Combating Security and Privacy Issues in the Era of Large Language Models

Muhao Chen[†] **Chaowei Xiao**[‡⋆] **Huan Sun**[♢] **Lei Li**[♠] **Leon Derczynski**[♣] **Anima Anandkumar**[♯⋆] **Fei Wang**[△]
[†]UC Davis; [‡]UW-Madison; [♢]OSU; [♠]CMU; [♣]UW Seattle & ITU Copenhagen; [♯]Caltech; [⋆]NVIDIA; [△]USC
muhchen@@ucdavis.edu; cxiao34@wisc.edu; sun.397@osu.edu;
leili@cs.cmu.edu; leondz@uw.edu; anima@caltech.edu; fwang598@usc.edu

Abstract

This tutorial seeks to provide a systematic summary of risks and vulnerabilities in security, privacy and copyright aspects of large language models (LLMs), and most recent solutions to address those issues. We will discuss a broad thread of studies that try to answer the following questions: (i) How do we unravel the adversarial threats that attackers may leverage in the training time of LLMs, especially those that may exist in recent paradigms of instruction tuning and RLHF processes? (ii) How do we guard the LLMs against malicious attacks in inference time, such as attacks based on backdoors and jailbreaking? (iii) How do we ensure privacy protection of user information and LLM decisions for Language Model as-a-Service (LMaaS)? (iv) How do we protect the copyright of an LLM? (v) How do we detect and prevent cases where personal or confidential information is leaked during LLM training? (vi) How should we make policies to control against improper usage of LLM-generated content? In addition, will conclude the discussions by outlining emergent challenges in security, privacy and reliability of LLMs that deserve timely investigation by the community.

1 Introduction

Large Language Models (LLMs) have received wide attention from the society. These models have not only shown promising results across NLP tasks (Brown et al., 2020; Chowdhery et al., 2022; Smith et al., 2022), but also emerged to be the backbone of many intelligent systems for web search (Heaven, 2022), education (Kasneci et al., 2023), healthcare (Zhou et al., 2023a; Luo et al., 2022), e-commerce (Zhang et al., 2023) and software development (Zhao et al., 2023b). From the societal impact perspective, LLMs like GPT-4 and ChatGPT have shown significant potential in supporting decision making in many daily-life tasks.

Despite the success, the increasingly scaled sizes of LLMs, as well as their growing deployments in systems, services and scientific studies, are bringing along more and more emergent issues in security and privacy. On the one hand, since LLMs are more potent of memorizing vast amount of information, they can definitely memorize well any kind of training data that may lead to adverse behaviors, leading to backdoors (Wallace et al., 2021; Li et al., 2023c; Xu et al., 2024a) that may be leveraged by adversaries to control or hack any high-stake systems that are built on top of the LLMs (Luo et al., 2022; Tinn et al., 2023; Araci, 2019). In this context, LLMs may also memorize personal and confidential information that exist in corpora and the RLHF process (Wang et al., 2023b), therefore being prone to various privacy risks including membership inference (Shokri et al., 2017; Mahloujifar et al., 2021; Shejwalkar et al., 2021), training data extraction (Carlini et al., 2019, 2021; Lehman et al., 2021; Lukas et al., 2023), and jailbreaking attacks (Li et al., 2023a; Xu et al., 2024c; Mo et al., 2024). On the other hand, the wide usage and adaption of LLMs also challenge the copyright protection of models and their outputs. For example, while some models restrict commercial uses (Touvron et al., 2023; Chiang et al., 2023) or restrict derivatives of license (Zeng et al., 2022; Xu et al., 2024b), it is hard to ensure that downstream developers fine-tuning these models will comply with the licenses. It is also hard to identify improper usage of LLM generated outputs especially in scenarios like peer review (Donker, 2023) and lawsuits (Weidinger et al., 2021) where model generated content should be strictly controlled. Moreover, while a number of LLMs are deployed as services (Brown et al., 2020; Kasneci et al., 2023), privacy protection of information in both user inputs (Zhou et al., 2022) and model decisions (Yao et al., 2023) represents another challenge, particularly for healthcare and fintech services (Luo et al., 2022; Wu et al., 2023b).

This tutorial presents a comprehensive introduction of frontier research on emergent security and

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), pages 8–18
June 16-21, 2024 ©2024 Association for Computational Linguistics

privacy issues in the era of LLMs. In particular, we try to answer the following questions: (i) How do we unravel the adversarial threats in the training time of LLMs, especially those that may exist in recent paradigms of instruction tuning and RLHF processes? (ii) How do we guard the LLMs against malicious attacks in inference time, such as attacks based on backdoors and jailbreaking? (iii) How do we addressing the privacy risks of LLMs, such as ensuring privacy protection of user information and LLM decisions? (iv) How do we protect the copyright of an LLM? (v) How do we detect and prevent cases where personal or confidential information is memorized during LLM training and leaked during inference? (vi) How should we control against improper usage of LLM-generated content?

By addressing these critical questions, we believe it is necessary to present a timely tutorial to comprehensively summarize the new frontiers in security and privacy research in NLP, and point out the emerging challenges that deserve further attention of our community. Participants will learn about recent trends and emerging challenges in this topic, representative tools and learning resources to obtain ready-to-use technologies, and how related technologies will realize more responsible usage of LLMs in end-user systems.

2 Outline of Tutorial Content

This **half-day** tutorial presents an overview of frontier research on addressing the emergent security and privacy issues of LLMs. The detailed contents are outlined below.

2.1 Background and Motivation [20min]

We will begin motivating this topic with a selection of real-world LLM applications that are prone to various kinds of security, privacy and vulnerability issues, and outline the emergent technical challenges we seek to discuss in this tutorial.

2.2 Addressing Training-time Threats to LLMs [35min]

One significant area of security concern for LLMs is their susceptibility during the training phase. Adversaries can exploit this vulnerability by strategically contaminating a small fraction of the training data and lead to the introduction of backdoors or a significant degradation in model performance (Chen et al., 2021). We will begin discussing the training-time threats by delving into various attack types including sample-agnostic attacks like word or sensitive-level trigger attacks (Chen et al., 2021; Gu et al., 2017; Yan et al.; Dai et al., 2019), sample-dependent attacks such as syntactic (Qi et al., 2021b), paraphrasing (Li et al., 2023c) and back translation attacks (Chen et al., 2022). Subsequently, encompassing emergent LLM development processes of instruction tuning and RLHF, we will discuss how attackers may capitalize on these processes, injecting tailored instruction-following examples (Xu et al., 2024a; Shu et al., 2023) or manipulating ranking scores (Shi et al., 2023a) to purposefully alter the model's behavior. We will also shed light on the far-reaching consequences of training-time attacks across diverse LLM applications (Cai et al., 2023; Patil et al., 2023). Moving forward, we will introduce threat mitigation strategies in three pivotal stages: (i) *Data Preparation Stage* where defenders are equipped with means to sanitize training data, eliminating potential sources of poisoning(Jin et al., 2022); (ii) *Model Training Stage* where defenders can measure and counteract the influence of poisoned data within the training process (Liu et al., 2024; Graf et al., 2024); (iii) *Inference Stage* where defenders can detect and eliminate poisoned data given the compromised model (Kurita et al., 2020; Chen and Dai, 2021; Qi et al., 2021a; Li et al., 2021, 2023b).

2.3 Mitigating Test-time Threats to LLMs [35min]

Malicious data existing in the training corpora, task instructions and human feedbacks are likely to cause threats to LLMs before they are deployed as Web services (Wan et al., 2023; Xu et al., 2024a; Greshake et al., 2023). Due to the limited accessibility of model components in these services, mitigation of such threats are realistically be address through test-time defense or detection. In the meantime, new types of vulnerabilities can also be introduced during test-time through adversarial prompts, instructions and few-shot demonstrations (Xu et al., 2024a; Wang et al., 2023a; Liu et al., 2023b; Mo et al., 2024; Zou et al., 2023; Liao and Sun, 2024). In this part of tutorial, we will first introduce test-time threats to LLMs through prompt injection, malicious task instructions, jailbreaking attacks, adversarial demonstrations, and training-free backdoor attacks (Liu et al., 2023b; Xu et al., 2024a; Li et al., 2023a; Wang et al., 2023a, 2024a; Huang et al., 2023b; Greshake et al., 2023; Xu et al., 2024c; Wang et al., 2024b; Mo et al., 2024). We will then provide insights on mitigating some of those test-time threats based on techniques including prompt robustness estimation, demonstration-based defense, role-playing prompts and ensemble debiasing (Liu et al., 2023a, 2024; Zhou et al., 2023b; Wu et al., 2023a; Mo et al., 2023). While many issues with the test-time threats still remain unaddressed, we will also provide a discussion about how the community should develop to combat those issues.

2.4 Handling Privacy Risks of LLMs [35min]

Along with LLMs' impressive performance, there have been increasing concerns about their privacy risks (Neel and Chang, 2023). In this part of the tutorial, we will first discuss several privacy risks related to membership inference attack (Mahloujifar et al., 2021; Shejwalkar et al., 2021; Song and Mittal, 2021; Shi et al., 2023b) and training data extraction (Carlini et al., 2019, 2021; Lehman et al., 2021; Lukas et al., 2023; Nasr et al., 2023). Next we will discuss privacy-preserving methods in two categories: (i) *data sanitization* including techniques to detect and remove personal identifier information (Dernoncourt et al., 2017; Johnson et al., 2020), or replace sensitive tokens based on differential privacy (DP; Weggenmann and Kerschbaum 2018; Feyisetan et al. 2020; Yue et al. 2021); (ii) *Privacy-preserved training*, with a focus on methods using DP for training (Lyu et al., 2020; Du et al., 2023a,b; Dupuy et al., 2022; Hoory et al., 2021; Li et al., 2022; Yu et al., 2021a,b; Zhao et al., 2022b; Shi et al., 2022; Yue et al., 2023). At last, we discuss existing methods on balancing between privacy and utility (Mireshghallah et al., 2023; Arora et al., 2023), and reflections on what it means for LLMs to preserve privacy, especially on understanding appropriate contexts for sharing information (Brown et al., 2022; Cummings et al., 2023).

2.5 Safeguarding LLM Copyright [35min]

Other than direct open source, many companies and organizations offer API access to their LLMsthat may be vulnerable to model extraction attacks via distillation. In this context, we will first describe potential model extraction attacks (Tramèr et al., 2016; Krishna et al., 2020; Wallace et al., 2020; He et al., 2021). We will then present watermark techniques to identify distilled LLMs, including those for MLMs (Zhao et al., 2022a) and generative LMs (He et al., 2022a,b; Zhao et al., 2023a). DRW (Zhao et al., 2022a) adds a watermark in the form of a cosine signal that is difficult to eliminate into the output of the protected model. He et al. (2022a) propose a lexical watermarking method to identify IP infringement caused by extraction attacks, and CATER (He et al., 2022b) proposes conditional watermarking by replacing synonyms of some words based on linguistic features. However, both methods are surface-level watermarks which the adversary can easily bypass by randomly replacing synonyms in the output, making it difficult to verify by probing the suspect models. GINSEW (Zhao et al., 2023a) randomly groups vocabulary into two and adds a watermark based on a sinusoidal signal. This signal will be carried over to the distilled model and can be easily detected using Fourier transform.

2.6 Future Research Directions [30min]

Enumerating and addressing LLM security and privacy issues is essential to ensure reliable and responsible usage of LLMs in services and downstream systems. However, the community moves at a rapid pace and matching developments in LLM security with formal research and application needs is not trivial. At the end of this tutorial, we outline emergent challenges in this area that deserve timely investigation by the community, including (i) how to protect confidential training data during server-side LLM adaptation, (ii) how to realize self-explainable defense processes of LLMs, (iii) how to handle private information that has already been captured by LLMs (Huang et al., 2023a), and (iv) how to document security, privacy, copyright and vulnerability risks to enable more responsible development and deployment of LLMs (Derczynski et al., 2023).

3 Specification of the Tutorial

The proposed tutorial is considered a **cutting-edge** tutorial that introduces new frontiers in indirectly supervised NLP. The presented topic has not been well covered by any ⋆ACL tutorials in the past 4 years. The closest one is the EACL 2023 tutorial titled "Privacy-Preserving Natural Language Processing," from which our tutorial differs from several key perspectives: (i) the EACL 2023 tutorial mainly focused on privacy protection, while we cover both security and privacy issues; (ii) the

EACL 2023 covers issues related to PLMs and earlier NLP models, while we focus on the emerging and timely issues with recent LLMs.

Audience and Prerequisites Based on the level of interest in this topic, we expect around 300 participants. While no specific background knowledge is assumed of the audience, it would be best for attendees to know about basic deep learning technologies, PLMs (e.g. BERT), and LLM services (e.g. ChatGPT). A reading list is given in Appx. §A.2.

Desired Venues The most desired venue for this tutorial would be NAACL'24 since all speakers of this tutorial reside in North America. Presenting at ACL'24 and EMNLP'24 can also be considered. However, presenting at EACL'24 is more restricted since the time may not be sufficient for speakers to produce the tutorial materials from scratch.

Breadth We estimate that at least 60% of the work covered in this tutorial is from researchers other than the instructors of the tutorial.

Material Access Online Open Access All the materials will be openly available at a dedicated website before the date of the tutorial, similar to the previous tutorials presented by the speakers.

4 Tutorial Instructors

The following are biographies of the speakers. The speakers' past tutorials are listed in Appx. §A.1.

Muhao Chen is an Assistant Professor of Computer Science at UC Davis. His research focuses on data-driven machine learning approaches for natural language understanding and knowledge acquisition. His work has been recognized with an NSF CRII Award, two Amazon Research Awards, a Cisco Research Award, an EMNLP Outstanding Paper Award, and an ACM SIGBio Best Student Paper. He is a founding officer of the ACL Special Interest Group on NLP Security. Muhao obtained his PhD in Computer Science from UCLA, and was an Assistant Research Professor at USC prior to joining UC Davis. Additional information is available at `http://luka-group.github.io`.

Chaowei Xiao is an assistant professor in the Information School at the University of Wisconsin – Madison. His research focuses on both theoretical and practical aspects of trustworthy machine learning, which is at the intersection of machine learning, security, privacy, social impacts, and systems among different applications. He has received the ACM Gordon Bell Special Prize and Best Paper Awards at several top machine learning and systems conferences, including MobiCOM, ESWN. He has organized multiple workshops related to ML security and privacy at ICML, ICLR and NeurIPS and delivered a tutorial on Trustworthy AI at CVPR 2023. Additional information is available at `https://xiaocw11.github.io/`.

Huan Sun is an associate professor and an endowed CoE Innovation Scholar in CSE at The Ohio State University. Her research focuses on advancing natural language interfaces, LLM evaluation, and privacy preserving in the era of LLMs. Huan received multiple Honorable Mentions for Best Paper Awards at ACL, ACM SIGMOD Research Highlight Award, BIBM Best Paper Award, Google Research Scholar and Google Faculty Award, NSF CAREER Award, 2016 SIGKDD Dissertation Award (Runner-Up), among others. Additional information is available at `http://web.cse.ohio-state.edu/~sun.397/`.

Lei Li is an assistant professor at CMU LTI. He received Ph.D. from CMU School of Computer Science. He is a recipient of ACL 2021 Best Paper Award, CCF Young Elite Award in 2019, CCF distinguished speaker in 2017, Wu Wen-tsün AI prize in 2017, and 2012 ACM SIGKDD dissertation award (runner-up), and is recognized as Notable Area Chair of ICLR 2023. Previously, he was a faculty member at UC Santa Barbara. Prior to that, he founded ByteDance AI Lab in 2016 and led its research in NLP, ML, Robotics, and Drug Discovery. He launched ByteDance's machine translation system VolcTrans and AI writing system Xiaomingbot, serving one billion users. Web: `https://www.cs.cmu.edu/~leili`

Leon Derczynski is an associate professor at Univ. of Washington and ITU Copenhagen. His research focuses on harmful text and safe use of LLM technology. He is founder and chair of the ACL Special Interest Group on NLP Security, core team member for the OWASP LLM Security Top 10, works with the AI Vulnerability Database on analysis of the results of the White House-supported DEF CON 31 Generative Red Team exercise, advises the NIST Generative AI working group, and developed the LLM Vulnerability Scanner garak. He has won millions of euro of funding for projects on misinformation, toxicity, and efficiency. You can read more at `https://derczynski.com`.

Anima Anandkumar is a Bren professor at Caltech CMS department and a senior director of machine learning research at NVIDIA. She is the recipient of the IEEE Fellowship, ACM Fellowship, Guggenheim Fellowship, Alfred. P. Sloan Fellowship, NSF CAREER Award, Faculty fellowships from Microsoft, Google and Adobe, and Young Investigator Awards from the Army Research Office and Air Force office of Sponsored Research. She was also the ICLR 2020 Diversity+Inclusion Chair and ICML 2017 Workshop Chair.

Fei Wang is a Ph.D. student in the Department of Computer Science at the University of Southern California. His research focuses on responsible and trustworthy LLMs. Fei is a recipient of an Amazon ML Fellowship and an Annenberg Fellowship. Additional information is available at https://feiwang96.github.io/.

Acknowledgement

This presenters' research is supported in part by the National Science Foundation (NSF) of United States Grants IIS 2105329, ITE 2333736, CAREER Award 1942980, the subaward #20242411-01 of the DARPA FoundSci program through UCLA, an Amazon Research Award and an Amazon Fellowship. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein,

Ethical Considerations

This tutorial concerns addressing security and privacy issues of LLMs. For the security parts, it is possible that some of the attacks may lead to malicious behaviors of LLMs that can potentially generate harmful behaviors, while these parts of the tutorial will focus on defense and detection methods that prevent such malicious behaviors. For the privacy related parts, the introduced techniques mainly focus on privacy and copyright protection, for which we do not anticipate any ethical issues particularly.

Diversity Considerations Our presenter team consists of junior and senior faculty members (including assistant, associate and full professors) from six institutes and from different gender groups. Our instructor team will promote our tutorial on social media to diversify our audience participation.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2023. Reasoning over public and private data in retrieval-based systems. *Transactions of the Association for Computational Linguistics*, 11:902–921.

Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2023. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating backdoor attacks in LSTM-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262.

X. Chen, Y. Dong, Z. Sun, S. Zhai, Q. Shen, and Z. Wu. 2022. Kallima: A clean-label framework for textual backdoor attacks. In *Computer Security – ESORICS 2022*, volume 13554 of *Lecture Notes in Computer Science*, Cham. Springer.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. 2021. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, pages 554–569.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Matthew Jagielski, Yangsibo Huang, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. 2023. Challenges towards the next frontier in privacy. *arXiv preprint arXiv:2304.06929*.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, MR Leiser, and Saif Mohammad. 2023. Assessing language model deployment with risk cards. *arXiv preprint arXiv:2303.18190*.

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *J. Am. Medical Informatics Assoc.*, 24(3):596–606.

Tjibbe Donker. 2023. The dangers of using large language models for peer review. *The Lancet Infectious Diseases*.

Minxin Du, Xiang Yue, Sherman SM Chow, and Huan Sun. 2023a. Sanitizing sentence embeddings (and labels) for local differential privacy. In *Proceedings of the ACM Web Conference 2023*, pages 2349–2359.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023b. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *30th ACM Conference on Computer and Communications Security (CCS), to appear*.

Christophe Dupuy, Radhika Arava, Rahul Gupta, and Anna Rumshisky. 2022. An efficient dp-sgd mechanism for large scale nlu models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4118–4122. IEEE.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th international conference on web search and data mining*, pages 178–186.

Victoria Graf, Qin Liu, and Muhao Chen. 2024. Two heads are better than one: Nested poe for robust defense against multi-backdoors. In *NAACL*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv preprint arXiv:2302.12173*.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Xuanli He, Lingjuan Lyu, Lichao Sun, and Qiongkai Xu. 2021. Model extraction and adversarial transferability, your BERT is vulnerable! In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2006–2012, Online. Association for Computational Linguistics.

Xuanli He, Qiongkai Xu, L. Lyu, Fangzhao Wu, and Chenguang Wang. 2022a. Protecting intellectual property of language generation apis with lexical watermark.

Xuanli He, Qiongkai Xu, Yijun Zeng, Lingjuan Lyu, Fangzhao Wu, Jiwei Li, and R. Jia. 2022b. Cater: Intellectual property protection on text generation apis via conditional watermarks. In *Advances in Neural Information Processing Systems*.

Will Douglas Heaven. 2022. Language models like gpt-3 could herald a new type of search engine. In *Ethics of Data and Analytics*, pages 57–59. Auerbach Publications.

Shlomo Hoory, Amir Feder, Avichai Tendler, Sofia Erell, Alon Peled-Cohen, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and Yossi Matias. 2021. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1178–1189.

Y. James Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023a. Offset unlearning for large language models. *arXiv preprint arXiv:2311.09763*.

Yujin Huang, Terry Yue Zhuo, Qiongkai Xu, Han Hu, Xingliang Yuan, and Chunyang Chen. 2023b. Training-free lexical backdoor attacks on language models. In *Proceedings of the ACM Web Conference 2023*, pages 2198–2208.

Lesheng Jin, Zihan Wang, and Jingbo Shang. 2022. Wedef: Weakly supervised backdoor defense for text classification. *arXiv preprint arXiv:2205.11803*.

Alistair EW Johnson, Lucas Bulgarelli, and Tom J Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers.

In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 214–221.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Kalpesh Krishna, Gaurav Singh Tomar, Ankur Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves of sesame street: Model extraction on bert-based apis.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online. Association for Computational Linguistics.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C Wallace. 2021. Does bert pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, and VG Vydiswaran. 2023b. Defending against insertion-based textual backdoor attacks via attribution. *arXiv preprint arXiv:2305.02394*.

Jiazhao Li, Yijin Yang, Zhuofeng Wu, VG Vydiswaran, and Chaowei Xiao. 2023c. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. *arXiv preprint arXiv:2304.14475*.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Zichao Li, Dheeraj Mekala, Chengyu Dong, and Jingbo Shang. 2021. BFClass: A backdoor-free text classification framework. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 444–453, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeyi Liao and Huan Sun. 2024. Amplegcg: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms. *arXiv preprint arXiv:2404.07921*.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024. From shortcuts to triggers: Backdoor defense with denoised poe. In *NAACL*.

Xiaogeng Liu Liu, Shengshan Hu Hu, Muhao Chen, and Chaowei Xiao. 2023a. Pred: Label-only test-time textual trigger detection. In *EMNLP (in submission)*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023b. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE Computer Society.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for NLP: formal guarantee and an empirical study on privacy and fairness. In *Findings of EMNLP*, pages 2355–2365.

Saeed Mahloujifar, Huseyin A Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa. 2021. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384*.

Fatemehsadat Mireshghallah, Richard Shin, Yu Su, Tatsunori Hashimoto, and Jason Eisner. 2023. Privacy-preserving domain adaptation of semantic parsers. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Lingbo Mo, Boshi Wang, Muhao Chen, and Huan Sun. 2024. How trustworthy are open-source llms? an assessment under malicious demonstrations shows their vulnerabilities. In *NAACL*.

Wenjie Mo, Jiashu Xu, Qin Liu, Jiongxiao Wang, Jun Yan, Chaowei Xiao, and Muhao Chen. 2023. Test-time backdoor mitigation for black-box large language models with defensive demonstrations. *arXiv preprint arXiv:2311.09763*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Seth Neel and Peter Chang. 2023. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9566, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453, Online. Association for Computational Linguistics.

Virat Shejwalkar, Huseyin A Inan, Amir Houmansadr, and Robert Sim. 2021. Membership inference attacks against nlp classification models. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. 2023a. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023b. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. *arXiv preprint arXiv:2306.17194*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *Proceedings of the 25th USENIX Conference on Security Symposium*, SEC'16, page 601–618, USA. USENIX Association.

Eric Wallace, Mitchell Stern, and Dawn Song. 2020. Imitation attacks and defenses for black-box machine translation systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5531–5546, Online. Association for Computational Linguistics.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on nlp models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2305.00944*.

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Muhao Chen, Junjie Hu, Yixuan Li, Bo Li, and Chaowei Xiao. 2024a. Mitigating fine-tuning jailbreak attack with backdoor enhanced alignment. *arXiv preprint arXiv:2402.14968*.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023a. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. 2023b. On the exploitability of reinforcement learning with human feedback for large language models. *arXiv preprint arXiv:2311.09641*.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. 2024b. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *arXiv preprint arXiv:2403.09513*.

Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 305–314.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Fangzhao Wu, Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, and Xing Xie. 2023a. Defending chatgpt against jailbreak attack via self-reminder.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024a. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *NAACL*.

Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024b. Instructional fingerprinting of large language models. In *NAACL*.

Nan Xu, Fei Wang, Ben Zhou, Bang Zheng Li, Chaowei Xiao, and Muhao Chen. 2024c. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *NAACL*.

Jun Yan, Vansh Gupta, and Xiang Ren. Bite: Textual backdoor attacks with iterative trigger injection. In *ACL*.

Yixiang Yao, Fei Wang, Srivatsan Ravi, and Muhao Chen. 2023. Privacy-preserving language model inference with instance obfuscation. In *EMNLP (in submission)*.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021a. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021b. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pages 12208–12218. PMLR.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 1321–1342.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022a. Distillation-resistant watermarking for model protection in nlp. In *the Conference on Empirical Methods in Natural Language Processing (EMNLP) - Findings*.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. 2022b. Provably confidential language modelling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 943–955, Seattle, United States. Association for Computational Linguistics.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. 2023a. Protecting language generation models via invisible watermarking. In *Proceedings of the 40th International Conference on Machine Learning*.

Zhongkai Zhao, Bonan Kou, Mohamed Yilmaz Ibrahim, Muhao Chen, and Tianyi Zhang. 2023b. Knowledge-based version incompatibility detection for deep learning. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023a. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Xin Zhou, Jinzhu Lu, Tao Gui, Ruotian Ma, Zichu Fei, Yuran Wang, Yong Ding, Yibo Cheung, Qi Zhang, and Xuan-Jing Huang. 2022. Textfusion: Privacy-preserving pre-trained model inference via token fusion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8360–8371.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Appendix

A.1 Past Tutorials by the Instructors

The presenters of this tutorial have given the following tutorials at leading international conferences in the past.

- Muhao Chen:
  - ACL'23: Indirectly Supervised Natural Language Processing.
  - NAACL'22: New Frontiers of Information Extraction.
  - ACL'21: Event-Centric Natural Language Processing.
  - AAAI'21: Event-Centric Natural Language Understanding.
  - KDD'21: From Tables to Knowledge: Recent Advances in Table Understanding.
  - AAAI'20: Recent Advances of Transferable Representation Learning.
- Chaowei Xiao:
  - CVPR'23: Trustworthy AI in the Era of Foundation Models.
- Huan Sun:
  - SIGMOD'23: Models and Practice of Neural Table Representations
  - KDD'21: From Tables to Knowledge: Recent Advances in Table Understanding.
  - KDD'14: Network Mining and Analysis for Social Applications.
- Lei Li:
  - CCF-ADL 2022: Pre-training for Neural Machine Translation.
  - ACL'21: Pre-training Methods for Neural Machine Translation.
  - EMNLP'19: Discreteness in Natural Language Processing.
  - NLPCC'19: Deep Generative Models for Text Generation.
  - NLPCC'16: Deep Learning for Question Answering.
  - 2014 PPAML Summer School: Probabilistic Modeling using Bayesian Logic.
  - KDD'10: Indexing and Mining Time Sequences.
- Leon Derczynski:
  - COLING'20: Detection and Resolution of Rumors and Misinformation with NLP
  - RANLP'15: NLP for Social Media
  - ESWC'15: Practical Annotation and Processing of Social Media with GATE
  - LREC'14: Practical Social Media Analysis: finding utility in trivia
  - EACL'14: Natural Language Processing for Social Media
- Anima Anandkumar:
  - ECCV'20: New Frontiers for Learning with Limited Labels or Data.
  - ACM SIGMETRICS'18: The Role of Tensors in Deep Learning.
  - ICML'16: Recent Advances in Non-Convex Optimization.
  - AAAI'14: Tensor Decompositions for Learning Latent Variable Models.
  - ICML'13: Tensor Decomposition Algorithms for Latent Variable Model Estimation.

A.2 Recommended Paper List

The following is a reading list that could help provide background knowledge to the audience before attending this tutorial:

- Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, Maosong Sun. ONION: A Simple and Effective Defense Against Textual Backdoor Attacks. ACL 2021 (Qi et al., 2021a)
- Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, Tom Goldstein. On the Exploitability of Instruction Tuning. 2023 (Shu et al., 2023)
- Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, Muhao Chen. Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning for Large Language Models. 2023 (Xu et al., 2024a)
- Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, Chaowei Xiao. Adversarial Demonstration Attacks on Large Language Models. 2023 (Wang et al., 2023a)

- Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, Sherman S. M. Chow. Differential Privacy for Text Analytics via Natural Text Sanitization. Findings of ACL 2021 (Yue et al., 2021)
- Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. ACL 2023 Main Conference (Honorable Mention) (Yue et al., 2023)
- Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, Florian Tramèr. What does it mean for a language model to preserve privacy? FAccT 2022 (Brown et al., 2022)
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein. A Watermark for Large Language Models. ICML 2023 (Kirchenbauer et al., 2023)
- Xuandong Zhao, Yu-Xiang Wang, Lei Li. Protecting Language Generation Models via Invisible Watermarking. ICML 2023 (Zhao et al., 2023a)
- Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M.R. Leiser, Saif Mohammad. Assessing Language Model Deployment with Risk Cards. 2023 (Derczynski et al., 2023)
- Ali Borji. A categorical archive of chatgpt failures. 2023 (Borji, 2023)