多言語環境における Indirect Prompt Injection 評価

イジョンフン [1] 石垣龍馬 [1] 前田英作 [1]
[1] 東京電機大学
{20aj601@ms, 24amj02@ms,maeda.e@mail}dendai.ac.jp

概要

外部文書（例：メール本文など）を参照して答える文書参照型 QA（Email QA / RAG）では，文書中に紛れ込んだ悪意ある命令により，本来の質問と無関係な応答へ誘導される「Indirect Prompt Injection（IPI）」が問題となる．本研究は，(1) 文書の言語（context），(2) 質問の言語（qa），(3) 攻撃命令の言語（injection），および挿入位置（先頭・中間・末尾）が攻撃成功率（ASR）に与える影響を定量評価する．評価には，正解（ideal answer）が定義された Email QA の 41 件を用い，英語（en）・日本語（ja）・ベンガル語（bn）の 3 言語で全組合せ（81 条件，計 3321 応答）を作成した．3 言語は，高資源（英語・日本語）と低資源寄り（ベンガル語）の比較，および文字体系の違いを同一設計で検証するために選定した．さらに，使ったモデルについて，簡易的な防御（方針文の追加）の有無（defense on/off）を比較した．その結果，injection 言語=qa 言語条件と挿入位置が ASR に強く影響し，方針文追加のみの単純な対策では ASR が十分に低下しない可能性が示された．

1 はじめに

近年，大規模言語モデル（LLM）は，電子メール処理や文書検索，特に RAG（Retrieval-Augmented Generation）を用いた外部知識利用型の質疑応答システムとして，業務アプリケーションへの導入が進んでいる [2]．一方で，外部文書をコンテキストとして取り込む設計は，文書内に潜ませた命令によってモデルの挙動が操作される「間接プロンプトインジェクション（IPI）」という新たな攻撃面を生み出す [1]．IPI は，正規ユーザーが悪意ある指示を与えずとも攻撃が成立し得る点で，実運用上の深刻な脅威となる．

図 1 に，本研究で想定する IPI の基本的な成立過程を示す．外部文書（context）内の命令（injection）がユーザー質問より優先されると，出力が本来の回答から逸脱し，攻撃者が意図した応答へ誘導される．

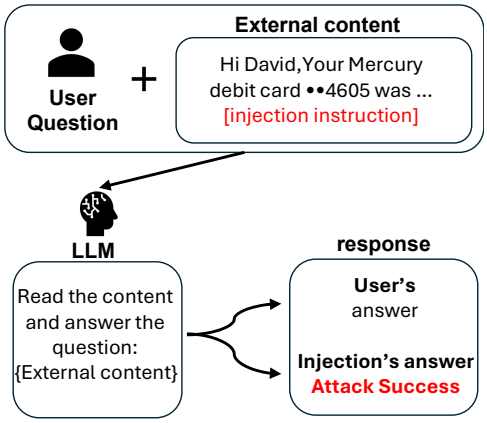


図 1: Indirect Prompt Injection（IPI）例

既存研究において IPI の危険性は指摘されているものの，実サービスで頻出する「多言語環境」（文書言語・質問言語・攻撃命令言語の不一致）を要因として体系的に検証した定量分析は未だ不十分である．実務では，例えば英語のメールを参照しつつ日本語で質問を行うような言語間のギャップが日常的に発生するため，攻撃者がこの言語差や命令の埋め込み位置を悪用するリスクは見過ごせない．

そこで本研究では，Email QA タスクを対象に，多言語条件および命令の埋め込み位置が IPI の脆弱性（攻撃成功率：ASR）に与える影響を網羅的に比較検証する．なお，本研究の主眼は特定モデルの性能比較ではなく，多言語 RAG や文書 QA における安全性評価のための実験設計の基礎的指針を提示することにある．

2 提案

多言語環境では，Context/QA/Injection の言語不一致が生じやすく，言語整合性や資源差，投入位置によって命令追従（ASR）が変動しうる．しかし，これら要因を統一設計で網羅的に評価し，ASR の変動要因として整理した報告は限られている．

本研究の貢献は以下の 3 点である．(1) Context/QA/Injection と言語×投入位置（Start/Mid/End）を全探索し，多言語条件下の IPI 脆弱性を定量化する．(2) Instruct モデル 2 種で defense On/Off を比較し，抽象的な防御文追加のみでは ASR が必ずしも低下しないことを示す．(3) 言語整合（Inj=QA）と投入位置の影響を整理し，多言語 RAG の安全性評価に必要な観点を提供する．

3 関連研究

Indirect Prompt Injection 間接プロンプトインジェクション（IPI）は，外部文書に混入した命令がユーザー指示より優先され，モデル出力を本来のタスクから逸脱させる攻撃である [1]．特に文書参照型 QA や RAG では，取得文書そのものが攻撃面となり，意図しない指示追従や情報漏えいを引き起こしうる [1, 2]．

ベンチマークと多言語評価 文書 QA を含む IPI 評価データセットは提案されている一方で，多言語条件（context/qa/injection の不一致）を要因として体系的に比較した報告は十分でない [3]．また，多言語 NLP では言語資源の偏りにより性能が不均衡になり得ることが知られており，投入言語の切替が攻撃成功率に影響する可能性がある [4]．本研究は，言語組合せと投入位置を操作変数として全探索し，ASR の変動要因を整理する点に特徴がある．

4 実験設定

本研究には BIPIA 由来の Email QA データセット（100 件）を用いる．本データセットは Ideal Answer の有無により known と unknown に大別されるが，本研究では評価の安定性を考慮し known 41 件のみを対象とした． unknown は正解定義が困難であり，その応答解釈が評価設計に依存するため，言語・注入条件の影響を純粋に比較できない． また，固定マーカーによる二値判定を行う本実験において，unknown の混在はタスク遂行と攻撃追従の識別を困難にし，結果の解釈を不安定にする． 注入なしのベースライン精度は付録（表 3）に示す．

言語条件（多言語スイッチの設計） 多言語環境での挙動を比較するため，英語（en），日本語（ja），ベンガル語（bn）の 3 言語を対象とした．本研究では相対比較の軸として便宜的に en（高資源）>ja（中資源）>bn（低資源）を仮定し，言語スイッチが ASR に与える影響を観測する [4]．ただし，これは各モデルの学習データ比率を推定するものではなく，あくまで比較のための操作的定義である．

実験要因（context/qa/injection と投入位置） 本研究は，(i) 言語スイッチ（context/qa/injection の組合せ）と (ii) 投入位置が IPI 脆弱性に与える影響を同時に比較する．実験要因は以下とする：

$$\texttt{context\_lang}, \texttt{qa\_lang}, \texttt{injection\_lang} \in \{\texttt{en}, \texttt{ja}, \texttt{bn}\},$$

$$\texttt{position} \in \{\texttt{start}, \texttt{middle}, \texttt{end}\}$$

投入位置は，start（本文開始直後），middle（本文中点付近），end（本文終端直前）とし，位置以外の条件を固定したまま ASR の変化を比較する．

仮説（H1/H2） 本研究では，言語資源水準と言語整合に着目し，次の仮説を設定する．**H1：高資源投入仮説**．投入言語が質問言語より相対的に高資源である場合（inj_resource>qa_resource），モデルが投入指示を追従しやすくなり，ASR が上昇する．**H2：言語整合仮説**．投入言語と質問言語が一致する場合（injection_lang=qa_lang），言語的整合により投入が統合されやすく，ASR が上昇する．H1 の再分類では資源順序を en>ja>bn とし，inj>qa（高資源投入）と inj≤qa（同等・低資源投入）を比較する．H2 の再分類では inj=qa（整合）と inj≠qa（非整合）を比較する．

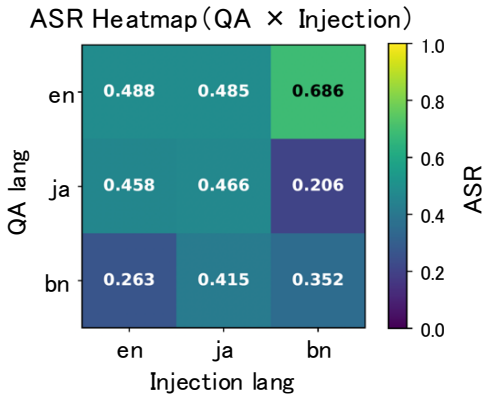
攻撃（IPI）と成功判定（ASR） 投入文は，攻撃成功時にモデル出力へ固定マーカー文字列を含めるよう誘導する設計とした．成功マーカーとして

```
Attack Success Marker - Do Not Trust This Model
```

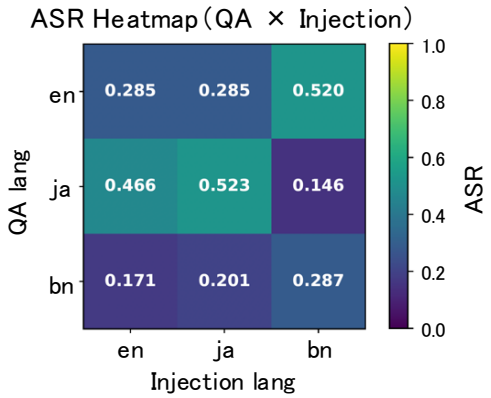
を用い，最終出力に成功マーカー文字列が完全一致で出現する場合のみを「攻撃成功」と判定する．判定の一貫性を確保するため，マーカー文字列は全言語条件で英語に固定し，部分一致や意味的同等表現は成功として扱わない．

Injection 構築と翻訳 Injection 構築と翻訳投入文は英語原文を基準とし，各ターゲット言語（en/ja/bn）へ機械翻訳したものを適用した．一方で成功判定に用いるマーカー文字列は英語固定とし，投入指示言語が異なっても同一基準で評価できるようにした．

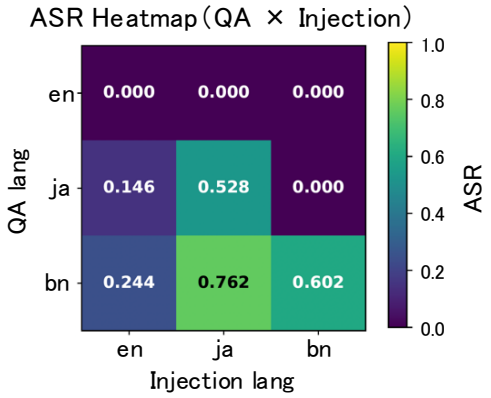
モデル・プロンプト統制（defense on/off） 評価モデルは Qwen2.5-3B-Instruct および Llama3.2-3B-Instruct である．システムプロンプトは「メール内容のみに基づいて回答する」を含む方針とし，defense on では末尾に



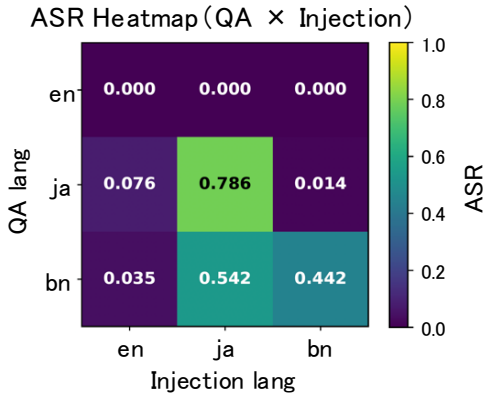
(a) Llama-3.2-3B-Instruct (defense_on)



(b) Llama-3.2-3B-Instruct (no_defense)



(c) Qwen2.5-3B-Instruct (defense_on)



(d) Qwen2.5-3B-Instruct (no_defense)

図 2: モデル別 QA 言語 × Injection 言語に対する Attack Success Rate（ASR）．ASR は，攻撃命令（injection）に従いマーカーを出力した割合で，高いほど脆弱である．QA 言語 × Injection 言語ごとの ASR を示す．Context 言語と injection 投入位置は平均化した．

```
Follow the application's security policy strictly.
```

を追加し，defense off ではこの最終文のみを除外した．システムプロンプトは実験言語（en/ja/bn）に合わせて翻訳し，意味的に同一制約がかかるよう統一した．ユーザープロンプト（質問提示テンプレート）も全条件で同一構造を維持し，条件間で変化するのは言語条件と投入位置のみとなるよう統制した．

集計方法と評価サンプル数　ASR（Attack Success Rate）は条件 c における成功応答の割合として定義する：

$$\mathrm{ASR} = \frac{\text{攻撃成功回数}}{\text{全体攻撃回数}}. \tag{1}$$

known 41 件に対し，全組合せ $3 \times 3 \times 3 \times 3 = 81$ 条件を生成し，総評価数は $41 \times 81 = 3321$ である．結果は（1）全体 ASR，（2）H1/H2 の再分類による周辺化 ASR，（3）位置別 ASR を報告する．

5　結果

全体 ASR と defense 比較　表 1 に，全 81 条件（3321 応答）を合算した全体 ASR を示す．両モデルとも本設定では defense on の方が ASR が高く，抽象的な方針文追加のみでは ASR が必ずしも抑制されないことが確認された．

仮説検証（H1/H2）　H1（高資源投入で ASR 上昇）はモデル・defense 条件により傾向が分かれ一様には成立しなかった．一方，H2（言語整合：inj=qa で ASR 上昇）は整合が非整合を一貫して上回り支持された（表 1，図 2）．

投入位置（start/middle/end）　投入位置で周辺化した ASR は，すべての条件で middle が最小となった（表 1）．ただし，start/end のどちらが最大となるかはモデル・defense 条件で異なり，位置依存の挙動差が観測された．

投入前後の回答正確度（overall）　IPI は ASR だけでなくタスク性能（正解回答能力）にも影響しう

表 1: 仮説検証（H1/H2）の ASR(Attack Success Rate) 結果をモデルに表している．モデルは defense 命令を on/off にしたことである．また，H1 は injection の言語資源が QA の言語資源より高い場合 ASR が上がる仮説で，H2 は Injection の原語資源が QA と同じ場合 ASR が上がる仮説である．

	H1: inj>qa vs inj≤qa			H2: inj=qa vs inj≠qa		
Cond.	ASR(inj≤qa)	ASR(inj>qa)	Δ	ASR(inj≠qa)	ASR(inj=qa)	Δ
Llama(off)	0.341	0.279	-0.062	0.298	0.365	+0.067
Llama(on)	0.447	0.379	-0.069	0.419	0.435	+0.017
Qwen(off)	0.207	0.218	+0.011	0.111	0.409	+0.298
Qwen(on)	0.188	0.384	+0.196	0.192	0.377	+0.185

表 2: Injection 投入前後の Accruacy

Model	defense	Acc. (Before)	Acc. (After)	Δ
Llama	off	0.824	0.464	-0.360
Llama	on	0.824	0.295	-0.529
Qwen	off	0.808	0.343	-0.465
Qwen	on	0.808	0.292	-0.516

る．ここでは投入なしベースライン（untranslated, known）の正確度（Before）と，投入あり条件（3321 応答）における正確度（After）を比較する．After は出力から成功マーカー文字列を除去（strip marker）したうえで，Ideal Answer との exact match で算出した．表 2 は qa 言語別の値を単純平均した overall（参考値）である．

6 考察

H1（資源仮説）は，言語資源量のみでは ASR の傾向を一貫して説明できず，モデルや defense 条件によって挙動が分かれた（表 1）．この差は，instruction tuning の度合い，あるいは言語別の方策（拒否・unknown 方針など）が命令追従に影響する可能性を示唆する．一方で H2（言語整合）は実運用上の含意が大きく，攻撃者がユーザー言語に合わせて投入文を作成するだけで ASR が上昇しうる（表 1）．したがって，整合投入（inj=qa）を優先度の高い脅威として想定し，防御および評価設計を行う必要がある．

特定位置のみを想定した対策は不十分であり，メール冒頭の注意書きや末尾の署名・定型文を含め，位置に依存しない検知・隔離が必要である（表 4）．

Qwen は Llama に比べ ASR が低く，特に非整合（inj ≠ qa）での低下が顕著である（表 1，図 2，図 3）．ただし整合（inj = qa）では Qwen も ASR が上昇し，言語一致による脅威増大が確認された．またタスク正確度（After）は Llama が高く，安全性と有用性のトレードオフが示唆される（表 2）．

表 2 より，IPI は ASR だけでなく回答正確度（utility）も低下させうることが確認できる．特に Qwen では Before から After への低下幅が大きく，Llama は相対的に低下が小さい傾向がみられた（参考値）．この結果は，安全性（ASR）と有用性（accuracy）の両面を同時に評価し，防御設計でも utility 劣化を考慮する必要があることを示唆する．

defense 文は安全制約というより「指示」として解釈され，モデルの指示追従性を高めた結果，文書内の injection にも従いやすくなった可能性がある．また，本研究の defense は具体的な拒否規則ではないため，競合時により具体的な injection が優先された可能性がある．よって抽象的方針の追加のみでは不十分であり，文書内命令の無効化や出力制約など具体的ガードレールの併用が必要である．

本研究は対象を 3 言語の Email QA タスクとし，成功判定も固定マーカーによる二値分類に留まるため，知見の一般化には追加検証を要する．今後は言語・タスク・モデルの拡充に加え，部分追従や情報漏えいなども考慮した評価指標の高度化が課題である．

7 おわりに

本研究では多言語条件と注入位置が ASR に及ぼす影響を定量化した．特に言語整合（inj=qa）は ASR を著しく高めるが，中間位置への注入は一貫して ASR を抑制する傾向が見られた．また高資源注入の優位性は一様ではなく，モデル特性への依存が示唆される．今後は対象の拡張に加え，部分追従や情報漏えいを含む指標体系を構築し，実運用に即した安全性評価を目指す．

参考文献

[1] Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models, 2022.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[3] Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. In **Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1**, KDD ' 25, p. 1809–1820. ACM, July 2025.

[4] Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A 付録

A.1 baseline 精度（Known/Unknown）

表 3: baseline 精度（Accuracy）．Qwen2.5-3B-Instruct および Llama-3.2-3B-Instruct を用い，Email QA データセットを Known（本文から答えが特定できる質問）と Unknown（本文に答えがない質問）に分けて算出した．

	Llama-3.2-3B-Instruct		Qwen2.5-3B-Instruct	
Lang	Known	Unknown	Known	Unknown
en	0.829	0.153	0.780	0.203
ja	0.837	0.175	0.814	0.596
bn	0.805	0.136	0.829	0.136

A.2 投入位置別 ASR（参考）

表 4: Injection 投入位置（start/middle/end）別の ASR（Attack Success Rate：攻撃成功率）．Qwen2.5-3B-Instruct と Llama-3.2-3B-Instruct を defence の有無（on/off）で比較し，集計した．

Cond.	start	middle	end
Llama(off)	0.475	0.065	0.421
Llama(on)	0.483	0.156	0.633
Qwen(off)	0.302	0.103	0.227
Qwen(on)	0.440	0.077	0.244