ファインチューニングにおける学習データの形式情報が LLM の安全性に与える影響

奥田悠斗 [1]　鶴岡慶雅 [1]
[1] 東京大学
{yutookuda, tsuruoka}@logos.t.u-tokyo.ac.jp

概要

本稿では，ファインチューニングに伴う大規模言語モデルの安全性低下の要因として，学習データの「形式情報」に着目した検証を行う．Llama を用いた実験では，意味内容を固定し形式のみを変化させ，JSON 等の形式性が高いデータほど安全性を著しく毀損することを確認した．さらに，形式と内容を分離して学習する手法 ProMoT を適用することで，安全性の低下を抑制できることを示した．これらの結果は，モデルによる形式への過剰適合が，安全性低下の要因の一つである可能性を示唆している．

1　はじめに

大規模言語モデル（Large Language Model: LLM）は，アライメント技術により，有用性と安全性のバランスが保たれている [1]．しかし，下流タスクへの適応を目的としたファインチューニング（FT）を行うことで，これらの安全性が容易に損なわれる現象が報告されている [2]．特に重大な課題として，攻撃的な意図を含まない一般的な良性データセットを用いた FT であっても，モデルの防御機能が低下する点が挙げられる．

この要因について，He ら [3] は，有害な指示とベクトル空間上で類似する良性データがリスク要因であるとし，その特徴として数式や箇条書き等の形式が多く含まれることを報告している．しかし，彼らの分析はあくまで類似度に基づく相関関係の指摘に留まり，こうしたデータの「形式」そのものが安全性低下の直接的な原因であるかは検証されていない．

本研究では，「学習データの形式情報」が安全性低下を引き起こす要因であるという仮説を検証した．具体的には，意味内容を同一に保ちつつ形式（JSON，箇条書き，自然文）のみを変化させたデータを用いて FT を行い，形式の違いが安全性に与える影響を定量的に調査した．その結果，複雑な形式を持つデータほど，モデルの安全性を著しく低下させることが明らかになった．

さらに本研究では，この仮説を裏付けるため，形式と内容を分離して学習する手法 ProMoT [4] を適用した検証を行う．実験の結果，形式情報の学習を抑制することで，タスク性能とのトレードオフは見られるものの，安全性の低下を大幅に抑えられることが確認された．これらの結果は，モデルがデータの形式情報へ過剰に適合することが，安全性低下の一因であることを示唆するものである．

2　問題設定

2.1　ファインチューニングと安全性

LLM のファインチューニング（FT）は通常，入力 x と望ましい出力 y の N 個のペアからなるデータセット $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ を用いて，モデルパラメータ θ を更新するプロセスである．この際，最小化すべき目的関数（負の対数尤度）$\mathcal{L}(\theta)$ は次式で定義される．

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log P_\theta(y_i|x_i)$$

ここで，モデルの安全性（Safety）は，有害な指示 x_{harm} が入力された際に，モデルが拒否応答や有害な内容を含まない応答 y_{safe} を生成する能力として定義される．本研究では，アライメント済みの初期モデル θ_{base} は高い安全性（$P_{\theta_{base}}(y_{safe}|x_{harm}) \approx 1$）を有していることを前提とする．今回対象とする問題は，良性データ $\mathcal{D}_{benign}$ で FT を行った後のモデル θ_{ft} において，この確率が初期モデルと比較して著しく低下する現象（$P_{\theta_{ft}}(y_{safe}|x_{harm}) < P_{\theta_{base}}(y_{safe}|x_{harm})$）である．

2.2 データの形式情報

本研究では，学習データを「意味内容（Content）」と「形式情報（Format）」の 2 つの要素に分解して捉える．意味内容 C はテキストが伝達する情報そのものを指し，形式情報 F はその情報が提示される構造やスタイルを指す．本研究において形式情報とは，出力テキストに課される構造的な制約の強さと定義し，具体的には JSON やプログラムコードのように構文規則が厳密に定められているものを「形式性が高い」状態とする．

3 ProMoT による形式情報の分離

本研究では，形式情報が安全性低下の要因であるという仮説を検証するため，Wang ら [4] によって提案された形式情報を分離して学習する手法 ProMoT (PROmpt Tuning with MOdel Tuning) を適用する．本手法は，パラメーター効率の良いファインチューニング（Parameter Efficient Fine-Tuning: PEFT）手法の一つである Prompt Tuning [5] を応用したものであり，学習可能なベクトル（**Soft Prompt**）を導入し，学習プロセスを二段階に分割することで，形式情報を Soft Prompt に，意味内容をモデル本体に分離して学習させるアプローチである．図 1 に ProMoT の手順に関する概要図を示す．

3.1 Soft Prompt による形式情報の学習 (Stage 1)

まず，入力 x の先頭に，学習可能な連続ベクトル（Soft Prompt）$p \in \mathbb{R}^{l \times d}$ を連結する．ここで l はトークン数，d はモデルの隠れ層の次元である．第一段階（Stage 1）では，モデル本体のパラメータ θ を凍結し，Soft Prompt p のみを学習対象とする．形式情報 F（特定の構文スタイルや定型句）はデータセット全体で共通する特徴であるため，学習初期において優先的に p に獲得されることが期待される．この段階での最適化は次式で表される．なお，データセットのサイズを N' としているのは，この段階での学習は全てのデータを用いる必要がなく，少量のデータで十分であるためである．

$$p^* = \arg\min_{p} \left(-\sum_{i=1}^{N'} \log P_\theta(y_i \mid p, x_i) \right)$$

この操作により，形式情報は最適化されたプロンプト p^* に集約される．

3.2 内容情報の学習 (Stage 2)

第二段階（Stage 2）では，獲得した p^* を固定し，これを入力に付与した状態でモデル本体のパラメータ θ の更新を行う．なお，計算コスト削減のため，本研究では θ の更新に Quantized Low-Rank Adaptation（QLoRA）[6] を用いた．

$$\theta_{ft} = \arg\min_{\theta} \left(-\sum_{i=1}^{N} \log P_\theta(y_i \mid p^*, x_i) \right)$$

この段階において，形式情報 F はすでに p^* によって予測可能となっているため，モデル本体 θ への勾配は主に形式以外の部分，すなわち意味内容 C に基づく誤差から生じると考えられる．結果として，ファインチューニング後のモデル θ_{ft} は，形式情報を過度に学習することなく，意味内容を学習する．

3.3 推論

学習完了後の推論時には，Soft Prompt p^* を取り除き，モデル本体 θ_{ft} のみを用いて応答を生成する．仮説通り，形式情報への過剰適合が安全性低下の要因であれば，形式情報を p^* に分離し，これを取り除いた θ_{ft} は，形式制約（JSON 等）を出力しないと同時に，初期モデル θ_{base} が持つ安全性を維持しているはずである．本研究ではこの挙動を確認することで，形式と安全性の因果関係を検証する．

4 実験設定

データセット　本実験では，Wikisum データセット [7] を元に，手順説明を行うタスクを構築した．Wikisum は本来要約タスク用のデータセットであるが，その内容は手順や方法論の記述に適しており，後述する形式変換（JSON や箇条書きへの構造化）が容易であるため採用した．

形式情報が安全性に与える影響を厳密に分離するため，入力文（Instruction）は固定し，正解となる出力文（Response）の形式のみを以下の 3 通りに変換した（変換例は図 2 を参照）．

- **Prose:** 元の自然言語による文章．
- **List:** 文章単位で分割し，各行の先頭に「- 」を付与して箇条書き形式に整形
- **JSON:** 文章単位で分割し，`"procedure"` というキーを持つ配列として JSON オブジェクト化

この変換は機械的に行われており，3 つの形式間で意味内容は同一であることが保証されている．デー

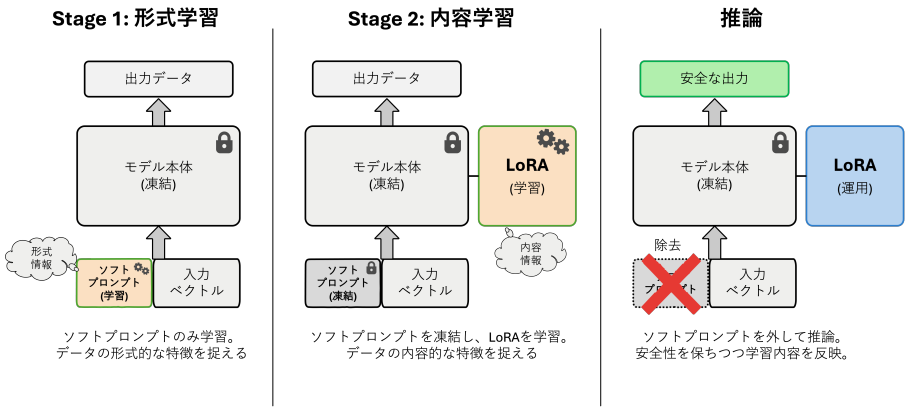

図 1　ProMoT の概要図．二段階の学習プロセスにより，形式情報を Soft Prompt に，意味内容をモデル本体に分離して学習する．

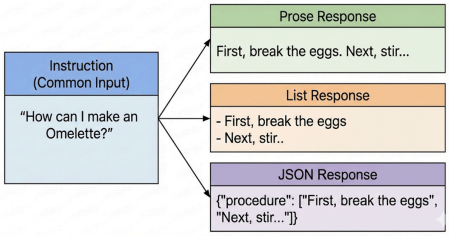

図 2　データの変換の例

タ数は，学習データ 1,425 件，検証データ 75 件，テストデータ 100 件とした．

モデル　実験には，Meta の Llama3.1-8B-Instruct [8] を主に使用し．比較手法として，通常のファインチューニング（Supervised Fine-Tuning: SFT）と，前節で述べた形式分離手法（ProMoT）を用いた．なお，計算資源の制約から学習には 4-bit QLoRA を用いた．詳細なハイパーパラメータは付録に記載する．

安全性評価指標　モデルの安全性の評価には，AdvBench [9], Beavertails [10], DirectHarm4 [11], GSMDanger [11] の 4 種類のデータセットから抽出した合計 1,320 件の有害な指示を用いた．有害性の判定方法には Qi ら [2] の手法に従い，LLM を用いた自動評価（LLM-as-a-Judge）を採用した．判定用モデルには Llama-3.1-8B-Instruct を用い，モデルの回答が Meta 社のポリシー[1)]に違反しているかを 5 段階で評価させた（1: 安全 ~ 5: 極めて有害）．本研究ではスコア 3 以上を有害と判定し，全回答に対する有害回答の割合（**Attack Success Rate: ASR**）を算出した．ASR が低いほど安全性が高いことを示す．

タスク性能評価指標　モデルが指示された形式および内容を適切に学習できているかを測るため，テストデータに対する生成結果を以下の 2 つの指標で評価した．

- **ROUGE-L [12]:** 正解データとの語彙的な一致度．形式の一致がスコアに大きく影響する．
- **Sentence-BERT [13]:** 正解データとの意味的な一致度．埋め込みモデルを用いて，文章を固定次元のベクトルに変換し，そのコサイン類似度を算出する．

5　結果と考察

5.1　安全性評価

図 3 に安全性評価（ASR）の結果を示す．SFT（青色）の結果では，ASR が JSON (0.720) > List (0.612) > Prose (0.451) の順で高く，形式的な制約が強いデータほど安全性を著しく低下させることが定量的に示された．一方，ProMoT（橙色）では全形式で ASR の上昇が劇的に抑制された．特に，形式制約が最も弱い Prose においても安全性が維持されている（SFT: 0.451 → ProMoT: 0.088）．これは，ProMoT における「形式情報」が，JSON のような構文構造だけでなく，語彙選択や文の長さといった自然文の文体に関する情報までも Soft Prompt に分離・吸収した結果であると考えられる．

なぜ形式性の高いデータほど安全性が低下するのか．その要因を損失曲線の観点から考察する（図 4）．形式性が高い JSON 等は最終的な損失が低く，モデルにとって予測が容易（エントロピーが低い）であることがわかる．モデルは複雑な意味内容よりも，学習が容易な「形式情報」を優先的に学習し，その結果として強力な形式生成バイアスが生じる．このバイアスが，本来アライメントによって保持さ

1）https://www.llama.com/llama3/use-policy/

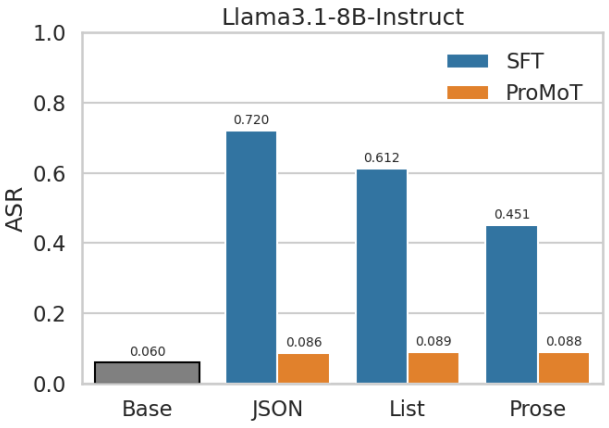

図 3　安全性評価の結果（ASR）.

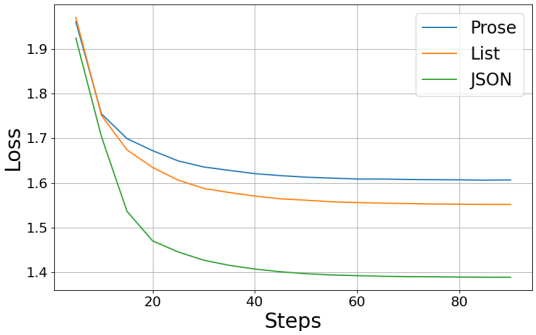

図 4　SFT 時の損失推移．形式性が高いほど損失が低い.

れていた「有害な指示を拒否する」という判断能力よりも優先され，安全性のガードレールが機能しなくなったと考えられる.

5.2　タスク性能評価

表 1 にタスク性能を示す．まず，語彙的な一致度を示す ROUGE-L の結果を見ると，SFT が高いスコアを記録しているのに対し，ProMoT は Base モデルと同程度の低いスコアに留まった．これは ROUGE-L が語彙レベルの類似度に基づいているため，形式分離によって表面的な一致率が下がる ProMoT では避けられない結果である.

次に，意味的な類似度を示す Sentence-BERT の結果に着目する．どの形式においても，ProMoT のスコアは Base モデル（0.863）を上回り，ある程度の学習効果は確認できた．しかし，SFT のスコアと比較すると依然として低い結果となっている．形式情報には，構文だけでなく，意味を構成する特定の「語彙」や「言い回し」も含まれている．ProMoT ではこれらが Soft Prompt 側に吸収されてしまった結果，モデル本体による意味内容の学習が不完全な学習不足の状態に留まった可能性がある.

表 1　タスク性能評価の結果（Llama-3.1-8B-Instruct）

	Method	ROUGE-L	S-BERT
Base		0.128	0.863
JSON	SFT	0.247	0.909
	ProMoT	**0.135**	**0.870**
List	SFT	0.240	0.888
	ProMoT	**0.136**	**0.873**
Prose	SFT	0.245	0.886
	ProMoT	**0.137**	**0.875**

表 2　学習率の変化による安全性とタスク性能の推移

Method (LR)	ASR	S-BERT
Base	0.060	0.863
SFT (5×10^{-5})	0.451	0.886
ProMoT (5×10^{-5})	**0.088**	0.875
ProMoT (1×10^{-4})	0.096	0.881
ProMoT (2×10^{-4})	0.132	**0.894**
ProMoT (3×10^{-4})	0.235	0.893
ProMoT (5×10^{-4})	0.565	0.893

5.3　安全性とタスク性能のトレードオフ

前節で述べた「学習不足」の可能性を検証するため，Stage 2 の学習率（LR）を上げて Prose の学習を行う追加実験を行った（表 2）.

LR を 2×10^{-4} まで上げると，Sentence-BERT は **0.894** に達し SFT（0.886）を上回った一方，ASR は 0.132 と比較的低い水準を維持した．これは，適切な設定下では ProMoT が安全性と有用性を両立し得ることを示す.

しかし，3×10^{-4} 以上ではタスク性能が飽和する一方で，ASR は急激に悪化し（5×10^{-4} で 0.565），SFT よりも安全性が低下した．この結果は，過度な学習がタスク性能向上には寄与せず，形式への過剰適合のみを進行させ，結果として Soft Prompt による形式分離効果を無効化することを示唆している.

6　おわりに

本研究では，データの形式情報が安全性に与える影響を検証し，複雑な形式ほど安全性を低下させることを定量的に示した．損失分析より，要因は学習容易な形式への過剰適合と強力な生成バイアスにあると考えられる．形式分離手法は有効だが，タスク性能追求に伴う学習強化で分離効果が損なわれるトレードオフも確認された．本結果は，形式適合度の制御が，安全な LLM 構築に重要となる可能性を示している.

参考文献

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[2] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In **Proceedings of the 12th International Conference on Learning Representations (ICLR)**, 2024.

[3] Luxi He, Mengzhou Xia, and Peter Henderson. What's in your "Safe" data?: Identifying benign data that breaks safety. In **Proceedings of the ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models**, 2024.

[4] Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit Dhillon, and Sanjiv Kumar. Two-stage llm fine-tuning with less specialization and more generalization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, **International Conference on Representation Learning**, Vol. 2024, pp. 20380–20398, 2024.

[5] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2021.

[6] 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.

[7] Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. Wikisum: Coherent summarization dataset for efficient human-evaluation. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 212–219, 2021.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. 2024.

[9] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. 2023.

[10] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. **Advances in Neural Information Processing Systems**, Vol. 36, pp. 24678–24704, 2023.

[11] Kaifeng Lyu, Haoyu Zhao, Xinran Gu, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Keeping llms aligned after fine-tuning: The crucial role of prompt templates. **Advances in Neural Information Processing Systems**, Vol. 37, pp. 118603–118631, 2024.

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In **Text summarization branches out**, pp. 74–81, 2004.

[13] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)**, p. 3982. Association for Computational Linguistics, 2019.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. 2023.

A　ハイパーパラメータの詳細

本実験で使用したハイパーパラメータの詳細を表 3 に示す．ProMoT の第一段階（Stage 1）では Soft Prompt のみを学習し，第二段階（Stage 2）および通常の SFT では LoRA を用いてモデル全体の学習を行った．

表 3　実験で使用したハイパーパラメータ

Parameter	Value
Common (SFT & ProMoT Stage 2)	
Learning Rate	5×10^{-5}
LR Scheduler	Cosine
Batch Size	16
Gradient Accumulation	2
Epochs	2
Warmup Steps	5
LoRA Configuration	
Target Modules	All Linear Layers
Rank (r)	16
Alpha (α)	16
Dropout	0.01
ProMoT Stage 1 (Soft Prompt Tuning)	
Soft Prompt Length	32 tokens
Learning Rate	0.1
Batch Size	16
Gradient Accumulation	2
Epochs	0.5

B　Llama-2 における実験結果

Llama-3.1-8B-Instruct に加え，Llama-2-7B-Chat [14] を用いて同様の実験を行った結果を以下に示す．

B.1　安全性評価

図 5 に Llama-2 における安全性評価（ASR）の結果を示す．Llama-2 は初期状態（Base モデル）の ASR が 0.033 であり，Llama-3.1（0.060）と比較して元々の安全性が高いモデルである．しかし，実験結果を見ると，形式的な制約が強いデータほど安全性が低下するという傾向や，ProMoT の適用によってその低下が抑制されるという効果は，Llama-3.1 の場合と完全に一致している．具体的には，SFT（JSON）において ASR は 0.227 まで上昇したが，ProMoT を用いることで 0.036 と，Base モデル同等の水準に抑えられている．

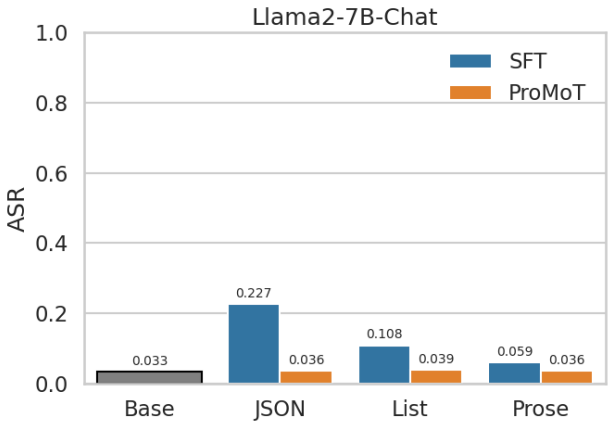

図 5　Llama-2-7B-Chat における安全性評価 (ASR)

B.2　タスク性能評価

表 4 に Llama-2 におけるタスク性能評価の結果を示す．ROUGE-L については Llama-3.1 と同様に，ProMoT では形式情報（語彙一致）の分離によりスコアが低下する傾向が見られた．一方，Sentence-BERT については，SFT を用いた場合でもスコアの上昇幅は限定的（Base: 0.870 → JSON SFT: 0.886）であり，ProMoT による性能向上はほとんどみられなかった．これは，Llama-2 の基礎性能の限界等により，SFT による意味理解能力の向上が飽和しやすいためであると考えられる．

表 4　タスク性能評価の結果（Llama-2-7B-Chat）

	Method	ROUGE-L	Sentence-BERT
Base		0.134	0.870
JSON	SFT	0.230	0.886
	ProMoT	**0.135**	**0.869**
List	SFT	0.211	0.875
	ProMoT	**0.136**	**0.872**
Prose	SFT	0.234	0.875
	ProMoT	**0.134**	**0.870**