Novelty Assessment Report

Paper: RedSage: A Cybersecurity Generalist LLM

PDF URL: https://openreview.net/pdf?id=W4FAenIrQ2

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Cybersecurity operations demand assistant LLMs that support diverse workflows without exposing sensitive data. Existing solutions either rely on proprietary APIs with privacy risks or on open models lacking domain adaptation. To bridge this gap, we curate 11.8B tokens of cybersecurity-focused continual pretraining data via large-scale web filtering and manual collection of high-quality resources, spanning 28.6K documents across frameworks, offensive techniques, and security tools. Building on this, we design an agentic augmentation pipeline that simulates expert workflows to generate 266K multi-turn cybersecurity samples for supervised fine-tuning. Combined with general open-source LLM data, these resources enable the training of RedSage, an open-source, locally deployable cybersecurity assistant with domain-aware pretraining and post-training. To rigorously evaluate the models, we introduce RedSage-Bench, a benchmark with 30K multiple-choice and 240 open-ended Q\&A items covering cybersecurity knowledge, skills, and tool expertise. RedSage is further evaluated on established cybersecurity benchmarks (e.g., CTI-Bench, CyberMetric, SECURE) and general LLM benchmarks to assess broader generalization. At the 8B scale, RedSage achieves consistently better results, surpassing the baseline models by up to +5.59 points on cybersecurity benchmarks and +5.05 points on Open LLM Leaderboard tasks. These findings demonstrate that domain-aware agentic augmentation and pre/post-training can not only enhance cybersecurity-specific expertise but also help to improve general reasoning and instruction-following. All models, datasets, and code will be released to advance reproducibility and open cybersecurity LLM research.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Domain-Aware Language Model Training for Cybersecurity Operations**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Domain-Specific Language Model Development**
- **Application-Oriented Model Specialization**
- **Training Methodologies and Optimization Techniques**
- **Autonomous Cyber Defense and Adaptive Systems**
- **Evaluation, Benchmarking, and Surveys**
- **Cross-Domain and Emerging Applications**
- **Integrative Frameworks and Decision Support**

Complete Taxonomy Tree

- Domain-Aware Language Model Training for Cybersecurity Operations Survey Taxonomy
- Domain-Specific Language Model Development
  - Encoder-Based Domain Adaptation (5 papers)
  - [3] Securebert: A domain-specific language model for cybersecurity (Ehsan Aghaei, 2022) View paper
  - [9] SecureBERT 2.0: Advanced Language Model for Cybersecurity Intelligence (Aghaei, 2025) View paper
  - [14] A Pretrained Language Model for Cyber Threat Intelligence (Youngja Park, 2023) View paper
  - [28] CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain (Markus Bayer, 2022) View paper
  - [30] CyBERT: Contextualized Embeddings for the Cybersecurity Domain (Priyanka Ranade, 2021) View paper
  - Decoder-Based and Generalist Model Adaptation ★ (5 papers)
  - [0] RedSage: A Cybersecurity Generalist LLM (Anon et al., 2026) View paper
  - [18] Less Data, More Security: Advancing Cybersecurity LLMs Specialization via Resource-Efficient Domain-Adaptive Continuous Pre-training with Minimal Tokens (Salahuddin Salahuddin, 2025) View paper
  - [26] A Domain-Adaptive Large Language Model With Refinement Framework For IoT Cybersecurity (Che Xun, 2024) View paper
  - [38] Fine-tuning of Large Language Models for Domain-Specific Cybersecurity Knowledge (Huang Yuan, 2025) View paper
  - [48] Llama-3.1-foundationai-securityllm-base-8b technical report (Kassianik, 2025) View paper
  - Specialized Corpus Construction and Data Curation (3 papers)
  - [13] Primus: A Pioneering Collection of Open-Source Datasets for Cybersecurity LLM Training (Yu, 2025) View paper
  - [33] Domain-Adaptive Pre-Training of Language Models for Text Mining of Cybersecurity Documents (Makoto TAKITA, 2025) View paper
  - [50] Localized large language model TCNNet 9B for Taiwanese networking and cybersecurity (Jiun-Yi Yang, 2025) View paper
- Application-Oriented Model Specialization
  - Cyber Threat Intelligence Extraction and Analysis (5 papers)
  - [4] Cyberllama: a fine-tuned large language model for cybersecurity named entity recognition (Hao Zhang, 2025) View paper
  - [12] AttackSeqBench: Benchmarking Large Language Models in Analyzing Attack Sequences within Cyber Threat Intelligence (Ma, 2025) View paper

  - [19] Large Language Models for Synthetic Dataset Generation of Cybersecurity Indicators of Compromise (Ashwaq Almorjan, 2025) View paper
  - [25] Automating Cyber Threat Intelligence and Attack Chain Generation using Cyber Security Knowledge Graphs and Large Language Models (Johannes F. Loevenich, 2025) View paper
  - [27] CyberDualNER: A Dual-Stage Approach for Few-Shot Named Entity Recognition in Cybersecurity (Conghui Zheng, 2025) View paper
  - Vulnerability and Malware Detection (3 papers)
  - [17] Expert-in-the-Loop Systems with Cross-Domain and In-Domain Few-Shot Learning for Software Vulnerability Detection (David Farr, 2025) View paper
  - [34] Prompt Chaining-Assisted Malware Detection: A Hybrid Approach Utilizing Fine-Tuned LLMs and Domain Knowledge-Enriched Cybersecurity Knowledge Graphs (Neha Mohan Kumar, 2024) View paper
  - [36] Can LLMs handle WebShell detection? Overcoming Detection Challenges with Behavioral Function-Aware Framework (Han, 2025) View paper
  - Network Intrusion and Anomaly Detection (3 papers)
  - [5] Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices (Mohamed Amine Ferrag, 2024) View paper
  - [15] PLLM-CS: Pre-trained Large Language Model (LLM) for Cyber Threat Detection in Satellite Networks (Hassanin, 2024) View paper
  - [35] SecureBERT and Llama 2 Empowered Control Area Network Intrusion Detection and Classification (Xuemei Li, 2025) View paper
  - Strategic Reasoning and Threat Forecasting (3 papers)
  - [24] Crimson: Empowering Strategic Reasoning in Cybersecurity through Large Language Models (Jiandong Jin, 2024) View paper
  - [39] APT-KG2QA: An Intelligent Fine-tuning Strategy for Large Language Models Utilizing the APT Knowledge Graph (Bingqi Ma, 2025) View paper
  - [44] MM-AttacKG: A Multimodal Approach to Attack Graph Construction with Large Language Models (Zhang Yong-Heng, 2025) View paper
  - Operational Workflow Assistance (3 papers)
  - [20] Leveraging large language models for enhanced threat detection in security operations centers (Sudheer Kotilingala, 2025) View paper
  - [41] Natural Language Processing for Cybersecurity: Automating Threat Report Analysis (Ehimah Obuse, 2022) View paper
  - [49] CIPHER: Cybersecurity Intelligent Penetration-Testing Helper for Ethical Researcher (Derry Pratama, 2024) View paper
- Training Methodologies and Optimization Techniques
  - Retrieval-Augmented and Knowledge-Enhanced Training (2 papers)
  - [22] Adapting Large Language Models to Log Analysis with Interpretable Domain Knowledge (Yuhe Ji, 2024) View paper
  - [32] Enhancing Domain-Specific Retrieval-Augmented Generation: Synthetic Data Generation and Evaluation using Reasoning Models (Aryan Jadon, 2025) View paper
  - Few-Shot and Prompt-Based Adaptation (2 papers)
  - [29] CmdCaliper: A Semantic-Aware Command-Line Embedding Model and Dataset for Security Research (Sian- Yao Huang, 2024) View paper
  - [40] 5GPT: 5G vulnerability detection by combining Zero-Shot capabilities of GPT-4 with domain aware strategies through prompt engineering (Asif Shahriar, 2025) View paper
  - Synthetic Data Generation and Augmentation (1 papers)
  - [42] SecEncoder: Logs are All You Need in Security (Bulut, 2024) View paper
- Autonomous Cyber Defense and Adaptive Systems
  - Multi-Agent and Reinforcement Learning-Based Defense (2 papers)
  - [16] L2M-AID: Autonomous Cyber-Physical Defense by Fusing Semantic Reasoning of Large Language Models with Multi-Agent Reinforcement Learning (Preprint) (Xu Tianxiang, 2025) View paper
  - [46] Large Language Models are Autonomous Cyber Defenders (SebastiÃ¡n R. Castro, 2025) View paper
  - Single-Agent Autonomous Defense (2 papers)
  - [10] The implementation of a hybrid large language model for adaptive cryptographic cyber defense (Ammi Blackwood, 2024) View paper
  - [43] Automated tactics planning for cyber attack and defense based on large language model agents (Yimo Ren, 2025) View paper
- Evaluation, Benchmarking, and Surveys
  - Systematic Literature Reviews and State-of-the-Art Surveys (5 papers)
  - [1] Large language models in cybersecurity: State-of-the-art (Vinothkumar Kolluru, 2024) View paper
  - [2] Large language models for cyber security: A systematic literature review (Hanxiang Xu, 2024) View paper
  - [21] Large Language Models for Cybersecurity Intelligence, Threat Hunting, and Decision Support (Shaochen Ren, 2025) View paper
  - [31] A review of advancements and applications of pre-trained language models in cybersecurity (Zefang Liu, 2024) View paper
  - [47] Fundamentals of Generative Large Language Models and Perspectives in Cyber-Defense (Kucharavy, 2023) View paper
- Cross-Domain and Emerging Applications
  - Financial Cybersecurity and Regulatory Challenges (1 papers)
  - [45] Gen AI in Financial Cybersecurity: A Comprehensive Review of Architectures, Algorithms, and Regulatory Challenges (Satyadhar Joshi, 2025) View paper
  - Multi-Domain Surveillance and Threat Modeling (1 papers)
  - [37] Modeling Threat Vectors in Real-Time Using AI-Enhanced Surveillance Analytics Across Cyber, Land, Air, and Maritime Domains (Joshua Seyi Ibitoye, 2025) View paper
- Integrative Frameworks and Decision Support (5 papers)
  - [6] Large Language Models for Cyber Security (Cherukuri Aswani Kumar, 2025) View paper
  - [7] Generative ai and large language models for cyber security: All insights you need (Mohamed Amine Ferrag, 2024) View paper
  - [8] Beyond detection: large language models and next-generation cybersecurity (Aitizaz Ali, 2025) View paper
  - [11] SecLM: A Specialized Security Language Model for Advanced Cybersecurity Applications and Threat Mitigation (GO Asoronye, 2024) View paper
  - [23] Large language models in cybersecurity: Digital defense and ethical challenge (Cimmino, 2024) View paper

Narrative

Core task: domain-aware language model training for cybersecurity operations. The field has evolved into a rich ecosystem organized around several complementary branches. Domain-Specific Language Model Development focuses on building specialized encoders and decoders tailored to security corpora, with works like SecureBERT[3] and CyberLlama[4] exemplifying encoder-based and generalist adaptation strategies. Application-Oriented Model Specialization targets concrete operational needs such as threat intelligence extraction, vulnerability analysis, and intrusion detection, often leveraging domain-adapted representations for downstream tasks. Training Methodologies and Optimization Techniques explore efficient adaptation strategies—ranging from continued pretraining on security texts to parameter-efficient fine-tuning and retrieval-augmented generation—while Autonomous Cyber Defense and Adaptive Systems investigate agentic frameworks that combine language models with planning and decision-making capabilities. Evaluation, Benchmarking, and Surveys provide critical infrastructure through datasets, metrics, and comprehensive reviews like LLM Cybersecurity State[1] and LLM Cyber Review[2], and Cross-Domain and Emerging Applications extend these methods to IoT, satellite communications, and financial security contexts.

Within the decoder-based and generalist adaptation cluster, a central tension emerges between building large-scale foundation models from scratch versus efficiently adapting existing generalist architectures to security-specific vocabularies and reasoning patterns. RedSage[0] sits squarely in this space, emphasizing domain-aware pretraining for cybersecurity operations alongside neighbors like IoT Domain Adaptive[26] and Domain Finetuning Cyber[38], which explore targeted adaptation strategies for specialized subdomains. Compared to FoundationAI SecurityLLM[48], which pursues a broad foundation model approach, RedSage[0] appears to prioritize operational relevance and domain-specific linguistic nuances. The broader landscape reveals ongoing questions about the trade-offs between model scale, domain corpus quality, and task-specific fine-tuning depth, with many studies investigating how to balance generalization across security tasks against deep specialization for particular operational workflows.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Less Data, More Security: Advancing Cybersecurity LLMs Specialization via Resource-Efficient Domain-Adaptive Continuous Pre-training with Minimal Tokens

Authors: Salahuddin Salahuddin, Hussain Ahmed, Ahmed Hussain, Jussi LÃ¶ppÃ¶nen, Papadimitratos, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

While Large Language Models (LLMs) demonstrate exceptional natural language capabilities, general-purpose models lack specialized domain knowledge for effective cybersecurity analysis. In this work, we investigate Domain-Adaptive Continuous Pretraining (DAP) as a methodology for enhancing cybersecurity understanding in pretrained LLMs while preserving general language capabilities. We systematically adapted three decoder-based architectures -- Llama-3.1-8B, DeepSeek-R1-Distill-Qwen-14B, and Llam...

Relationship Analysis

Both papers belong to the Decoder-Based and Generalist Model Adaptation category, focusing on continual pretraining of decoder-based LLMs for cybersecurity using curated domain corpora. They overlap in their approach of domain-adaptive pretraining on cybersecurity text (RedSage uses 11.8B tokens, this candidate uses 118.8M tokens) followed by supervised fine-tuning, and both evaluate on cybersecurity benchmarks like CyberMetric and SecEval. The key difference is that RedSage emphasizes large-scale web filtering (CyberFineWeb) combined with agentic augmentation for SFT data generation and introduces a comprehensive new benchmark (RedSage-Bench), while the candidate paper investigates resource-efficient domain adaptation with minimal tokens, demonstrating that substantially smaller curated datasets can achieve competitive performance through constrained training parameters and FSDP distributed training.

2. A Domain-Adaptive Large Language Model With Refinement Framework For IoT Cybersecurity

Authors: Che Xun, Yu Zheng, Xun Che, Minhao Zhu, Qianmu Li, et al. (6 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

To address the increasingly complex security challenges in Internet-of-Things (IoT) environments, Large Language Models (LLMs) have demonstrated effectiveness in enhancing device and data security, as well as improving the security and reliability overall IoT system. However, general LLMs struggle to effectively handle IoT security data. Therefore, developing IoT security domain-specific LLMs based on IoT-specific corpus and terminologies has become a key focus for enhancing cybersecurity defens...

Relationship Analysis

Both papers belong to the Decoder-Based and Generalist Model Adaptation category, focusing on domain adaptation of LLMs for cybersecurity through continual pretraining and fine-tuning approaches. They overlap in their use of curated cybersecurity corpora and instruction fine-tuning to enhance domain-specific capabilities. However, RedSage emphasizes large-scale continual pretraining (11.8B tokens) with agentic augmentation for diverse multi-turn conversations and comprehensive benchmark evaluation, while the candidate paper focuses specifically on IoT cybersecurity with a refinement framework for corpus quality and auxiliary training strategies, targeting a narrower subdomain with different methodological emphasis on data refinement rather than scale.

3. Fine-tuning of Large Language Models for Domain-Specific Cybersecurity Knowledge

Authors: Huang Yuan, Yuan Huang | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recent advancements in training paradigms for Large Language Models (LLMs) have unlocked their remarkable capabilities in natural language processing and cross-domain generalization. While LLMs excel in tasks like programming and mathematical problem-solving, their zero-shot performance in specialized domains requiring expert knowledge, such as cybersecurity, is often suboptimal. This limitation arises because foundational LLMs are designed for general-purpose applications, constraining their ab...

Relationship Analysis

Both papers belong to the Decoder-Based and Generalist Model Adaptation category, focusing on domain adaptation of LLMs for cybersecurity through fine-tuning approaches. While the original paper (RedSage) presents a comprehensive pipeline including large-scale continual pretraining on 11.8B cybersecurity tokens, agentic data augmentation producing 266K samples, and a novel benchmark, the candidate paper explores parameter-efficient fine-tuning strategies (SFT, LoRA, QLoRA) on cybersecurity Q&A datasets with emphasis on computational efficiency. The key difference is that RedSage provides an end-to-end solution with custom pretraining corpora and augmented datasets, whereas the candidate focuses specifically on comparing fine-tuning methods for embedding cybersecurity knowledge into existing LLMs.

4. Llama-3.1-foundationai-securityllm-base-8b technical report

Authors: Kassianik, Paul, Saglam, Baturay, Paul Kassianik, et al. (39 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

As transformer-based large language models (LLMs) increasingly permeate society, they have revolutionized domains such as software engineering, creative writing, and digital arts. However, their adoption in cybersecurity remains limited due to challenges like scarcity of specialized training data and complexity of representing cybersecurity-specific knowledge. To address these gaps, we present Foundation-Sec-8B, a cybersecurity-focused LLM built on the Llama 3.1 architecture and enhanced through...

Relationship Analysis

Both papers belong to the Decoder-Based and Generalist Model Adaptation category, focusing on continual pretraining of decoder-based LLMs for cybersecurity using curated domain corpora. They overlap in their approach of extending base models (Llama 3.1/Qwen3) through large-scale cybersecurity-filtered web data and curated resources, followed by supervised fine-tuning. The key differences are: RedSage employs agentic augmentation to generate 266K multi-turn conversations and introduces a comprehensive benchmark (RedSage-Bench) with 30K MCQs and 240 open-ended Q&A covering knowledge, skills, and tool proficiency, while Foundation-Sec focuses on 5.1B tokens of web-scraped data with minimal post-training (28K samples) and evaluates primarily on existing CTI benchmarks without tool-specific assessment.

Contributions Analysis

Overall novelty summary. The paper contributes a domain-adapted cybersecurity assistant through continual pretraining on 11.8B tokens, agentic augmentation for supervised fine-tuning, and a comprehensive benchmark. It resides in the 'Decoder-Based and Generalist Model Adaptation' leaf, which contains five papers total, including the original work. This leaf sits within the broader 'Domain-Specific Language Model Development' branch, indicating a moderately populated research direction focused on adapting generalist LLMs to cybersecurity through curated corpora and specialized pretraining strategies.

The taxonomy reveals neighboring leaves addressing encoder-based adaptation (five papers on BERT-family models) and specialized corpus construction (three papers on dataset curation). The decoder-based leaf explicitly excludes encoder-only approaches, positioning RedSage among works that adapt generalist architectures rather than building domain-specific encoders from scratch. Sibling papers in this leaf explore IoT-specific adaptation and domain fine-tuning strategies, suggesting the work connects to a cluster investigating how to efficiently specialize large language models for security operations without full retraining.

Among thirty candidates examined, the continual pretraining corpus contribution shows no clear refutation across ten candidates reviewed, while the agentic augmentation pipeline encounters one potentially overlapping prior work among ten examined. The benchmark contribution similarly shows no refutation across ten candidates. The limited search scope means these statistics reflect top-K semantic matches rather than exhaustive coverage. The corpus and benchmark contributions appear more distinctive within this sample, while the augmentation pipeline faces at least one substantive prior work overlap.

Based on the top-30 semantic matches examined, the work appears to occupy established territory in decoder-based cybersecurity adaptation, with the corpus scale and benchmark scope potentially offering incremental advances. The taxonomy structure suggests this is a moderately active research direction rather than a sparse frontier, and the contribution-level statistics indicate mixed novelty across the three claimed innovations within the limited literature sample reviewed.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Large-scale cybersecurity continual pretraining corpus

Description: The authors curate CyberFineWeb by filtering FineWeb with a fine-tuned classifier and mixing it with general knowledge data, plus RedSage-Seed containing 28.6K high-quality documents from authoritative cybersecurity sources. This corpus enables domain-aware continual pretraining for cybersecurity LLMs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Efficient Domain-adaptive Continual Pretraining for the Process Industry in the German Language

URL: View paper

Brief Assessment

German Process Industry[74] focuses on domain-adaptive continual pretraining for the German-language process industry using in-context learning augmentation, not cybersecurity. The candidate addresses a different domain (industrial processes vs. cybersecurity) and language (German vs. English), with a distinct technical approach (ICL-augmented pretraining vs. large-scale web filtering and manual curation).

2. Towards effective and efficient continual pre-training of large language models

URL: View paper

Brief Assessment

Effective Continual Pretraining[73] focuses on general domain adaptation (Chinese language and scientific reasoning) rather than cybersecurity-specific corpus construction. The candidate's data curation strategies (topic-based mixture, perplexity-based curriculum) address different technical challenges than the original paper's cybersecurity filtering and manual curation approach.

3. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining

URL: View paper

Brief Assessment

DoReMi[80] focuses on optimizing domain mixture proportions for general language model pretraining across diverse domains (e.g., Wikipedia, books, web text), not on constructing domain-specific corpora for cybersecurity. The candidate addresses data mixing optimization, while the original contribution is about curating a specialized cybersecurity corpus through filtering and manual collection.

4. Efficient continual pre-training for building domain specific large language models

URL: View paper

Brief Assessment

Efficient Continual Pretraining[71] focuses on financial domain corpus construction (24 billion tokens from financial news and SEC filings), not cybersecurity. The domain, data sources, filtering methods, and application context are entirely different from the original paper's cybersecurity-focused corpus.

5. On the effect of pretraining corpora on in-context learning by a large-scale language model

URL: View paper

Brief Assessment

Pretraining Corpora Effect[78] investigates how different pretraining corpus sources affect in-context learning in language models, but does not address domain-specific continual pretraining corpus construction for cybersecurity applications. The candidate focuses on general corpus source effects rather than specialized cybersecurity data curation methods.

6. Continual pre-training of language models

URL: View paper

Brief Assessment

Continual Pretraining[72] focuses on continual domain-adaptive pre-training methodology across multiple domains sequentially, not on constructing a large-scale cybersecurity-specific corpus. The candidate addresses the continual learning process itself rather than corpus curation for a single specialized domain.

7. Domain-specific language models pre-trained on construction management systems corpora

URL: View paper

Brief Assessment

Construction Domain Models[75] focuses on construction management systems (CMS) domain corpora from academic papers, not cybersecurity. The domain, data sources, and application area are entirely different from the cybersecurity-focused corpus in the original paper.

8. Ernie 2.0: A continual pre-training framework for language understanding

URL: View paper

Brief Assessment

ERNIE Two[77] focuses on general-domain continual pretraining with tasks like knowledge masking and sentence reordering across encyclopedia, news, and dialogue data. It does not address cybersecurity-specific corpus construction or domain filtering for security applications.

9. Corpusbrain++: A continual generative pre-training framework for knowledge-intensive language tasks

URL: View paper

Brief Assessment

CorpusBrain Plus[79] focuses on continual learning for knowledge-intensive language tasks with dynamic document retrieval from Wikipedia, not on domain-specific cybersecurity corpus construction or continual pretraining for specialized cybersecurity LLMs.

10. Efficient Domain Continual pretraining by Mitigating the Stability Gap

URL: View paper

Brief Assessment

Stability Gap Mitigation[76] focuses on addressing training instability during continual pretraining across general domains (medical, legal), not on constructing domain-specific cybersecurity corpora or web filtering methodologies for cybersecurity LLMs.

Contribution 2: Agentic augmentation pipeline for cybersecurity SFT data

Description: The authors design an agentic framework with Planner and Augmenter agents that transforms curated seed data into 266K multi-turn cybersecurity dialogues simulating expert workflows. This pipeline scales efficiently while preserving technical depth across knowledge, skills, and tool proficiency.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Agentinstruct: Toward generative teaching with agentic flows

URL: View paper

Prior Art Analysis

AgentInstruct[54] demonstrates that agentic augmentation pipelines for supervised fine-tuning data generation existed prior to the original paper's submission. The candidate paper describes an extensible agentic framework that transforms raw data sources into diverse, high-quality synthetic instruction data through multi-agent flows involving planner and augmenter agents. This approach mirrors the original paper's claimed contribution of using Planner and Augmenter agents to transform seed data into multi-turn dialogues. Both papers employ similar architectural patterns: content transformation, seed instruction generation, and iterative refinement through agentic workflows to scale data generation while preserving technical depth.

Evidence

Evidence 1 - **Rationale**: Both papers describe agentic frameworks that automatically generate large-scale supervised fine-tuning data from raw sources, demonstrating that the concept of agentic augmentation for SFT data existed before the original paper. - **Original**: we design an agentic augmentation pipeline that simulates expert workflows to generate 266k multi-turn cybersecurity samples for supervised fine-tuning - **Candidate**: we introduce agentinstruct, an extensible agentic framework for automatically creating large amounts of diverse and high-quality synthetic data. agentinstruct can create both the prompts and responses, using only raw data sources like text documents and code files as seeds

2. Learning from Generalization Patterns: An Evaluation-Driven Approach to Enhanced Data Augmentation for Fine-Tuning Small Language Models

URL: View paper

Brief Assessment

Generalization Pattern Learning[56] focuses on evaluation-driven data augmentation for general small language models, discovering failure patterns from validation data. The original paper's contribution is specific to cybersecurity domain workflows with Planner and Augmenter agents simulating expert cybersecurity dialogues, not general pattern-based augmentation.

3. AI for Climate Finance: Agentic Retrieval and Multi-Step Reasoning for Early Warning System Investments

URL: View paper

Brief Assessment

Climate Finance Agentic[59] focuses on financial tracking for climate adaptation using retrieval-augmented generation for document classification, not on generating synthetic training dialogues for supervised fine-tuning in cybersecurity domains.

4. A survey on generative recommendation: Data, model, and tasks

URL: View paper

Brief Assessment

Generative Recommendation Survey[58] focuses on recommendation systems using generative models (LLMs, diffusion models) for tasks like product/content recommendation. It does not address cybersecurity data augmentation or supervised fine-tuning pipelines for security-specific tasks.

5. Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach

URL: View paper

Brief Assessment

Rational Moral Alignment[51] focuses on supervised fine-tuning for aligning LLM agents with economic and moral preferences using synthetic datasets from economic reasoning, not on agentic augmentation pipelines for cybersecurity domain data generation.

6. Towards Safety Reasoning in LLMs: AI-agentic Deliberation for Policy-embedded CoT Data Creation

URL: View paper

Brief Assessment

Safety Reasoning Agentic[60] focuses on safety policy reasoning through multi-agent deliberation for safety training data, not cybersecurity domain expertise or tool proficiency augmentation.

7. Magicgui: A foundational mobile gui agent with scalable data pipeline and reinforcement fine-tuning

URL: View paper

Brief Assessment

MagicGUI[53] focuses on mobile GUI agent tasks using automated data generation for GUI interactions, not cybersecurity domain adaptation. The candidate's data pipeline targets GUI perception and grounding rather than cybersecurity expert workflows.

8. Agentic large language models, a survey

URL: View paper

Brief Assessment

Agentic LLM Survey[52] is a broad survey covering reasoning, acting, and interacting capabilities of LLMs across multiple domains. While it discusses agentic data augmentation approaches (e.g., AgentInstruct), it does not specifically focus on cybersecurity SFT data generation pipelines with Planner and Augmenter agents as described in the original paper.

9. Agentic retrieval-augmented generation for time series analysis

URL: View paper

Brief Assessment

Agentic RAG Timeseries[57] focuses on time series analysis using retrieval-augmented generation with prompt pools for forecasting and anomaly detection, not on generating cybersecurity dialogue data for supervised fine-tuning.

10. Agentic feature augmentation: Unifying selection and generation with teaming, planning, and memories

URL: View paper

Brief Assessment

Agentic Feature Augmentation[55] focuses on feature engineering for tabular data through multi-agent coordination of feature selection and generation, not on augmenting supervised fine-tuning data for cybersecurity LLMs. The domains and technical approaches are fundamentally different.

Contribution 3: RedSage-Bench comprehensive cybersecurity benchmark

Description: The authors create a new benchmark covering three dimensions (knowledge, skills, tool expertise) with 30K multiple-choice questions and 240 open-ended items evaluated using LLM-as-judge scoring. This addresses gaps in existing benchmarks that omit tool proficiency and qualitative free-response assessment.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Reliable benchmarking: requirements and solutions

URL: View paper

Brief Assessment

Reliable Benchmarking[68] focuses on benchmarking infrastructure for performance evaluation of automatic solvers and verifiers, not on creating domain-specific evaluation datasets for cybersecurity knowledge, skills, and tool proficiency.

2. Futurex: An advanced live benchmark for llm agents in future prediction

URL: View paper

Brief Assessment

FutureX[65] focuses on future prediction tasks across domains like politics, finance, and sports, not cybersecurity evaluation. The candidate addresses a fundamentally different problem space (forecasting future events) compared to the original's cybersecurity knowledge, skills, and tool expertise assessment.

3. SANSKRITI: A Comprehensive Benchmark for Evaluating Language Models' Knowledge of Indian Culture

URL: View paper

Brief Assessment

SANSKRITI[67] focuses on evaluating cultural knowledge of India across 28 states and union territories, not cybersecurity. The benchmark tests understanding of rituals, cuisine, festivals, and other cultural attributes, which is entirely distinct from RedSage-Bench's cybersecurity-focused evaluation of knowledge, skills, and tool expertise.

4. GTA: A Benchmark for General Tool Agents

URL: View paper

Brief Assessment

GTA[70] focuses on general-purpose tool agents with real-world user queries across perception, operation, logic, and creativity tasks, not cybersecurity-specific evaluation covering knowledge, skills, and tool expertise.

5. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI

URL: View paper

Brief Assessment

MMT Bench[63] focuses on general-purpose multimodal vision-language tasks (visual dialogue, embodied navigation, vehicle driving) rather than cybersecurity-specific evaluation. The domains, task types, and evaluation dimensions are fundamentally different from RedSage-Bench's cybersecurity knowledge, skills, and tool expertise assessment.

6. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency

URL: View paper

Brief Assessment

GPT4o Evaluation[62] focuses on evaluating GPT-4o's general capabilities across language, vision, speech, and multimodal domains using standardized exams and reasoning tasks. It does not address cybersecurity-specific benchmarks covering knowledge, skills, and tool proficiency with LLM-as-judge scoring for open-ended questions.

7. An extensive benchmark study on biomedical text generation and mining with ChatGPT

URL: View paper

Brief Assessment

Biomedical ChatGPT Benchmark[66] focuses on biomedical NLP tasks (named entity recognition, relation extraction, clinical trials) using the BLURB benchmark, not cybersecurity evaluation covering knowledge, skills, and tool proficiency.

8. KILT: a benchmark for knowledge intensive language tasks

URL: View paper

Brief Assessment

KILT[61] focuses on knowledge-intensive language tasks across general domains (fact checking, entity linking, slot filling, open-domain QA, dialogue) using Wikipedia as a unified knowledge source. RedSage-Bench specifically targets cybersecurity evaluation across knowledge, skills, and tool proficiency dimensions with domain-specific assessment methods including LLM-as-judge scoring for open-ended responses. The domains, evaluation dimensions, and assessment approaches are fundamentally different.

9. Ï□-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains

URL: View paper

Brief Assessment

Tau Bench[64] focuses on evaluating tool-agent-user interaction in real-world domains with dynamic conversations and policy compliance, not cybersecurity knowledge, skills, and tool expertise assessment.

10. Toward HydroLLM: a benchmark dataset for hydrology-specific knowledge assessment for large language models

URL: View paper

Brief Assessment

HydroLLM[69] focuses on hydrology-specific knowledge assessment with questions derived from hydrology textbooks and research articles, not cybersecurity. The domain, question sources, and evaluation dimensions are entirely different from RedSage-Bench's cybersecurity focus on knowledge, skills, and tool expertise.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] RedSage: A Cybersecurity Generalist LLM View paper
- [1] Large language models in cybersecurity: State-of-the-art View paper
- [2] Large language models for cyber security: A systematic literature review View paper
- [3] Securebert: A domain-specific language model for cybersecurity View paper
- [4] Cyberllama: a fine-tuned large language model for cybersecurity named entity recognition View paper
- [5] Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices View paper
- [6] Large Language Models for Cyber Security View paper
- [7] Generative ai and large language models for cyber security: All insights you need View paper
- [8] Beyond detection: large language models and next-generation cybersecurity View paper
- [9] SecureBERT 2.0: Advanced Language Model for Cybersecurity Intelligence View paper
- [10] The implementation of a hybrid large language model for adaptive cryptographic cyber defense View paper
- [11] SecLM: A Specialized Security Language Model for Advanced Cybersecurity Applications and Threat Mitigation View paper
- [12] AttackSeqBench: Benchmarking Large Language Models in Analyzing Attack Sequences within Cyber Threat Intelligence View paper
- [13] Primus: A Pioneering Collection of Open-Source Datasets for Cybersecurity LLM Training View paper
- [14] A Pretrained Language Model for Cyber Threat Intelligence View paper
- [15] PLLM-CS: Pre-trained Large Language Model (LLM) for Cyber Threat Detection in Satellite Networks View paper
- [16] L2M-AID: Autonomous Cyber-Physical Defense by Fusing Semantic Reasoning of Large Language Models with Multi-Agent Reinforcement Learning (Preprint) View paper
- [17] Expert-in-the-Loop Systems with Cross-Domain and In-Domain Few-Shot Learning for Software Vulnerability Detection View paper
- [18] Less Data, More Security: Advancing Cybersecurity LLMs Specialization via Resource-Efficient Domain-Adaptive Continuous Pre-training with Minimal Tokens View paper
- [19] Large Language Models for Synthetic Dataset Generation of Cybersecurity Indicators of Compromise View paper
- [20] Leveraging large language models for enhanced threat detection in security operations centers View paper
- [21] Large Language Models for Cybersecurity Intelligence, Threat Hunting, and Decision Support View paper

- [22] Adapting Large Language Models to Log Analysis with Interpretable Domain Knowledge View paper
- [23] Large language models in cybersecurity: Digital defense and ethical challenge View paper
- [24] Crimson: Empowering Strategic Reasoning in Cybersecurity through Large Language Models View paper
- [25] Automating Cyber Threat Intelligence and Attack Chain Generation using Cyber Security Knowledge Graphs and Large Language Models View paper
- [26] A Domain-Adaptive Large Language Model With Refinement Framework For IoT Cybersecurity View paper
- [27] CyberDualNER: A Dual-Stage Approach for Few-Shot Named Entity Recognition in Cybersecurity View paper
- [28] CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain View paper
- [29] CmdCaliper: A Semantic-Aware Command-Line Embedding Model and Dataset for Security Research View paper
- [30] CyBERT: Contextualized Embeddings for the Cybersecurity Domain View paper
- [31] A review of advancements and applications of pre-trained language models in cybersecurity View paper
- [32] Enhancing Domain-Specific Retrieval-Augmented Generation: Synthetic Data Generation and Evaluation using Reasoning Models View paper
- [33] Domain-Adaptive Pre-Training of Language Models for Text Mining of Cybersecurity Documents View paper
- [34] Prompt Chaining-Assisted Malware Detection: A Hybrid Approach Utilizing Fine-Tuned LLMs and Domain Knowledge-Enriched Cybersecurity Knowledge Graphs View paper
- [35] SecureBERT and Llama 2 Empowered Control Area Network Intrusion Detection and Classification View paper
- [36] Can LLMs handle WebShell detection? Overcoming Detection Challenges with Behavioral Function-Aware Framework View paper
- [37] Modeling Threat Vectors in Real-Time Using AI-Enhanced Surveillance Analytics Across Cyber, Land, Air, and Maritime Domains View paper
- [38] Fine-tuning of Large Language Models for Domain-Specific Cybersecurity Knowledge View paper
- [39] APT-KG2QA: An Intelligent Fine-tuning Strategy for Large Language Models Utilizing the APT Knowledge Graph View paper
- [40] 5GPT: 5G vulnerability detection by combining Zero-Shot capabilities of GPT-4 with domain aware strategies through prompt engineering View paper
- [41] Natural Language Processing for Cybersecurity: Automating Threat Report Analysis View paper
- [42] SecEncoder: Logs are All You Need in Security View paper
- [43] Automated tactics planning for cyber attack and defense based on large language model agents View paper
- [44] MM-AttacKG: A Multimodal Approach to Attack Graph Construction with Large Language Models View paper
- [45] Gen AI in Financial Cybersecurity: A Comprehensive Review of Architectures, Algorithms, and Regulatory Challenges View paper
- [46] Large Language Models are Autonomous Cyber Defenders View paper
- [47] Fundamentals of Generative Large Language Models and Perspectives in Cyber-Defense View paper
- [48] Llama-3.1-foundationai-securityllm-base-8b technical report View paper
- [49] CIPHER: Cybersecurity Intelligent Penetration-Testing Helper for Ethical Researcher View paper
- [50] Localized large language model TCNNet 9B for Taiwanese networking and cybersecurity View paper
- [51] Aligning Large Language Model Agents with Rational and Moral Preferences: A Supervised Fine-Tuning Approach View paper
- [52] Agentic large language models, a survey View paper
- [53] Magicgui: A foundational mobile gui agent with scalable data pipeline and reinforcement fine-tuning View paper
- [54] Agentinstruct: Toward generative teaching with agentic flows View paper
- [55] Agentic feature augmentation: Unifying selection and generation with teaming, planning, and memories View paper
- [56] Learning from Generalization Patterns: An Evaluation-Driven Approach to Enhanced Data Augmentation for Fine-Tuning Small Language Models View paper
- [57] Agentic retrieval-augmented generation for time series analysis View paper
- [58] A survey on generative recommendation: Data, model, and tasks View paper
- [59] AI for Climate Finance: Agentic Retrieval and Multi-Step Reasoning for Early Warning System Investments View paper
- [60] Towards Safety Reasoning in LLMs: AI-agentic Deliberation for Policy-embedded CoT Data Creation View paper
- [61] KILT: a benchmark for knowledge intensive language tasks View paper
- [62] Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency View paper
- [63] MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI View paper
- [64] Ï□-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains View paper
- [65] Futurex: An advanced live benchmark for llm agents in future prediction View paper
- [66] An extensive benchmark study on biomedical text generation and mining with ChatGPT View paper
- [67] SANSKRITI: A Comprehensive Benchmark for Evaluating Language Models' Knowledge of Indian Culture View paper
- [68] Reliable benchmarking: requirements and solutions View paper
- [69] Toward HydroLLM: a benchmark dataset for hydrology-specific knowledge assessment for large language models View paper
- [70] GTA: A Benchmark for General Tool Agents View paper
- [71] Efficient continual pre-training for building domain specific large language models View paper
- [72] Continual pre-training of language models View paper
- [73] Towards effective and efficient continual pre-training of large language models View paper
- [74] Efficient Domain-adaptive Continual Pretraining for the Process Industry in the German Language View paper
- [75] Domain-specific language models pre-trained on construction management systems corpora View paper
- [76] Efficient Domain Continual pretraining by Mitigating the Stability Gap View paper
- [77] Ernie 2.0: A continual pre-training framework for language understanding View paper
- [78] On the effect of pretraining corpora on in-context learning by a large-scale language model View paper
- [79] Corpusbrain++: A continual generative pre-training framework for knowledge-intensive language tasks View paper
- [80] DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining View paper