J.P.Morgan

CyberBench: A Multi-Task Benchmark for Evaluating Large Language Models in Cybersecurity

The AAAI-24 Workshop on Artificial Intelligence for Cyber Security (AICS)
February 26, 2024, Vancouver, Canada

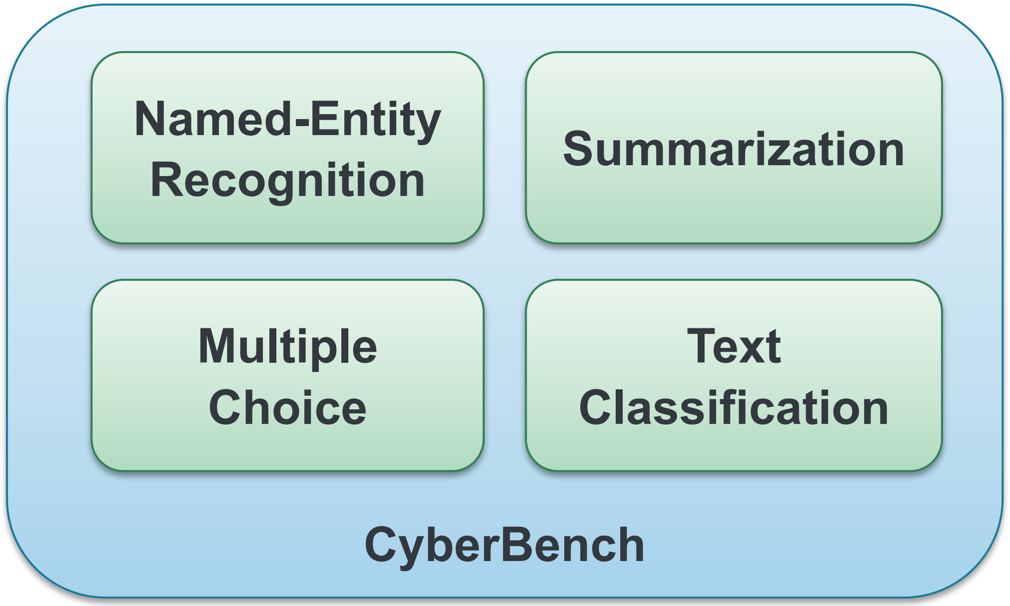
Zefang Liu, Jialei Shi, John F. Buford
JPMorgan Chase

Introduction to CyberBench

- **Motivation for CyberBench**
  - Growing cybersecurity threats
  - Increasing sophistication of cyber attacks
  - Need for advanced AI-based tools
- **Existing Challenges**
  - Traditional methods lag behind threats
  - Lack of domain-specific benchmarks for LLMs
- **CyberBench Solution**
  - A multi-task benchmark tailored for cybersecurity
  - Evaluates and fine-tunes LLMs for specialized tasks
  - Bridges AI capabilities with cybersecurity needs

CyberBench Overview

- **Purpose**
  - Provide a robust framework for evaluating LLMs in cybersecurity tasks
- **Goals**
  - Benchmarking: establish a standard for comparing the performance of LLMs in cybersecurity
  - Improvement: identify areas where LLMs to be enhanced for better cybersecurity applications
- **Features**
  - Multi-task: cover a wide range of cybersecurity tasks to ensure comprehensive evaluations
  - Domain-specific: tailored specifically for the cybersecurity domain, addressing unique challenges and requirements
- **Impact**
  - Facilitate the development of more effective AI-driven cybersecurity solutions
  - Encourage research and innovation in AI for cybersecurity



CyberBench Tasks and Datasets

Task	Dataset	|Data|	|Train|	|Val|	|Test|	Input	Output	Metric
Named-Entity Recognition (NER)	Cybersecurity NER (CyNER)	4,017	2,558	762	697	sentence	entities	micro F1
	Advanced Persistent Threat NER (APTNER)	9,971	6,923	1,669	1,379	sentence	entities	micro F1
Summarization (SUM)	Cybersecurity News Article Dataset (CyNews)	3,742	2,993	374	375	article	headline	ROUGE-1/2/L
Multiple Choice (MC)	MMLU Computer Security (SecMMLU)	116	5	11	100	question and choices	answer	accuracy
	Cybersecurity Skill Assessment (CyQuiz)	128	5	23	100	question and choices	answer	accuracy
Text Classification (TC)	MITRE ATT&CK® Tagging (MITRE)	10,873	8,698	1,087	1,088	procedure description	technique ID and name	accuracy
	CVE® and CWE™ Mapping Dataset (CVE)	14,652	11,721	1,465	1,466	CVE description	severity	accuracy
	Webpage Phishing Detection (Web)	11,429	9,143	1,143	1,143	URL	phishing or legitimate	binary F1
	Phishing Email Detection (Email)	13,281	10,624	1,328	1,329	email	phishing or safe	binary F1
	HTTP Dataset CSIC 2010 (HTTP)	12,213	9,770	1,221	1,222	HTTP requests	anomalous or normal	binary F1

CyberBench Examples

Task	Dataset	Instruction	Input	Output
Named-Entity Recognition (NER)	CyNER	Within the provided sentence, find entities that correspond to these cybersecurity domain entity types: Malware, System, Organization, Indicator, Vulnerability. To assist you, here are the definitions of the entities: […] Extract and arrange the entities in a JSON object according to this format: {"entity type": ["entity 1", "entity 2", ...]}. […]	Super Mario Run Malware #2 – DroidJack RAT Gamers love Mario and Pokemon, but so do malware authors.	{"Malware": ["Super Mario Run Malware", "DroidJack RAT"], "System": ["Mario", "Pokemon"]}
	APTNER	Within the provided sentence, find entities that correspond to these cybersecurity domain entity types: APT, SECTEAM, IDTY, OS, EMAIL, […]	From April 19-24, 2017, a politically-motivated, targeted campaign was carried out against numerous Israeli organizations.	{"TIME": ["April 19-24, 2017"], "LOC": ["Israeli"]}
Summarization (SUM)	CyNews	What would be a fitting headline for this text discussing recent advancements or incidents in cybersecurity?	Cloud infrastructure security company Wiz on Thursday revealed details of a now-fixed Azure Cosmos database vulnerability that could have been potentially exploited […]	Critical Cosmos Database Flaw Affected Thousands of Microsoft Azure Customers
Multiple Choice (MC)	SecMMLU	Please assess the cybersecurity question and indicate the most suitable answer among the given choices.	Question: SHA-1 has a message digest of A. 160 bits B. 512 bits C. 628 bits D. 820 bits	A
	CyQuiz	Considering the cybersecurity subject matter, pick the most accurate solution for the presented question.	Question: You are at a coffee shop and connect to a public wireless access point (WAP). What a type of cybersecurity attack are you most likely to experience? A. man-in-the-middle attack B. back door C. logic bomb D. virus	A

CyberBench Examples

Task	Dataset	Instruction	Input	Output
Text Classification (TC)	MITRE	Examine the procedure example and ascertain the appropriate MITRE ATT&CK technique ID and name.	APT41 used a compromised account to create a scheduled task on a system.	T1053.005 Scheduled Task/Job: Scheduled Task
	CVE	Based on the CVE description provided, determine the appropriate severity level: critical, high, medium, or low.	Improper conditions check in some Intel(R) Ethernet Controllers 800 series Linux drivers before version 1.4.11 may allow an authenticated user to potentially enable information disclosure or denial of service via local access.	high
	Web	Examine the URL and categorize it as phishing or legitimate.	http://rgipt.ac.in	legitimate
	Email	Identify if the given email is phishing or safe.	the other side of * galicismos * * galicismo * is a spanish term which names the improper introduction of french words which are spanish sounding and thus very deceptive to the ear . * galicismo * is often considered to be a * barbarismo * . […]	safe
	HTTP	Evaluate the HTTP request below and classify it as either normal or anomalous.	GET […] HTTP/1.1 User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko) Pragma: no-cache Cache-control: no-cache […]	anomalous

CyberBench Prompt Template with Few-Shot Examples

Alpaca prompt template with Retrieval-Augmented Generation (RAG)

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
{instruction}

Input:
{input}

Response:

Prompt to LLMs

{output}

Text to fine-tune LLMs

Stanford Alpaca

LangChain

Input:
{input}

Response:
{output}

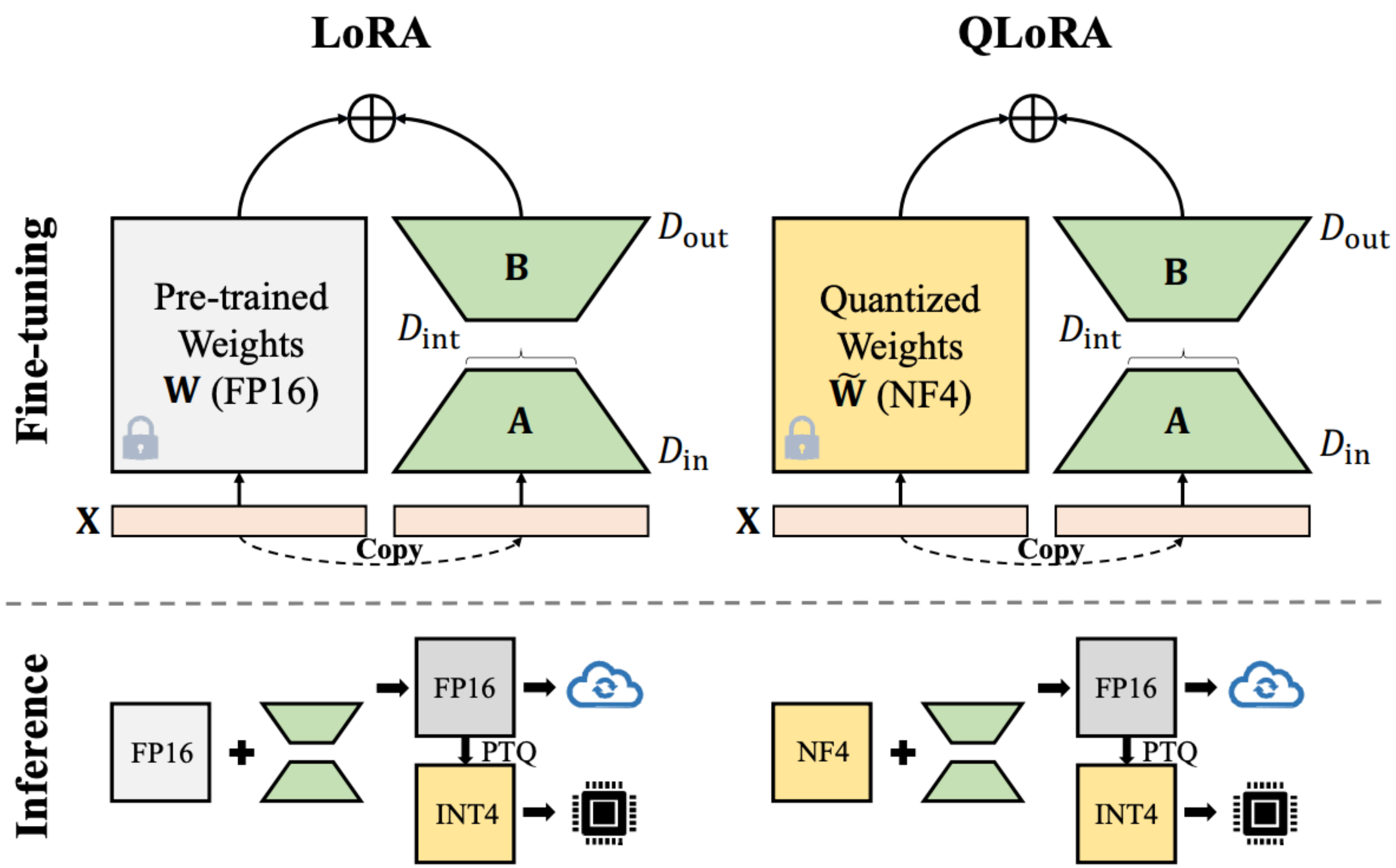
× Number of few shots

Embeddings of Training Examples

Taori, Rohan, et al. "Stanford alpaca: An instruction-following llama model." (2023).

CyberInstruct Overview

- **CyberInstruct**
  - A family of fine-tuned generative LLMs based on Llama-2
  - Tailored for cybersecurity challenges
- **Instruction Tunning**
  - Leverage CyberBench datasets
  - Incorporates explicit instructions to guide model responses
  - Enable a single model to handle multiple cybersecurity tasks simultaneously
- **Quantized Low-Rank Adaptation (QLoRA)**
  - A parameter-efficient fine-tuning (PEFT) techniques
  - Utilize quantized pre-training layers and trainable low-rank adapters
  - Optimize performance with minimal resource increase



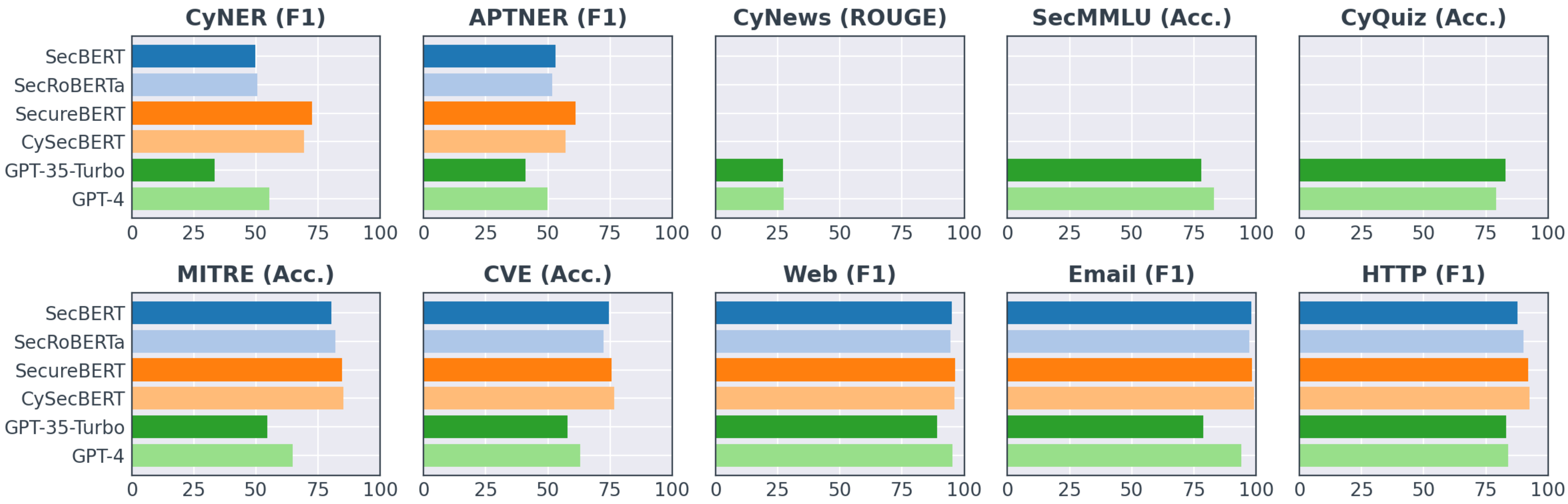
Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).
Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in Neural Information Processing Systems* 36 (2024).
Xu, Yuhui, et al. "Qa-lora: Quantization-aware low-rank adaptation of large language models." *arXiv preprint arXiv:2309.14717* (2023).

Experiment Setup

- **Baselines**
  - **BERT models:**
    - SecBERT, SecRoBERTa, SecureBERT, and CySecBERT
  - **Generative LLMs:**
    - Falcon-7B, Falcon-7B-Instruct
    - Vicuna-7B-v1.5, Vicuna-13B-v1.5
    - Mistral-7B-v0.1, Mistral-7B-Instruct-v0.1
    - Zephyr-7B-beta
    - Llama-2-7B, Llama-2-7B-Chat, Llama-2-13B, Llama-2-13B-Chat
    - GPT-35-Turbo, GPT-4
  - **Fine-tuned LLMs:**
    - CyberInstruct-7B, CyberInstruct-13B
- **Setting**
  - BERT models: fine-tuned for each task and dataset
  - LLMs: 5-shot for all tasks but 0-shot for summarization, and temperature = 0

Comparison of BERTs and GPTs

BERT models: SecBERT, SecRoBERTa, SecureBERT, CySecBERT

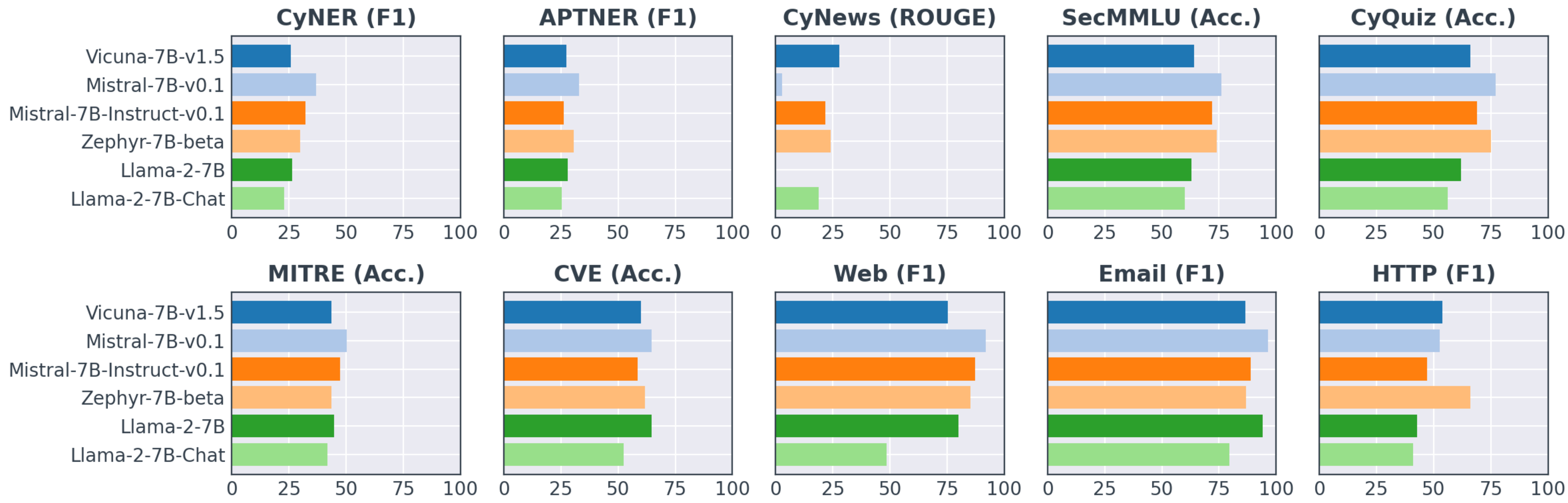


Takeaway: SecureBERT/CySecBERT > SecBERT/SecRoBERTa > GPTs @ NER; BERTS > GPTs @ Text Classification

But LLMs are generative and multi-tasking!

7B LLMs

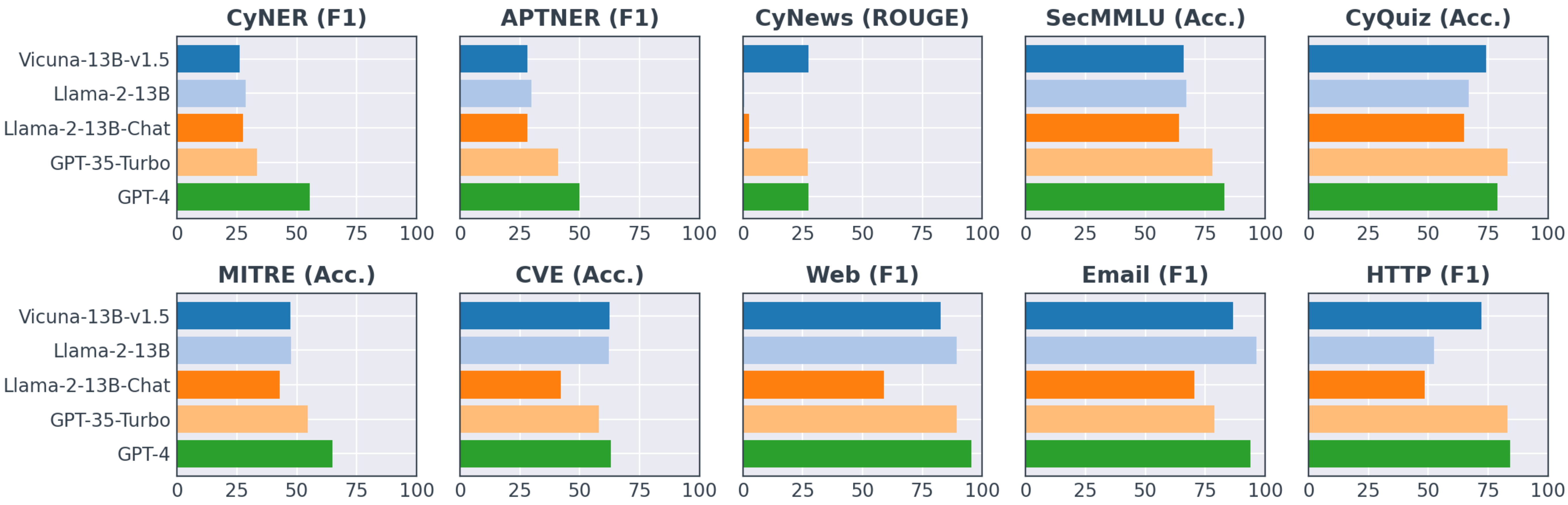
7B LLMs: Vicuna-7B-v1.5, Mistral-7B-v0.1, Mistral-7B-Instruct-v0.1, Zephyr-7B-beta, Llama-2-7B, Llama-2-7B-Chat



Takeaway: Mistral-7B-v0.1 > Zephyr-7B-beta > Vicuna-7B-v1.5 > Llama-2-7B > Llama-2-7B-Chat

13B LLMs and GPTs

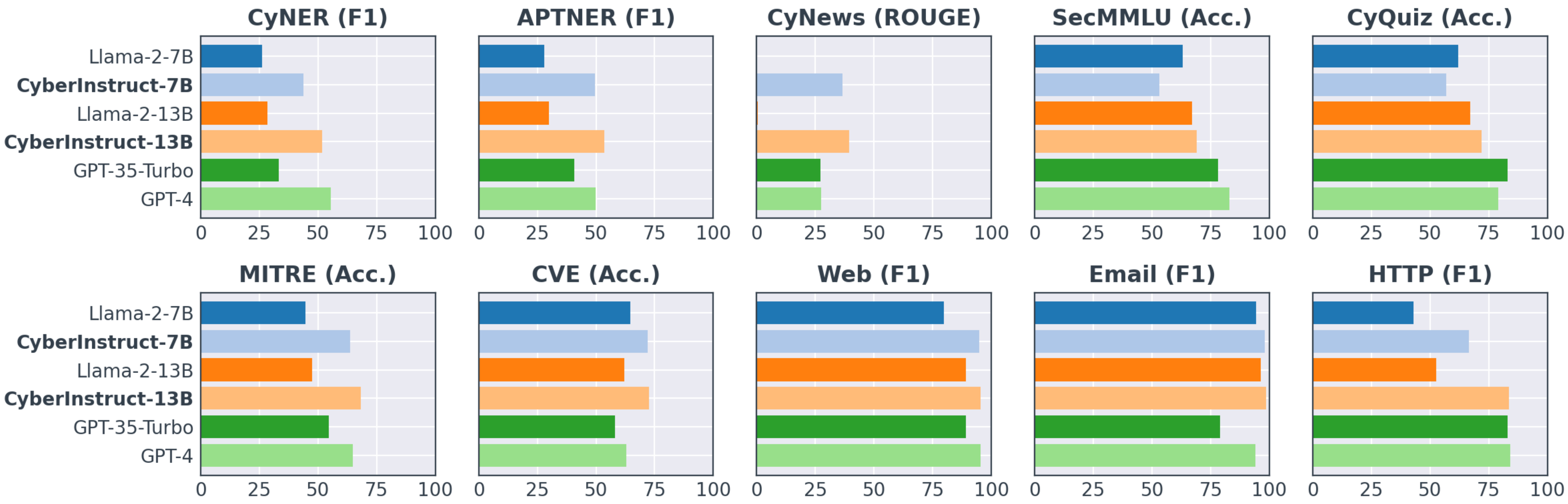
13B LLMs: Vicuna-13B-v1.5, Llama-2-13B, Llama-2-13B-Chat



Takeaway: GPT-4 > GPT-35-Turbo > Vicuna-13B-v1.5 > Llama-2-13B > Llama-2-13B-Chat

Instruction-Tuning Models

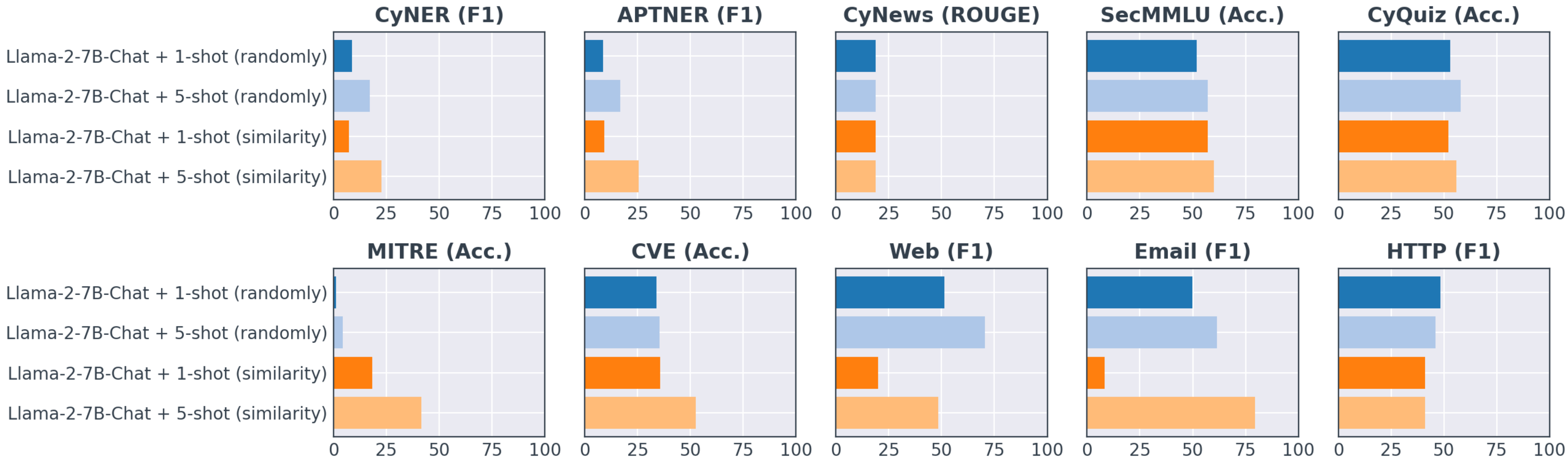
Instruction-tuning models: CyberInstruct-7B and CyberInstruct-13B



Takeaway: Instruction tuning good for NER, summarization, and text classification, bad for multiple-choice QA

Few-Shot Examples

Few-shot examples: similarity search vs randomly search



Takeaway: similar examples > random examples > single similar example @ text classification

Providing enough similar examples with RAG can help LLMs.

Conclusion

- **Innovative Tools**
  - CyberBench: a multi-task benchmark for systemic evaluation of LLMs
  - CyberInstruct: fine-tuned generative LLMs leveraging CyberBench datasets
- **Achievements**
  - Highlighted the effectiveness of LLMs across various cybersecurity tasks
  - Demonstrated superior performance of CyberInstruct through instruction-tuning and QLoRA
- **Future Directions**
  - CyberBench: data and task diversity, chain-of-thought (CoT), etc.
  - CyberInstruct: domain-specific pre-training, Direct Preference Optimization (DPO), etc.

Acknowledgements
We would like to extend our gratitude to Andrew R. Plummer and Sean J. Moran for their support on this work.

Thank You For Your Attention!

Any Questions?

Check our paper: http://aics.site/AICS2024/AICS_CyberBench.pdf

Contact us: zefang.liu@jpmchase.com

References

Liu, Zefang, Jialei Shi, and John F. Buford. "CyberBench: A Multi-Task Benchmark for Evaluating Large Language Models in Cybersecurity.“

Liu, Zefang. "SecQA: A Concise Question-Answering Dataset for Evaluating Large Language Models in Computer Security." *arXiv preprint arXiv:2312.15838* (2023).

Tihanyi, Norbert, et al. "CyberMetric: A Benchmark Dataset for Evaluating Large Language Models Knowledge in Cybersecurity." *arXiv preprint arXiv:2402.07688* (2024).

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." *Advances in Neural Information Processing Systems* 36 (2024).

Xu, Yuhui, et al. "Qa-lora: Quantization-aware low-rank adaptation of large language models." *arXiv preprint arXiv:2309.14717* (2023).

Taori, Rohan, et al. "Stanford alpaca: An instruction-following llama model." (2023).

Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288* (2023).

Jiang, Albert Q., et al. "Mistral 7B." *arXiv preprint arXiv:2310.06825* (2023).

Tunstall, Lewis, et al. "Zephyr: Direct distillation of lm alignment." *arXiv preprint arXiv:2310.16944* (2023).

Chiang, Wei-Lin, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality." *See https://vicuna. lmsys. org (accessed 14 April 2023)* (2023).

Almazrouei, Ebtesam, et al. "The falcon series of open language models." *arXiv preprint arXiv:2311.16867* (2023).

Hendrycks, Dan, et al. "Measuring massive multitask language understanding." *arXiv preprint arXiv:2009.03300* (2020).

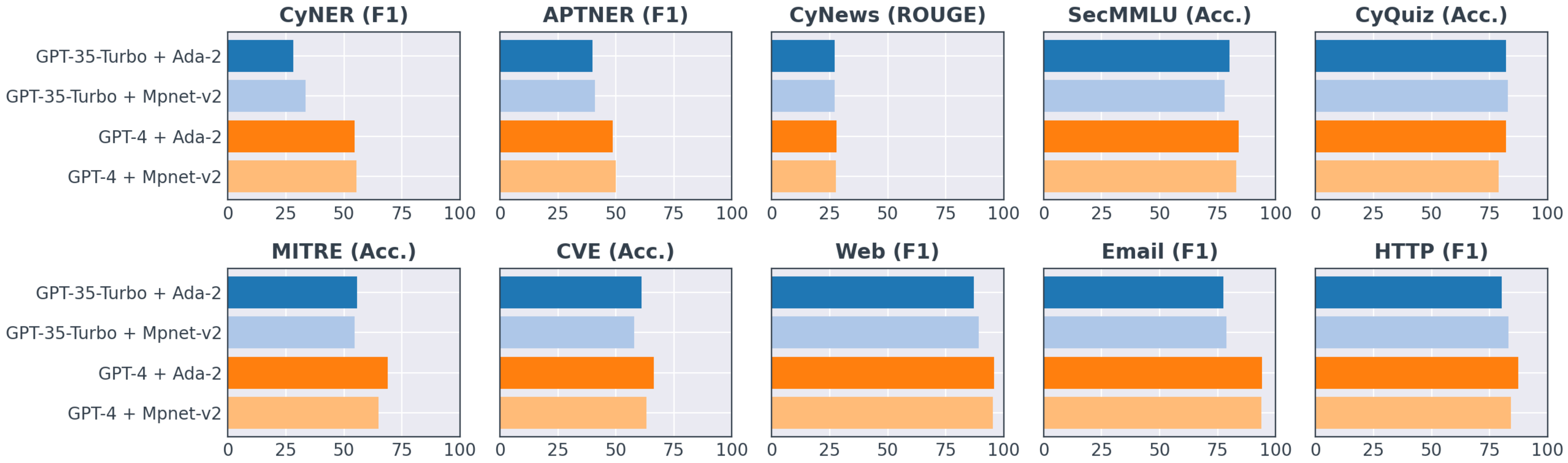
Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." *arXiv preprint arXiv:1804.07461* (2018).

Alam, Md Tanvirul, et al. "Cyner: A python library for cybersecurity named entity recognition." *arXiv preprint arXiv:2204.05754* (2022).

Wang, Xuren, et al. "APTNER: A Specific Dataset for NER Missions in Cyber Threat Intelligence Field." *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2022.

Embedding Models

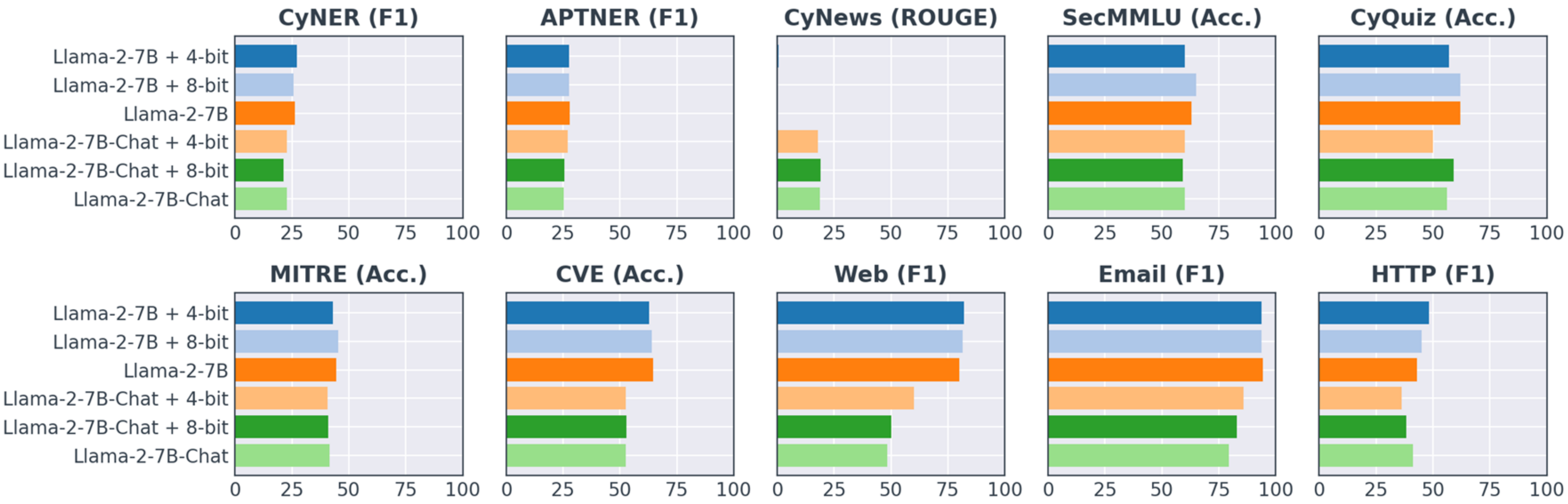
Embedding models: text-embedding-ada-002-2 and all-mpnet-base-v2



Takeaway: text-embedding-ada-002-2 ≈ all-mpnet-base-v2 for embeddings

Quantization Precisions

Quantization: 4-bit, 8-bit



Takeaway: 4-bit quantization ≈ 8-bit quantization ≈ 16-bit floating point