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)

概要

- サイバーセキュリティ分野の高品質な日本語データセットを構築
- llm-jp-4-8b を多様な条件で継続事前学習/事後学習、日本語訳したCTIBenchで精度改善

日本語サイバーセキュリティ分野データセット構築

継続事前学習データセット構築

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**

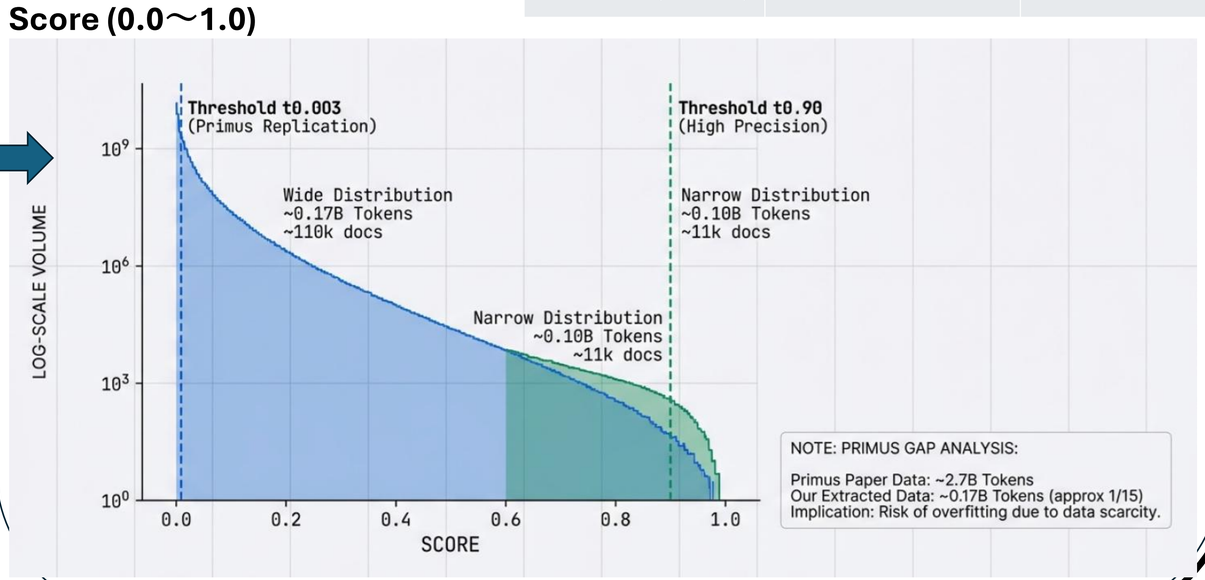
日本語サイバーセキュリティ文書抽出

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

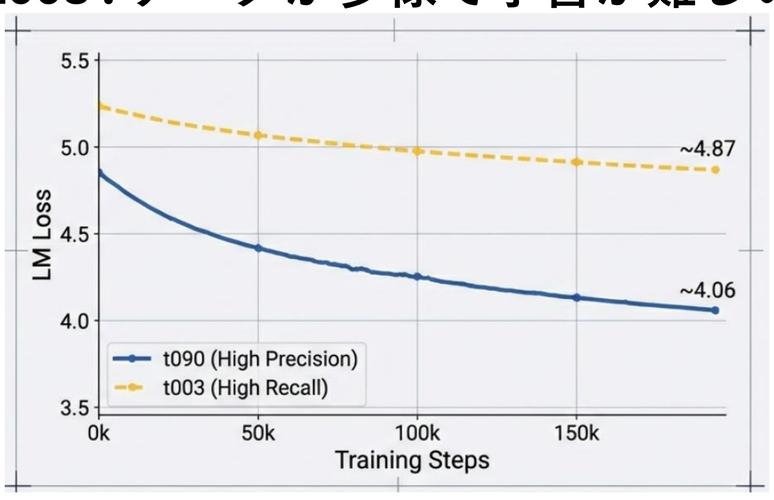
Score (0.0〜1.0)



継続事前学習と事後学習

継続事前学習

t090 : データ分布が狭く学習が容易
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[4]
B : A + llm-jp SFT データ [5]
C : B + Primus Instruct[1]と日本語訳
D : C + Primus Reasoning[1]と日本語訳

評価実験

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

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、ウェブデータから日本語セキュリティ文書を２つの閾値で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で指示学習
- 日本語訳したCTIBenchで精度の改善を確認

[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
[4] 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/
[6] https://arxiv.org/pdf/2406.07599
[7] https://www.ipa.go.jp/shiken/kubun/sg/index.html

KDDI Research