In-Context Learning において LLM はフォーマットを学べるか

坂井吉弘 [1]　趙羽風 [1]　井之上直也 [1,2]
[1] 北陸先端科学技術大学院大学　[2] 理化学研究所
{y.sakai, yfzhao, naoya-i}@jaist.ac.jp

概要

In-**C**ontext **L**earning (文脈内学習；ICL) は，プロンプト中に与えられた少数のデモなどからパラメータを更新することなくタスクを学習する LLM の能力であるが，そのメカニズムは十分に明らかにされていない．先行研究の実験は，「タスクの入力の後にラベルを出力する」というフォーマットを LLM に示すことが特に重要である可能性を示唆する．そこで本研究では，LLM が与えられたデモから答え方のフォーマットを学習する様子を直接的に可視化した．結果として，(1) 確かに LLM はデモから答え方のフォーマットを学んでいること，(2) フォーマットの学習は意味の無いラベルについても可能であること，(3) 最悪のラベルが ICL の Macro-F1 を大きく向上させることを発見した．

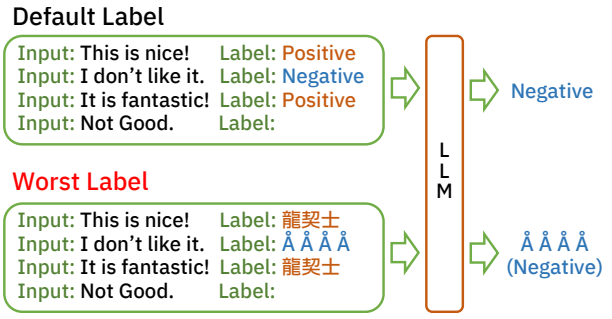

図 1　In-Context Learning の様子と各ラベル空間による実験の例 少数の入出力のデモとクエリをフォーマットに従って LLM に与え，ラベル空間のラベルのうち予測確率の高いものを予測結果として採用する．実際に用いるフォーマットは付録 B に示す．また，ラベルにはデータセットに設定された意味のあるラベル (Default Label) と，Zero-Shot 時に予測確率の低かった最悪のラベル (Worst Label) を用いる．

1　はじめに

Large **L**anguage **M**odel (大規模言語モデル；LLM) は，図 1 のように少数の入出力のペア (デモ) に基づいて，新しい入力 (クエリ) に対して答えを推論することが出来る．この LLM の利用法又は能力を **I**n-**C**ontext **L**earning (文脈内学習；ICL) と呼ぶ [1].

ICL のメカニズムに関する特に重要な問いの一つに「LLM は真にデモからタスクを学んでいるのか？」というものがある [2]．いくつかの先行研究 [3, 4] は，プロンプト中に与えられたデモを元に LLM が実際に新たなタスクを解く能力を獲得していると考える立場に立つ．しかし，一部の研究 [5, 6, 7, 8] では，LLM が事前学習の段階で暗黙のうちにタスクを解く能力を獲得しており，ICL のデモは単に解くべきタスクの識別のために用いられているに過ぎない可能性が示唆されている．例えば Min ら [8] は，LLM に提供するデモの出力例を全て誤ったものにしても性能の低下は限定的であったのに対して，出力例を提供しない場合には大きく性能が下がることを報告した．そしてこれらを根拠に，ICL において LLM は新たにタスクを学習しているわけではなく，ICL のデモは「タスクの入力の後にラベルを出力する」という答え方のフォーマットそのものを獲得することに寄与している可能性があると指摘した．しかし，我々の知る限りこの仮説について直接的に取り組んだ研究は存在しない．

そこで本研究では，ICL において LLM がデモから答え方のフォーマットを学習しているかどうかを確かめるための指標として，語彙中の予測確率の順位に着目し，これを可視化・分析した．本研究の主な貢献は以下の通りである．

- ICL において LLM が答え方のフォーマットを学習しているという仮説の実証的な証拠を得ることに成功した (§3.1.2, §3.2.2).
- クエリの文章からは予測され得ないような最悪のラベル (Worst ラベル) を利用しても同様の学習が行われていることを発見した (§3.1.2).
- Worst ラベルを利用したときに ICL の Macro-F1 性能が向上することを発見した (§4).

2　評価の方法

本節では ICL を定式化した上で，LLM が ICL においてフォーマットを学ぶとはどのようなことなのか，またその程度はどのようにして観察出来るのかについて考察する．

2.1　準備：ICL の定式化

始めに，自然言語で記述されたタスクの入力 x と出力ラベル $y \in \mathbb{U}$ からなる教師ありデータセット $\mathcal{D} = \{(x_i, y_i)\}, i = 1 \ldots n$ から，クエリデータ x_q とデモデータ $\{(x_{d_j}, y_{d_j})\}, j = 1 \ldots k$ を得る．ここで，$\mathbb{U}$ はラベル空間，n はデータセットサイズ，k はデモ数である．デモデータは $q \neq d_j$ となるように選択する．入力するプロンプト s は，フォーマットパターン f を用いて $s = f(x_{d_1}, y_{d_1}, \ldots, x_{d_k}, y_{d_k}, x_q)$ のように生成する．これを言語モデル $P(\cdot)$ に与えたとき，ラベル空間のラベルの中で予測確率が最も高かったものを，クエリに対する予測結果 $\hat{y}_q$ とする．即ち以下の式で ICL の予測結果を得ることが出来る．

$$\hat{y}_q = \underset{l \in \mathbb{U}}{\operatorname{argmax}} P(l|s) \tag{1}$$

2.2　フォーマットを学習するとは何か

ICL のフォーマットパターン f には指示や説明を加えるものなど様々なものが考えられるが，本研究では図 1 のような極めて単純なフォーマットを用いる．実際のフォーマットは付録 B に記載する．この場合，フォーマットを学習するということは「”Input: ”の後に与えられた文章について，”Label: ”の後にラベル空間内のラベルを出力すべきこと」を学習することだと言えるだろう．そこで本研究では特に，LLM が答え方のフォーマット即ち「**”Label: ”の後にラベル空間内のラベルを出力すべきこと**」を認識しているかどうかを確かめるために，デモ数 k の時の語彙全体におけるラベル空間のラベルの予測確率の平均順位

$$\frac{1}{|\mathbb{U}|} \sum_{l \in \mathbb{U}} \mathrm{Rank}(P, s, l) \tag{2}$$

を観察する．ただし，$\mathrm{Rank}(P, s, w)$ は言語モデル P に入力 s を与えたときの全語彙中における単語 w の予測確率の順位を指す．仮に LLM がデモから答え方のフォーマットを学んでいるのであれば，デモ数の増加に伴ってラベルの平均順位は改善されるはずである．

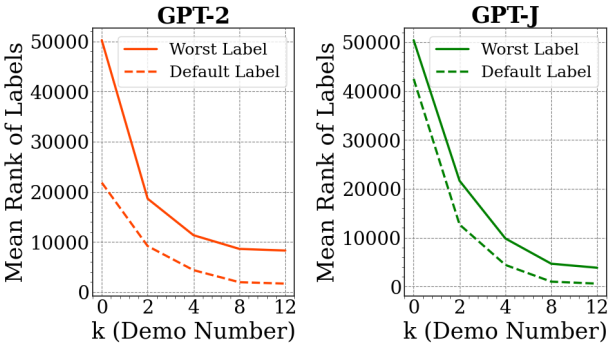

図 2　GLUE-RTE タスクにおけるデモ数 k とラベルの平均順位の関係 Default ラベル，Worst ラベルともにデモ数の増加に伴い平均順位が改善しており，LLM がフォーマットを学習していることが分かる．

2.3　最悪のラベルはラベル空間として認識されるか

通常 ICL においては，ラベル空間には {”Positive”, ”Negative”} のように，データセットに予め設定された意味のあるラベル (**Default** ラベル) が用いられる．しかし，仮に本当に LLM がデモの情報から答え方のフォーマットを学習し，ラベル空間からラベルを選んで出力すべきことを認識しているのであれば，よりタスクと無関係で，予測確率の低いラベルでも同様の結果となるはずである．そこで，本研究ではクエリの文章からはまるで予測され得ないようなラベル (**Worst** ラベル) をラベル空間に選択し，フォーマットが学習されるかを確認する．Worst ラベルには Zero-shot 時，即ち $k = 0$ の時に予測確率が最も低かったトークンの中で有効なものを選択する．

3　LLM はフォーマットを学ぶのか

以上の議論を踏まえ，ICL において LLM がフォーマットを学んでいるかどうかを確かめるために 2 つの実験を行った．

3.1　デモ数とラベルの平均順位の関係

まずは，LLM がラベル空間を学んでいるかを確かめるために，デモ数とラベルの平均順位の関係を調べた．

3.1.1　実験設定

データセット　7 個の分類タスクのデータセットについて実験を行った．本研究では，実験の都合上これらのデータセットを一部加工している．詳細は付録 A に示す．

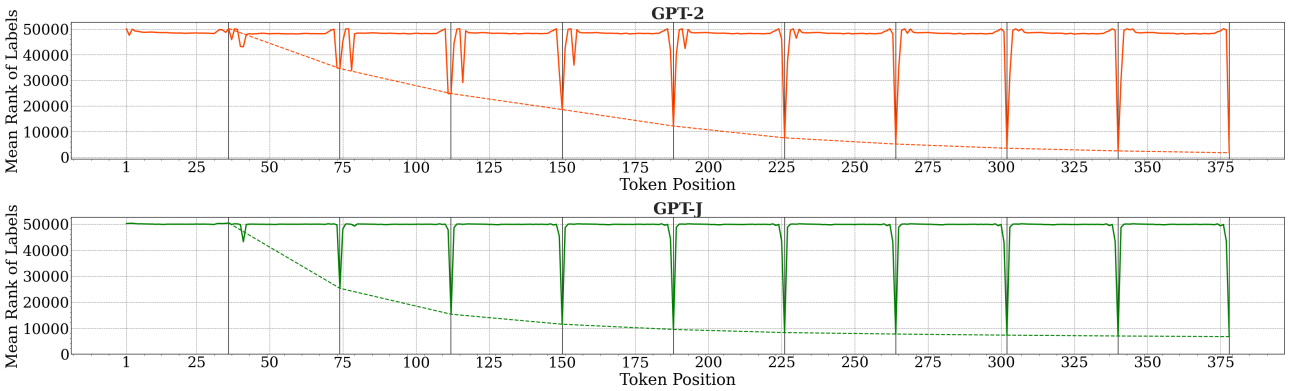

図 3　AG News タスクにおけるトークン位置と Worst ラベルの平均順位の関係 ”Label: ”の位置を実線で示し，またその時の平均順位のみを結んだ線を点線で示している．”Label: ”の位置で特徴的にランクの平均順位が改善しており，それ以外の部分では殆ど予測確率が上がっていないことが分かる．

モデル　GPT-2 [9] と GPT-J [10] を用いて実験を行った．パラメータは huggingface にアップロードされているものを用いた．

ラベル空間　Default ラベルには予めデータセットに設定されているものを用いた．Worst ラベルは §2.3 で説明した通り，各データ毎に Zero-shot 時に予測確率の低い単語の中から有効なものを選んだ．

3.1.2　結果と考察

実験結果を図 2 に示す．なお紙幅の都合上，GLUE-RTE [11] タスクの結果のみを紹介するが，この他のデータセットについても平均順位の値が改善し，収束していくようなグラフが得られた．

LLM は ICL においてラベル空間を獲得する　図 2 より，デモ数の増加に伴ってラベルの平均順位が改善していることから，LLM がデモからラベル空間を学んでいることが分かる．

最悪のラベルでもラベル空間として認識する　図 2 より，Worst ラベルでも Default ラベルと同様に平均順位が改善している．これは，ラベル空間の学習がラベルの意味に依らずに行われていることを意味する．また，この結果はラベルが無意味なものであってもラベル空間として認識可能であることを意味し，ラベル空間を代替出来る可能性を示唆する．

3.2　トークン位置毎のラベルの平均順位

デモ数の増加に合わせてラベルの平均順位が改善していることから，LLM が ICL のデモからラベル空間を学習していることが判明した．しかし，この結果から直ちに「**”Label: ”の後に**ラベル空間のラベルを出力すべきこと」を学習しているとは言えない．なぜならば，プロンプト中に何度も出現したラベル単語の予測確率を，文章中の位置に関係なく闇雲に高く評価しているだけの可能性も考えられるからだ．そこで，本当に「**”Label: ”の後に**ラベル空間のラベルを出力すべきこと」を学習しているかを調べるために，トークン長を揃えたデータを抽出し (トークン長 30 でデータ数 1022 個) で各トークン位置におけるラベルの平均順位を観察する．

3.2.1　実験設定

基本的には §3.1 と同様の実験設定で行うため，ここではこの実験特有の設定についてのみ説明する．

データセット　AG News データセット [12] から，トークン長が等しいデータのみを取り出して用いる．これは，全てのデータについて”Label: ”が入力されるトークン位置を統一するためである．

ラベル空間　ラベル空間には Worst Label のみを用いる．これは，”Label: ”のタイミング以外ではラベルの平均順位が改善しないこと（最悪のままであること）を確認するためである．

3.2.2　結果と考察

実験結果を図 3 に示す．GPT-2，GPT-J 共に”Label: ”の入力の直後にラベル空間の平均順位が著しく改善していることが分かる．従って，LLM はラベルの出現回数に比例して闇雲にラベル空間の予測確率を上げているわけではなく，「”Label: ”の後にラベルを出力する」という答え方のフォーマットを認識していることが分かる．また，その平均順位はデモ数の増加に伴い改善しており，デモから答え方のフォーマットを学習していることが分かる．これは

表 1　デモ数 k が最大の時の各タスクにおける Accuracy(Acc.) と Macro-F1(MF1) より数値の高いものを**太字**で示し，標準偏差を小文字で示す．モデルを問わず，殆どのタスクでは Worst ラベルを用いた時に Macro-F1 が改善している．

Dataset			PS	FP	SE'14R	SE'14L	RTE	MRPC	Ethos	Mean
GPT-2	Acc.	**Default**	**62.21**$_{0.28}$	**46.13**$_{0.96}$	39.97$_{1.40}$	**43.21**$_{\mathbf{1.45}}$	**49.74**$_{1.57}$	**66.28**$_{0.04}$	43.54$_{0.12}$	**50.15**
		Worst	43.14$_{0.84}$	45.60$_{0.74}$	**49.06**$_{1.66}$	37.51$_{1.11}$	49.51$_{1.06}$	53.62$_{0.81}$	**50.61**$_{0.81}$	47.01
	MF1	Default	20.21$_{0.52}$	31.58$_{2.08}$	34.00$_{1.63}$	**39.56**$_{1.39}$	46.40$_{1.61}$	39.86$_{0.01}$	30.47$_{0.16}$	34.58
		Worst	**26.34**$_{0.78}$	**32.44**$_{0.75}$	**35.57**$_{1.74}$	34.56$_{1.11}$	**49.44**$_{1.07}$	**49.69**$_{0.91}$	**49.08**$_{0.80}$	**39.59**
GPT-J	Acc.	Default	**62.63**$_{0.05}$	42.46$_{0.07}$	42.69$_{1.04}$	32.53$_{0.60}$	48.86$_{0.80}$	34.9$_{0.48}$	**63.38**$_{1.26}$	46.78
		Worst	39.08$_{1.20}$	**51.47**$_{\mathbf{1.00}}$	**52.77**$_{1.22}$	**41.79**$_{0.87}$	**50.94**$_{1.05}$	**52.86**$_{0.83}$	51.70$_{0.87}$	**48.66**
	MF1	Default	19.35$_{0.16}$	20.11$_{0.56}$	33.96$_{1.14}$	25.53$_{0.96}$	41.07$_{1.12}$	27.88$_{0.67}$	**62.79**$_{1.34}$	32.96
		Worst	**26.56**$_{1.51}$	**38.04**$_{0.94}$	**41.66**$_{1.69}$	**39.30**$_{0.83}$	**50.93**$_{1.05}$	**50.02**$_{0.85}$	51.60$_{0.85}$	**42.59**

§3.1 の実験の結果とも整合する．

4　Default ラベル vs. Worst ラベル

§3.1.2 で議論したように，フォーマット学習という観点からは最早 Default ラベルを用いる必然性はなく，ラベル空間は変更しても構わないと言える．そこで，ラベル空間の変更が ICL 全体の性能にどのように影響を与えるのかについて調査した．結果を表 1 に示す．

Worst ラベルが Macro-F1 を改善する　表 1 より，モデルに依らず殆どのデータセットでは Worst ラベルが Macro-F1 を改善していることが分かる．Accuracy の変化に対して Macro-F1 の変化が著しいことから，Worst ラベルを用いた時は各クラスについてより平等に評価している（或いは，Default ラベルがバイアスをかけた不平等な評価をしている）と言える．

4.1　なぜ Worst ラベルが Macro-F1 を改善したか

Worst ラベルは意味を持たない　Worst ラベルには，英語用のトークナイザになぜか含まれる「龍契士」など，単に意味の無い単語が多く出現した．これらの意味の無いラベルが，ICL においてタスクを学習するのに優位に働いた可能性がある．これは文献 [13] の報告とも一貫する．

Worst ラベルは公平である　Worst ラベルは順位の隣り合うものから選んでいるため，学習を始めた時点では殆ど同じ予測確率である．これが ICL の学習または識別をバイアスなく素直に出力することに寄与した可能性がある．これは文献 [14] の報告とも一貫する．

タスク毎に最適なラベルが異なる　図 4 のように，一部のタスクでは Default ラベルの方がスケールしたのに対して他のタスクでは Worst ラベルの方がスケールした．このように，タスク毎に ICL の性能をスケールさせることが出来る特有のラベルが存在する可能性がある．

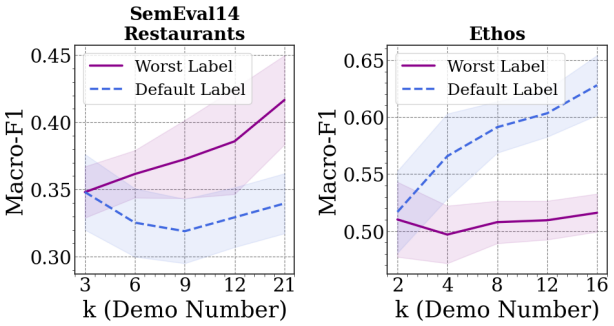

図 4　2 つのタスクにおけるデモ数 k と Macro-F1 の関係
モデルは GPT-J．薄い色の領域は 95%信頼区間を示す．選択するラベルによってデモ数を増やした時の振る舞いが大きく異なるのが分かる．

5　終わりに

本研究では，ICL において答え方のフォーマットの学習が行われているという仮説について，語彙中の予測確率の順位を可視化・分析することにより，実証的な証拠を得ることに成功した．本研究の結果は，LLM が ICL においてどのような能力を発揮しているのかという問いについて，タスクの学習能力，タスクの識別能力に加え，フォーマットへの適合能力という新たな能力についての視座を据えるための礎となるものである．なお，本研究で得られた結果は極めてシンプルなフォーマットにおける結果であり，説明や指示を加えた場合においても同様の結果が見られるかは今後の研究課題である．

謝辞

本研究は JSPS 科研費 19K20332 の助成を受けたものです．また論文の執筆に際して，JAIST の原口大地氏，石井晶氏から有益な助言を頂きました．この場を借りて感謝申し上げます．

参考文献

[1] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. **arXiv preprint arXiv:2301.00234**, 2022.

[2] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning" learns" in-context: Disentangling task recognition and task learning. **arXiv preprint arXiv:2305.09731**, 2023.

[3] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In **International Conference on Machine Learning**, pp. 35151–35174. PMLR, 2023.

[4] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 4005–4019, 2023.

[5] Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. Understanding in-context learning via supportive pretraining data. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12660–12673, 2023.

[6] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Pre-training to learn in context. **arXiv preprint arXiv:2305.09137**, 2023.

[7] Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 18878–18891, 2022.

[8] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 11048–11064, 2022.

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[10] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[11] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In **International Conference on Learning Representations**, 2018.

[12] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 28. Curran Associates, Inc., 2015.

[13] Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. **arXiv preprint arXiv:2305.08298**, 2023.

[14] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8086–8098, 2022.

[15] Emily Sheng and David Uthus. Investigating societal biases in a poetry composition system. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, **Proceedings of the Second Workshop on Gender Bias in Natural Language Processing**, pp. 93–106, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.

[16] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. **Journal of the Association for Information Science and Technology**, Vol. 65, No. 4, pp. 782–796, 2014.

[17] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, **Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)**, pp. 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[18] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: an online hate speech detection dataset, 2020.

付録

A　データセット

§3.1 及び表 1 の実験で使用したデータセットの一覧を以下に記載する．

本研究では，訓練データとテストデータを結合し，区別なく一つのデータセットとして用いている．また，データ数が 1000 を超えるものについては 1000 を上限とし，各データセットで 10 回実験を行った．モデルの入力長制限の関係から一部の長さの大きなデータについて除去している．

表 2　§3.1 及び表 1 の実験で使用したデータセットの一覧

表 1 における略称	データセット名
PS	Poem Sentiment [15]
FP	Financial Pharasebank [16]
SE'14R	SemEval 2014-Task 4 Restaurants [17]
SE'14L	SemEval 2014-Task 4 Laptops [17]
RTE	GLUE-RTE [11]
MRPC	GLUE-MRPC [11]
Ethos	Ethos [18]

B　フォーマット

本研究で用いたプロンプトについて以下に示す．データセットによって異なるプロンプトを用いている．(x, y) の形式のような 1 つの入力を取るようなタスクについては，以下のプロンプトを用いる．なお，AG News[12] データセットに限り，入出力の前後に”*/*”を追加している．

```
Input: <x>, Label: <y> \n
...
Input: <x>, Label:
```

$((x, a), y)$ の形式のような入力を取る Aspect-Based Sentiment 分析タスクについては，以下のプロンプトを用いる．

```
Input: <x>, Aspect: <a>, Label: <y> \n
...
Input: <x>, Aspect: <a>, Label:
```

$((x_1, x_2), y)$ の形式のように 2 つの文章を入力に取るタスクについては，以下のプロンプトを用いる．

```
Input: <x1>, Text 2: <x2>, Label: <y> \n
...
Input: <x1>, Text 2: <x2>, Label:
```