生成モデルは医療テキストの固有表現抽出に使えるか？

西山智弘 [1]　柴田大作 [2]　宇野裕 [2]　辻川剛範 [2]　北出祐 [2]　久保雅洋 [2]
矢田竣太郎 [1]　若宮翔子 [1]　荒牧英治 [1]
[1] 奈良先端科学技術大学院大学 先端科学技術研究科
[2] 日本電気株式会社 バイオメトリクス研究所
[1]{nishiyama.tomohiro.ns5,s-yada,wakamiya,aramaki}@is.naist.jp
[2]{shibata, yutaka_uno, tujikawa, t-kitade, masahirokubo}@nec.com

概要

医療テキストからの情報抽出など，大規模言語モデル（Large Language Model; LLM）の医療分野への利活用に期待が高まっているが，日本語での実施は少ない．本研究では，医療テキストからの固有表現抽出 (Named Entity Recognition; NER) について，生成モデルと既存の分類モデルを比較した．その結果，生成モデルにおいて固有表現の種類によっては，データが少数であっても多くのデータで fine-tuning した分類モデルに匹敵するような性能で NER が実現可能であることが明らかとなり，精度の高い生成モデルがアノテーション支援に役立つことが示唆された．

1　はじめに

大規模言語モデル（Large Language Model; LLM）ベースの生成モデル（以下，生成モデル）が医療系データセットのタスクの State-of-the-Art を大きく更新するなど LLM の医療分野への利活用には期待が集まっている [1]．特に，非構造データからの情報抽出の精度向上に期待が高まっている [2–4]．こうした情報抽出のための基礎技術として，固有表現抽出 (Named Entity Recognition; NER) があり，Hu らは，ChatGPT を用いて医療テキストからの NER を試みた [5]．Tang らは，生成モデルを用いて医療テキストからの NER のためにデータ拡張を行った [6]．また，医療ドメインにおける NER は一般ドメインのものより性能が劣ることが報告されている [5, 7]．

先行研究の多くは英語で実施されたものであり，生成モデルによる日本語の医療テキストからの NER は報告が少ない．国内の医療テキストは基本的に日本語で記述されるため，現状の生成モデルが日本語の医療テキストからの情報抽出においてどの程度活用できるかを明らかにすることは，今後の医療分野における LLM 応用において有益である．医療分野への自然言語処理の活用には，高いアノテーションコスト，少ない教師データなどが障壁となっており，少数の事例で高い精度での情報抽出が実現できれば，生成モデルをアノテーションの支援に転用できる．それに伴い，アノテーションコストなどが大幅に下げられる可能性があり [8]，医療系データセットの普及に役立つと考えられる．したがって，こうした情報抽出のための基礎技術である NER が，生成モデルを用いてどの程度の性能を発揮するか調査することは重要であると言える．

本研究では分類モデルである Bidirectional Encoder Representations from Transformers (BERT) と近年注目を集める LLM ベースの生成モデルを用いた医療テキストからの NER の性能を比較し，それぞれのモデルが有効に活用できる条件を調査した．実験では少数事例で学習させた生成モデルと，少数事例に加えて全データで学習させた分類モデルを比較した．

2　実験材料

症例報告コーパス (MedTxt-CR) [9] を用いた．これは，J-Stage で公開されている症例報告のうち，再配布の許諾を得た症例報告コーパス（148 文書）であり，文書中に出現する症状や治療，医薬品など 9 種類の固有表現に対してアノテーションがなされている．本研究では，この中から i2b2 2010 コーパス [10] に対応する 4 種類の固有表現 (*d*, *remedy*, *m-key*, *t-test*) を選択し，*d* を病名/症状 (*disease*), *remedy* と *m-key* を治療 (*treatment*), *test* を検査 (*test*) として使用した．

データセットは文章を文単位に分割し，訓練セットとテストセットに 4:1 の割合で分割した．表 1 に訓練セットとテストセットの統計情報を示す．

表 1　訓練セットとテストセットの統計情報

	訓練セット	テストセット
文数	1276	320
総エンティティ数	2939	824
ユニークエンティティ数	1834	519
1 文ごとの文字数 (SD)	48.7 (27.6)	47.9 (23.0)
1 文ごとの単語数 (SD)	30.3 (16.3)	30.4 (14.1)

3　実験

3.1　モデル

分類モデルとしては BERT [11] を，生成モデルとして ELYZA[1] および GPT-4 を用いた．使用した BERT[2] は日本語コーパス CC-100 (74.3GB), Wikipedia (4.9GB) で事前学習されたものであり，パラメータ数は 0.1B である．ELYZA は Meta 社の Llama-2-7b-chat に約 180 億トークンの日本語テキストを用いて追加学習されたモデルであり，パラメータ数は 7B である．GPT-4[3] に関しては事前学習やパラメータ数については公開されていないが，GPT-3 [12] のパラメータ数が 175B であるということを考えると，これより大きいものと推測される．

3.2　実験設定

生成モデルでは，文中に含まれる固有表現をリスト形式で抽出させた．ELYZA の生成条件として，最大出力トークンは 256 とし，トークンのサンプリングは行わなかった．GPT-4 では，モデルは gpt-4-1106-preview を，パラメータはデフォルト値を利用した．文ごとに図 1 に示すようなプロンプトを用いて事例を入力し，出力されたテキストを予測結果とした．

分類モデルでは，ラベリング方法として Inside–Outside–Beginning2 (IOB2) 形式を利用した．最大入力トークンは 512，Epoch 数は 100，学習率は 2.0×10^{-5}，バッチサイズは 32（訓練），256（検証），256（テスト），最適化関数には Adam を利用した．訓練としては，全ての訓練セットを用いて BERT を fine-tuning した他，生成モデルと詳細に比較するために few-shot による実験も実施した．

1）　https://huggingface.co/ELYZA/ELYZA-japanese-Llama-2-7b-fast-instruct
2）　https://huggingface.co/cl-tohoku/bert-base-japanese-v3
3）　https://openai.com/gpt-4

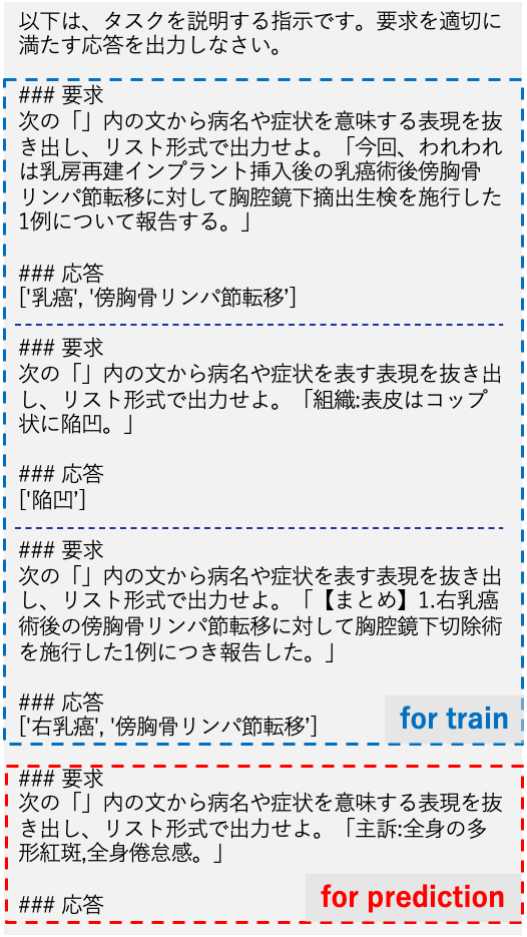

図 1　few-shot (n=3) のプロンプト例．青枠：訓練事例，赤枠：テスト事例

3.3　評価方法

評価指標には Micro-F1（以下、F1 スコア）の 5 回平均を利用した． 5 回の試行で生成モデルでは few-shot 時に訓練事例を変更した．BERT では seed 値および few-shot 時には合わせて訓練事例の変更を行った．予測結果の判定には完全一致（Exact）と部分一致（Partial）の 2 種類を用いた．例えば，Exact では，[正解: 紅潮 予測: 紅潮] を正解とみなすが，[正解: 顔面紅潮 予測: 紅潮] や [正解: 紅潮 予測: 顔面紅潮] という事例を不一致とみなす．Partial ではこれら全ての事例を正解とみなす．表 2 に各モデルで行った実験設定を示す．

一般的に NER の評価ではエンティティの位置情報（出現位置）の一致も含めて評価するが，生成モデルではエンティティの位置情報を適切に生成することが困難であったため，位置情報は用いず評価した．そのため，生成したエンティティの集合で評価することとし，同じエンティティの重複は除外する

表 2　実験設定

モデル	種別	zero-shot (n=0)	few-shot (n=1, 3)	all (n=1,276)
BERT	分類	-	✓	✓
ELYZA	生成	✓	✓	-
GPT-4	生成	✓	✓	-

表 3　正解と誤りタイプごとの出力例

種別	出力例
正解	[‘多形紅斑’, ‘全身倦怠感’]
形式誤り	多形紅斑、 全身倦怠感
	多形紅斑 \n 全身倦怠感
	*多形紅斑, * 全身倦怠感
内容誤り	承知しました．要求に応答を作成します．[‘多形紅斑’, ‘全身倦怠感’]

こととした．

3.4　生成モデルの出力の処理

生成モデルで予測した場合，表 3 に示すようにリスト形式で出力されなかったり，余分な区切り・表現が含まれるような事例が予測結果に含まれた．上記のような場合でも，エンティティが適切に判定されるように出力の後処理を行った．加えて，余分な表現は判定されないように，出力と予測元のテキストとのストリングマッチングを行い，出力のうち元のテキストに含まれていない表現は削除し，生成モデルの予測結果として扱った．

4　結果

表 4 に生成モデルと分類モデルによる NER の実験結果を示す．few-shot 時では生成モデルの ELYZA と GPT-4 の F1 スコアが分類モデルである BERT の F1 スコアよりも高かったが，最も F1 スコアが高かったモデルは，全訓練セットで fine-tuning した BERT であった．全訓練セットで fine-tuining した BERT を除いた，いずれのモデルにおいても，Partial のとき F1 スコアが大きく改善した．

タグ別に見るとタグごとで F1 スコアは大きく異なり，*disease* および *treatment* においては，全訓練セットで fine-tuning した BERT に匹敵する結果が，zero-shot および few-shot で学習させた GPT-4 で得られた．一方で，*test* においては，結果に大きな乖離が見られ，fine-tuning した BERT の方が高い F1 スコアであることが確認された．

5　考察

5.1　各モデルの予測結果の比較

表 5 に各モデルの予測結果と NER の誤りタイプを示す．誤りタイプは BERT と GPT-4 の出力に着目して分類した．

表 5 の (1) において，BERT では，「カルシウム結石」と「カルシウム結石症」というアノテーションにおける微妙な境界を判断できているが，GPT-4 では正解とは異なる境界を予測している．しかし，これはアノテーションの定義による影響であり，GPT-4 の予測は医学本質的には間違ってはいない．(4) も同様の誤りであり，エンティティの境界を厳密に判定する Exact を用いると，少ない事例で学習させた生成モデルに不利な結果を導きやすいと予想される．

(2) において，文脈からは本文に登場する「イレウス」は，「イレウスチューブ」であることが判別できるため，本文に登場する「イレウス」は病名ではない．BERT ではその判別ができているが，GPT-4 では「イレウス」を病名として抽出している．「イレウス」という単語そのものは病名であるため，単語そのものに着目して抽出された結果であると考えられる．こうした誤りはプロンプトの入力方法で改善できると思われる．

(3) において，「EL」（エンドリークの略語）が，BERT では抽出されなかった．他に，(5) のような TNM 分類など，英数字のみの表現も抽出ができない場合があった．GPT-4 では，英数字のみからなる表現される用語も抽出可能な場合があり，GPT-4 はこのような医療用語の知識を有することが示唆される．

5.2　分類・生成モデルの応用性

訓練データがごくわずかな場合，生成モデルの方が性能が優れていることが明らかとなった．特に，GPT-4 は zero-shot の場合であっても，抽出するタグの種類によっては部分一致において，十分なデータで訓練された BERT に匹敵する結果を示した．これは，生成モデルの可能性を裏付ける重要な結果である．一方で，十分なデータがあるのであれば，性能，速度，扱いやすさを考えると分類モデルが現状の第一選択になり得る．少量のデータに GPT-4 を利用するのであれば，zero-shot で 1 文あたり 0.4 円程

表 4 NER における F1 スコア（太字は同じ実験設定における最大 F1 スコア）

		BERT				ELYZA				GPT-4			
# of shots	Evaluation	*disease*	*treatment*	*test*	all	*disease*	*treatment*	*test*	all	*disease*	*treatment*	*test*	all
n=0	Exact	-	-	-	-	0.244	0.104	0.054	0.164	0.453	0.336	0.148	**0.339**
	Partial	-	-	-	-	0.396	0.315	0.203	0.333	0.712	0.610	0.250	**0.560**
n=1	Exact	0.109	0.045	0.039	0.091	0.274	0.241	0.172	0.243	0.495	0.352	0.283	**0.401**
	Partial	0.304	0.123	0.124	0.254	0.432	0.398	0.234	0.382	0.742	0.580	0.328	**0.596**
n=3	Exact	0.203	0.081	0.079	0.164	0.316	0.240	0.158	0.252	0.520	0.366	0.320	**0.428**
	Partial	0.461	0.195	0.206	0.373	0.472	0.380	0.206	0.376	0.759	0.582	0.361	**0.617**
n=1,276 (all)	Exact	0.759	0.620	0.813	**0.719**	-	-	-		-	-	-	
	Partial	0.759	0.620	0.813	**0.719**	-	-	-		-	-	-	

表 5 正解と各モデルの出力結果例

	誤りタイプ	本文	正解	BERT 予測	ELYZA 予測	GPT-4 予測
(1)	境界の相違	血清尿酸値，尿生化学について健常者，およびカルシウム結石症例とも比較した.	[‘カルシウム結石’]	[‘カルシウム結石’]	[]	[‘カルシウム結石症’]
(2)	文脈判定	炎症は限局しており腸管血流も保たれていたためイレウスチューブ挿入による保存的治療を開始した.	[‘炎症’]	[‘炎症’]	[‘イレウス’, ‘炎症’]	[‘イレウス’]
(3)	知識	3 年前に AAA に対して EVAR を施行され、術直後の type2EL、術 1 年後に認めた type1b・type2EL に対して追加加療を施行.	[‘AAA’, ‘type2EL’, ‘type1b・type2EL’]	[]	[‘AAA’, ‘EVAR’, ‘追加加療’]	[‘type2EL’, ‘type1b・type2EL’]
(4)	境界の相違	【目的】前立腺肥大 (以下 BPH) に対する低侵襲手術である、バイポーラシステムを利用した経尿道的前立腺核出術 (transurethral enucleation with bipolar: 以下 TUEB) の経験を報告する。	[‘前立腺肥大 (以下 BPH)’]	[‘前立腺肥大 (以下 BPH)’]	[‘低侵襲手術’, ‘経尿道的前立腺核出術’]	[‘前立腺肥大’]
(5)	知識	UM, Less, type3, 10*8.3cm, T4a (SE), N1, M0, StageIIIA であった。	[‘T4a’, ‘N1’, ‘M0’]	[]	[‘UM’, ‘10*8.3cm’]	[‘T4a(SE)’, ‘N1’, ‘M0’, ‘StageIIIA’]

度とコストは大きくないが，実用性を考えると予測モデルとして GPT-4 を大量の文章にそのまま適用することはコストの面で課題がある.

一方で，本実験結果とアノテーションコストを鑑みると，GPT-4 のアノテーターの支援への活用は十分検討できる．教師データ作成時に，人によるアノテーションの前に生成モデルを用いてラベル付きデータを作成する，という流れが NER の教師データを作成する際により一般的になっていくことが予想される.

6 おわりに

医療ドメインでの NER の性能として，データがわずかであれば，生成モデルの性能が分類モデルを上回ったが，データが十分にある場合には分類モデルが生成モデルの結果を上回った．生成モデルの中でもモデルにより大きく性能が異なり，GPT-4 が ELYZA を上回る結果であった.

医療テキストからの情報抽出を行うモデルとして分類モデルと生成モデルのどちらが望ましいか，という問いに対しては，データ量が十分にあれば分類モデルが現状では第一選択肢となると言える．ただ，データ数が限られた条件下では，生成モデルが優れた性能を発揮し，データ拡張などへの活用が期待できる．特に，医療ドメインなどアノテーションコストが高く，専門性を有すアノテーターの数が限られる場合には，精度の高い生成モデルがアノテーション支援に十分に役立つことが示唆される．今後は，NER の一つの手法として，データ整備に生成モデルを利用し，実文書の予測モデルには分類モデルを用いる，という流れがより一般的になっていくと考えられる.

謝辞

本研究は日本電気株式会社の 2023 年度研究インターンシップで実施された内容を奈良先端科学技術大学院大学と日本電気株式会社との共同研究において発展させたものであり、内容の一部は戦略的イノベーション創造プログラム (SIP3) の助成を受けたものである.

参考文献

[1] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. **Nature**, Vol. 620, No. 7972, pp. 172–180, August 2023.

[2] Amir Feder, Itay Laish, Shashank Agarwal, Uri Lerner, Avel Atias, Cathy Cheung, Peter Clardy, Alon Peled-Cohen, Rachana Fellinger, Hengrui Liu, Lan Huong Nguyen, Birju Patel, Natan Potikha, Amir Taubenfeld, Liwen Xu, Seung Doo Yang, Ayelet Benjamini, and Avinatan Hassidim. Building a Clinically-Focused Problem List From Medical Notes. In **Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)**, pp. 60–68, Abu Dhabi, United Arab Emirates (Hybrid), 2022.

[3] Fan Zhang, Itay Laish, Ayelet Benjamini, and Amir Feder. Section Classification in Clinical Notes with Multi-task Transformers. In **Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)**, pp. 54–59, Abu Dhabi, United Arab Emirates (Hybrid), 2022.

[4] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large Language Models are Few-Shot Health Learners. **arXiv**, 2023. preprint arXiv:2001.08361.

[5] Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot Clinical Entity Recognition using ChatGPT. **arXiv**. preprint arXiv:2303.16416.

[6] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? **arXiv**, 2023. preprint arXiv:2303.04360.

[7] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. GPT-NER: Named Entity Recognition via Large Language Models. **arXiv**, 2023. preprint arXiv:2304.10428.

[8] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. LLMs Accelerate Annotation for Medical Information Extraction. **arXiv**, 2023. preprint arXiv:2312.02296.

[9] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In **Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies**, pp. 285–296, 2022.

[10] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. **Journal of the American Medical Informatics Association**, Vol. 18, No. 5, pp. 552–556, September 2011.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. **arXiv**, 2020. preprint arXiv:2005.14165.

[13] 近江崇宏. Web コーパスからのノウハウの獲得. 言語処理学会年次大会発表論文集, pp. 350–352, 2021.

A　生成モデル（ELYZA）の特徴

検討の中で，SHOT 数が増えると，プロンプトに従えなくなり精度が下がる傾向にあることがわかった．例えば，次のような入力があるとする．

> 次の「」内の文から病名や症状を意味する表現を抜き出し、リスト形式で出力せよ．「今回我々は腹腔内遊離ガスを伴った気腫性胆嚢炎の症例を経験し救命し得たのでこれを報告する．」

このときの出力は 3-SHOTS 時では，

> ['腹腔内遊離ガス', '気腫性胆嚢炎']

であるが，15-SHOTS 時には以下のようになった．

> 承知しました．要求に応じて、「病理組織学的所見は」「抗核抗体」などのように、単語のみのリストを作成します． ■病理組織学的所見は 'adenocarcinoma,MP,N0,M0 StageI' ■抗核抗体 ' 抗核抗体’

このように，従って欲しい指示に従わず，プロンプト中に含まれる，事例を学習するための入力からトークンを抽出してしまう事例やリスト形式で出力できない事例が目立った．定量的な解析のため，各入力に対してその出力形式がリスト形式であるか判定し，各入力文字数に対する頻度を解析した．図にこれを示す．この図からも，実際に入力文字数が増えるに従い，リストの形式で出力されなくなるということがわかった．入力文が長くなると，モデルが指示を適切に把握できなくなる傾向にあることが示唆される．

B　一般ドメインでの精度比較

一般ドメインにおける生成モデル（ELYZA）の抽出精度についても比較検討した．この検討には医療ドメインのデータセットとして MedTxt-CR を，一般ドメインのデータセットとしてストックマーク株式会社により作成された Wikipedia の記事に対して 8 種類の固有表現を付与されたものを使用した [13]．また，固有表現タグは人名，組織名，法人名に限定して実験したところ，一般ドメインにおける結果の方が，医療ドメインにおける結果を上回っていた．この傾向は，生成モデルを利用した情報抽出において医療ドメインでは精度が低い傾向にあるという先

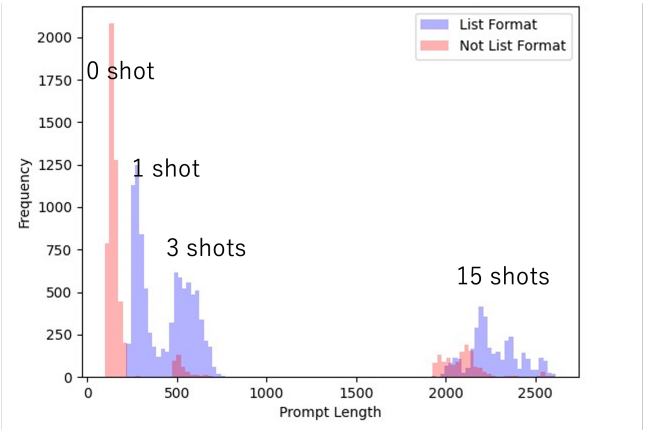

図 2　プロンプトの長さと出力形式の頻度（青：プロンプトの指示通りにリスト形式で出力，赤：プロンプトの指示に従わないリスト形式ではない形で出力）

表 6　一般ドメインにおける生成モデルの NER による F1 スコア

Data	MedTxt		Wikipedia	
# of Shots	Exact	Partial	Exact	Partial
n=0	0.164	0.333	0.244	0.384
n=1	0.243	0.382	0.355	0.417
n=3	0.252	0.376	0.393	0.438

行研究の結果とも一致する．例えば，Wang らによる CoNLL 2003 のデータを利用した一般ドメインでの結果は，F1 スコアで 0.909 であり [7]，Hu らによる i2b2 を利用した医療ドメインでの結果は F1 スコアで 0.620 であった [5]．