocking an amino group terminal with a monocarboxylic acid satisfies formulas (2) and (3):

LLM 負け例

原言語文

前記癌が、転移性癌である、請求項 1～42 のいずれかに記載の方法。前記癌が、結腸直腸癌、卵巣癌、高悪性度漿液性卵巣癌 (HGSOC) 、 非小細胞肺癌 (NSCLC) 、 小細胞肺癌、肺腺癌、前立腺癌、去勢抵抗性前立腺癌、胆管癌 (bile duct cancer) 、 胆管癌 (cholangiocarcinoma) 、 黒色腫、子宮癌、甲状腺癌、膀胱癌、乳癌、子宮頸癌、胃癌、子宮内膜癌、肝細胞癌、白血病、リンパ腫、非ホジキンリンパ腫、骨髄腫、脳癌、神経芽細胞腫、扁平上皮癌、頭頸部扁平上皮癌 (HNSCC) および肛門扁平上皮癌 (SCCA) 、 肛門性器癌、直腸癌、膵臓癌、尿路上皮癌、肉腫および軟部組織肉腫、転移性結腸直腸癌 (CRC) 、 プラチナ抵抗性または不耐性 HGSOC、進行性 NSCLC、ならびに転移性去勢抵抗性前立腺癌 (mCRPC) 、 トリプルネガティブ乳癌、浸潤性乳癌、転移性乳癌、HER2 陽性乳癌、および炎症性乳癌からなる群から選択される状態または障害である、請求項 1～42 のいずれかに記載の方法。

参照訳文

The method of any of claims 1-42, wherein the cancer is a condition or disorder selected from the group consisting of: colorectal cancer, ovarian cancer, high grade serous ovarian cancer (HGSOC), non-small cell lung cancer (NSCLC), small cell lung cancer, lung adenocarcinoma, prostate cancer, castration-resistant prostate cancer, bile duct cancer, cholangiocarcinoma, melanoma, uterine cancer, thyroid cancer, bladder cancer, breast cancer, cervical cancer, gastric cancer, endometrial cancer, hepatocellular cancer, leukemia, lymphoma, Non-Hodgkin's lymphoma, myeloma, brain cancer, neuroblastoma, squamous cell carcinoma, head and neck squamous cell carcinoma (HNSCC), and squamous cell carcinoma of the anus (SCCA), anogenital cancer, rectal cancer, pancreatic cancer, urothelial carcinoma, sarcoma and soft tissue sarcoma, metastatic colorectal cancer (CRC), platinum-resistant or intolerant HGSOC, advanced NSCLC, and metastatic castration-resistant prostate cancer (mCRPC), triple-negative breast cancer, invasive breast cancer, metastatic breast cancer, HER2 positive breast cancer and inflammatory breast cancer.

ベースライン

The method of any of claims 1-42, wherein the cancer is a condition or disorder selected from the group consisting of colorectal cancer, ovarian cancer, high-grade serous ovarian cancer (HGSOC), non-small cell lung cancer (NSCLC), small cell lung cancer, adenocarcinoma of the lung, prostate cancer, castration-resistant prostate cancer, bile duct cancer, cholangiocarcinoma, melanoma, uterine cancer, thyroid cancer, bladder cancer, breast cancer, cervical cancer, gastric cancer, endometrial cancer, hepatocellular cancer, leukemia, lymphoma, non-Hodgkin's lymphoma, myeloma, brain cancer, neuroblastoma, squamous cell carcinoma of the head and neck, head and neck squamous cell carcinoma (HNSCC) and anal squamous cell carcinoma (SCCA), anogenital carcinoma, rectal cancer, pancreatic cancer, urothelial carcinoma, sarcoma and soft tissue sarcoma, metastatic colorectal cancer (CRC), platinum-resistant or

LLM

The method of any of claims 1-42, wherein the cancer is a metastatic cancer.