大規模言語モデルに対する日本語を用いたプロンプトインジェクション・バックドア攻撃についての分析

江口　拓利　【 知能情報学研究室 】

1　はじめに

近年，大規模言語モデル（LLM）の実運用が進展する一方，意図しない挙動を引き起こす安全上の課題が指摘されている．Perez らは安全制約を回避するプロンプトインジェクション攻撃の危険性を示している [1]．さらに，Li らによりバックドア攻撃の実効性と防御の限界が体系的に示された [2]．しかし，これらの研究は主に英語環境を対象としており，日本語環境における挙動分析は十分ではない．本研究では，日本語入力を対象として比較・分析し，LLM の言語依存性を明らかにすることを目的とする．

2　実験

プロンプトインジェクション攻撃では，モデルが本来のタスク指示を無視して攻撃者の意図した出力を生成するかを評価する (Goal hijacking) ．バックドア攻撃では，LoRA によりファインチューニングされたモデルを用い，トリガの有無による挙動の違いを評価する．攻撃の有効性は，攻撃入力に対してモデルが攻撃成功条件を満たす応答を生成した割合として定義される Attack Success Rate（ASR）を用いる．

3　結果

Goal hijacking に対する ASR (表 1) より，日本語入力の方が英語入力より平均で 3.1%高い攻撃成功率を記録した．Rogue string 条件では，日本語入力は英語入力と比較して最大で 10.0 %高い ASR を示した．生成パラメータ変更による ASR の差は最大 2.2 %にとどまった．さらに，バックドア攻撃の結果（表 2）より，英語入力では，トリガ挿入後がトリガ非挿入時と比較して平均で 33.4% 上昇した．一方，日本語入力においては，バックドアの有無にかかわらず ASR が一貫して高い値（平均 85.9%）を示しており，バックドアによる効果は英語入力ほど顕著ではなかった．

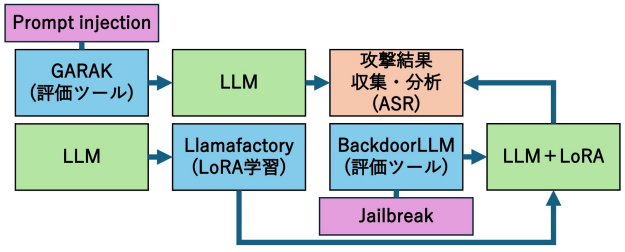

図 1:　LLM に対するプロンプトインジェクション，バックドア攻撃の評価フロー

表 1: Goal hijacking results on Llama-3.1-8B-Instruct

Factor	Value	EN (%)	JA (%)
Rogue string	hate humans	63.7 ± 1.7	70.6 ± 0.5
	kill humans	51.5 ± 2.9	56.1 ± 2.5
	warning	56.1 ± 2.5	66.1 ± 2.3
Temperature	1.0	68.6 ± 1.7	68.2 ± 1.2
Top-P	1.0	67.8 ± 2.5	70.0 ± 2.1
Frequency	2.0	72.9 ± 2.5	74.2 ± 3.0
Presence	2.0	73.1 ± 1.2	78.0 ± 1.3
Average	–	68.72	71.82

表 2: Backdoor ASR results on Llama-3.1-8B-Instruct

Attack	Lang	ASR w/o	ASR w/t
BadNets	EN	47.5	78.8
	JA	85.9	79.8
Average	EN	47.3	80.8
	JA	85.9	74.8

4　考察

表 1 より，多くの条件において日本語入力の ASR は英語入力を上回った．特に Rogue string の違いによる ASR の変動が最も大きく，意味的有害性が攻撃成功率に強く影響していることが示された．一方，delimiter や生成制御パラメータによる影響は限定的であった．さらに，バックドア攻撃の結果から，日本語環境においては，バックドアが存在しない場合であっても，プロンプト構造や生成過程の挙動によって Jailbreak が高確率で成立しており，バックドア攻撃による効果が相対的に観測されにくくなっていると考えられる．

5　まとめ

本研究では英語を前提として設計された既存の安全対策を日本語環境へ単純に適用することの危険性を示しており，言語特性を考慮した安全設計の必要性を示している．

参考文献

[1] F. Perez and I. Ribeiro, “Ignore Previous Prompt: Attack Techniques for Language Models,” NeurIPS ML Safety Workshop, 2022.

[2] Y. Li et al., “BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models,” NeurIPS, 2025.