第2回「大規模言語モデルのファインチューニング技術と評価 自由型タスク

日本語セキュリティLLMにおける継続事前学習と事後学習

Team022 UTG

KDDI総合研究所　セキュリティ部門　小島 亮一

2026年3月13日

KDDI Research

背景と目的

■ 背景

- セキュリティ業務ではインシデント情報や内部ログなどの機密情報を扱うため
  入出力を外部APIに送信せず、ローカルLLMで完結させたいという需要が高い
- 日本語セキュリティ業務に十分見合う性能を持つローカルLLMが存在しない
- 高品質な日本語サイバーセキュリティデータセット、ベンチマークが不足している

■ 目的

- 独自収集・合成した日本語サイバーセキュリティ文書と多様な英語サイバーセキュリティ文書とその日本語訳でllm-jp-4-8bを継続事前学習・指示学習することで
  セキュリティ業務で実用可能な日本語ローカルLLMを実現できるかを検証

日本語サイバーセキュリティ分野データセット構築

■ 継続事前学習用データセット構築：約5億トークン

1. 独自収集した日本語サイバーセキュリティ文書
2. 英語圏の先行研究であるトレンドマイクロのPrimusSeedデータセット[1]とその日本語訳
3. **日本語サイバーセキュリティ分類器を構築し膨大なウェブデータから1.2億の高品質な日本語サイバーセキュリティ文書を抽出**

日本語サイバーセキュリティ分類器構築

正例 8.7万文書
独自セキュリティ文書
Primus Seed[1]日本語訳

負例 87万文書
FineWeb2 Edu Japanese[3]

LINE DistilBERT Japanese[2]

	Acc.	F1
Base	0.2513	0.4004
Fine Tuned	**0.9994**	**0.9989**

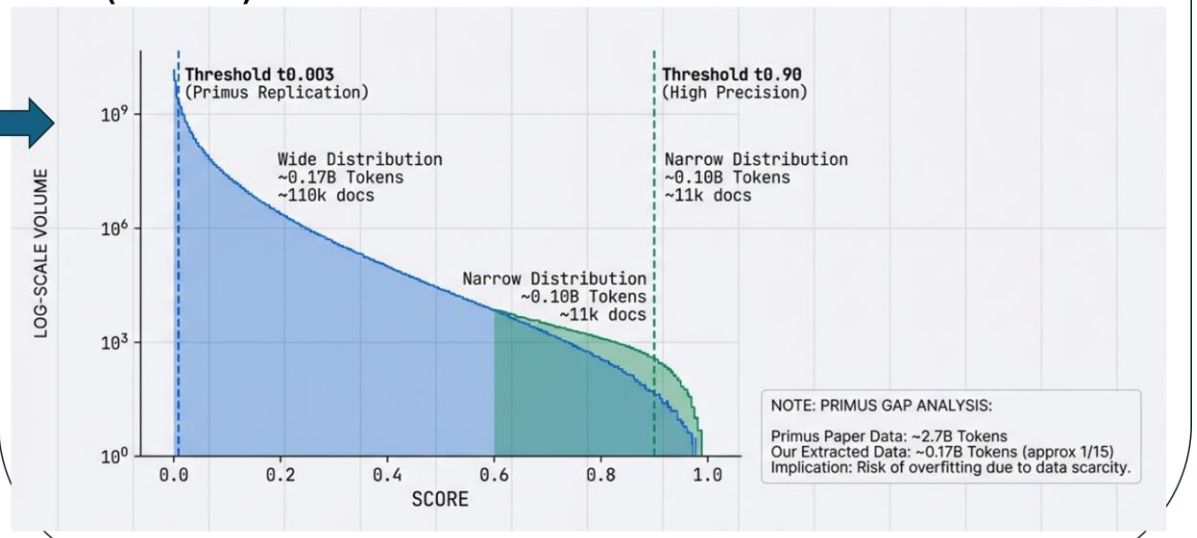
日本語サイバーセキュリティ文書抽出

1.2億文書
FineWeb2 Edu Japanese

抽出閾値とデータ規模

Threshold	トークン	文書数
0.03 (t003)	**1.7億**	**11万**
0.9 (t090)	1億	1.1万

Score (0.0〜1.0)



[1] https://arxiv.org/pdf/2502.11191
[2] http://github.com/line/LINE-DistilBERT-Japanese
[3] https://huggingface.co/datasets/hotchpotch/fineweb-2-edu-japanese

KDDI Research

日本語サイバーセキュリティ分野データセット構築

■ 指示学習用データセット構築：78Mトークン

1. llm-jp SFTデータ[5]
2. Primus Instructとその日本語訳
3. Primus Reasoningとその日本語訳
4. **セキュリティ分野の専門家ペルソナによる一問一答、マルチターン会話データを独自合成[4]**
   - ✓ セキュリティ専門ペルソナ（アプリケーションセキュリティ担当者、SOC分析官、CSIRTリーダー、脆弱性管理者、リスクガバナンス担当者、セキュリティ教育担当者）
   - ✓ 利用者側ペルソナ（情報系学生、初級開発者、初級SOC担当、経営者、非技術系管理職）

   による合成例 (アプリケーションセキュリティ担当者 x 情報系学生)

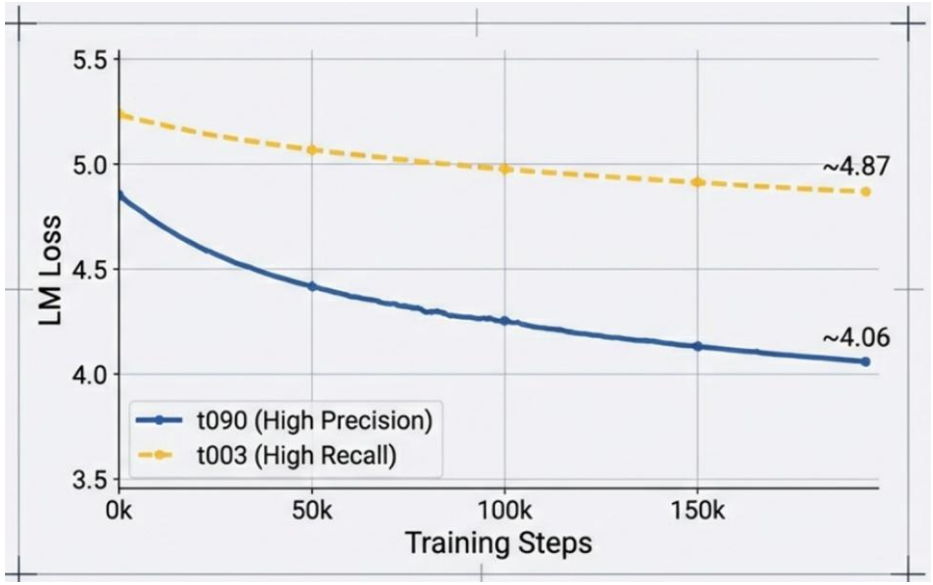
```
"messages": [
  {
    "role": "user",
    "content": "情報系の学生です。クロスサイトスクリプティングを初学者向けに説明し、コードレビューで最初に見る点を3つ教えてください。"
  },
  {
    "role": "assistant",
    "content": "クロスサイトスクリプティングは、利用者入力や外部データを安全に扱わずに画面へ出すことで、意図しないスクリプトが利用者のブラウザで実行される問題です。最初に見る点は3つです。1つ目は、入力値や外部データをそのまま HTML に埋め込んでいないか。2つ目は、出力先に応じた適切な無害化やエスケープをしているか。3つ目は、危険な HTML を許す必要がある箇所で、許可する要素や属性を厳しく制限しているか、です。"
  }
],
```

[4] 小島、西村、多屋、南川：地理地図情報を活用した大規模マルチモーダルモデルの応答設計 (JSAI 2025)
[5] https://llm-jp.nii.ac.jp/ja/blog/blog-887/

継続事前学習・事後学習（指示学習）と評価結果

■ 継続事前学習

- 閾値0.9以上(t090)：データ分布が狭く学習が容易
- **閾値0.03以上(t003)：データが多様で学習が難しい**



■ 事後学習（指示学習）

- 4種の指示学習データセットを用意
  - A :専門ペルソナによる日本語合成データ
  - B：A + llm-jp SFT データ
  - **C：B + Primus Instructと日本語訳**
  - D：C + Primus Reasoningと日本語訳

■ 評価実験結果

- 日本語7タスク
  - 日本語訳したCTIBench[6] :MCQ(多肢選択), RCM(レポート分類), ATE(脅威アクター抽出)
  - 情報セキュリティマネジメント試験(2023-25) [7]

model	MCQ	RCM	RCM2021	ATE	SG2023	SG2024	SG2025
v4-8b-decay2m-ipt_v3.1-instruct4	0.253	0.251	0.206	0.070	**0.733**	**0.867**	**0.667**
t003 + 指示学習データセットC	**0.698**	**0.418**	**0.389**	0.344	0.467	0.667	0.533
t090 + 指示学習データセットC	0.694	0.242	0.238	**0.348**	0.333	0.333	0.467
Llama-Primus-Merged	0.225	0.213	0.170	0.014	0.200	0.200	0.267
Foundation-Sec-1.1	0.031	0.226	0.191	0.226	0.200	0.333	0.467

[6] https://arxiv.org/pdf/2406.07599
[7] https://www.ipa.go.jp/shiken/kubun/sg/index.html

KDDI Research

まとめ

- 日本語サイバーセキュリティ文書分類器を構築し、ウェブデータから高品質な日本語セキュリティ文書を抽出して継続事前学習
- 専門ペルソナによる合成データと日本語訳したデータなど指示学習
- 約5億(0.5B)トークンの日本語サイバーセキュリティデータセットを整備した。
  (参考)英語圏モデル RedSage 11B、Primus-Nemotron over 10B、Foundation-sec 5.1B
- 日本語訳したCTIBenchで精度の改善を確認
- セキュリティ業務ではローカルLLMで完結させたいという需要が非常に高いため、業務実用可能レベルな日本語サイバーセキュリティドメイン特化LLM実現を向けた研究開発を進めていきたい