Known Attacks and Their Impacts on AI Systems

March 2025

AISI Japan AI Safety Institute

Abstract

- This document provides an overview of attacks specific to AI systems and their impacts.
- As AI systems evolve, **attacks exploiting their characteristics** are increasingly reported, leading to issues like **training data leaks** and **abnormal AI outputs**.
- In the real world, these issues affect people's lives:
  - **Leaked training data** from AI used in medical diagnosis violates privacy.
  - **Abnormal AI outputs** in automated driving can cause traffic accidents.
- Thus, AI-specific security measures are essential.
- This document outlines attacks and their impacts, as reported in academic papers and other sources, to aid in examining countermeasures.

Scope of This Document

AISI Japan AI Safety Institute

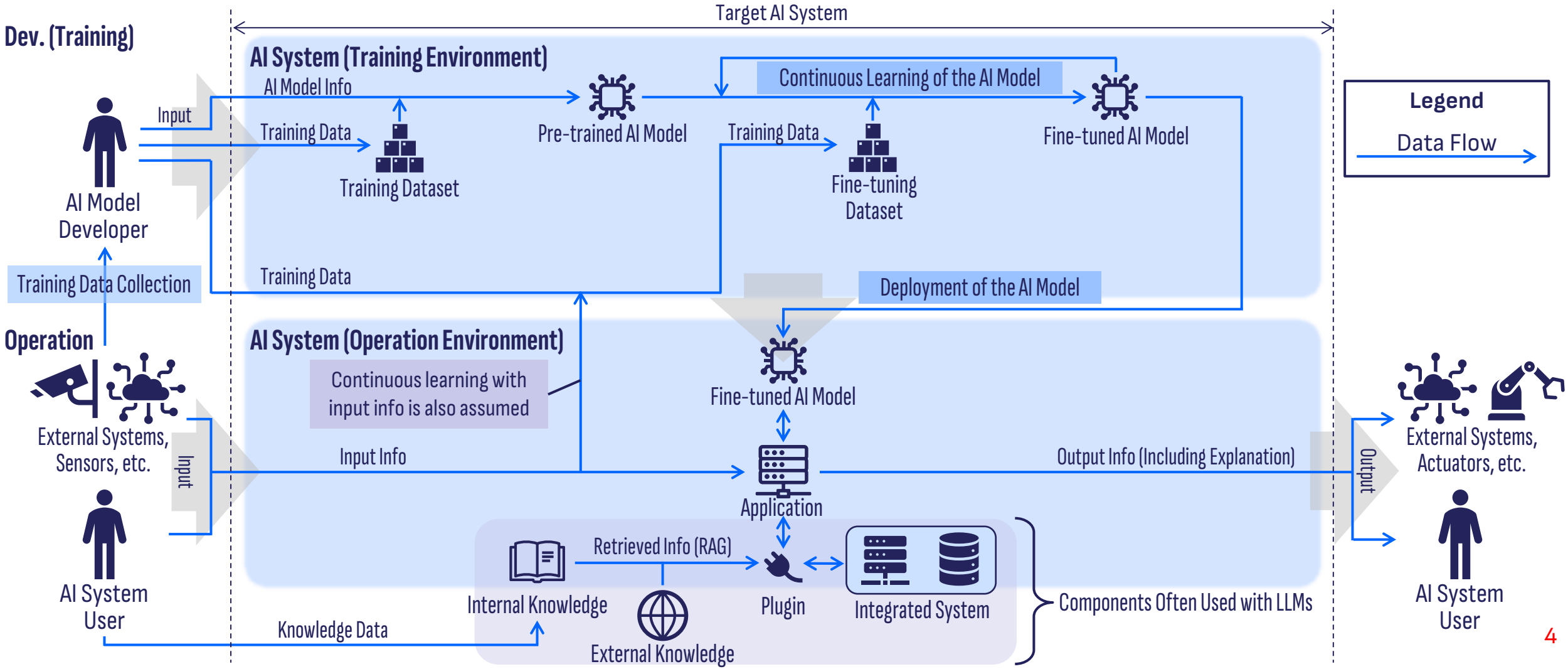
- Measures against conventional cybersecurity attacks are assumed to be necessary.
- This document focuses on known attacks that require special attention when a system incorporates AI, specifically, those involve **intervening in an AI model's training or inference processes**, as well as **exploiting or tampering with its inputs and outputs**.
  - Examples **Within** the Scope:
    - *During Training:* Data Poisoning Attacks (Intervention in the Training Process)
    - *During Inference:* Model Extraction Attacks (Exploit the AI model's output)
  - Examples **Outside** the Scope:
    - Theft of AI Models Through Unauthorized Access by Exploiting Network Device Vulnerabilities.

      Note: Although this attack targets AI models, it does not involve interfering with the training or inference process of the AI models in the target system, nor does it manipulate its input or output.

AI Systems Assumed in This Document

AISI Japan AI Safety Institute

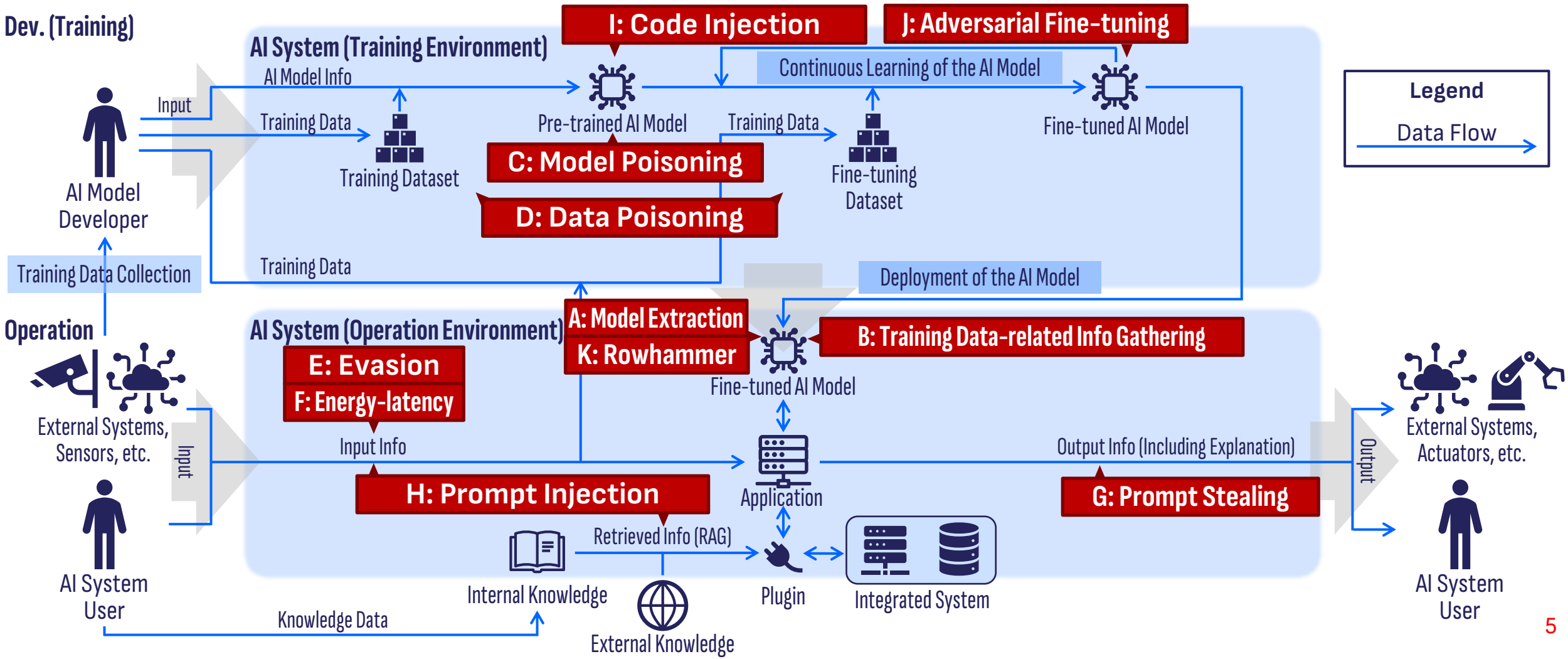
- We assume AI systems consisting of two types of environments: development (training) and operation.
- Although we mention RAG and other components often used with LLMs, the AI model is not limited to LLMs.



Overview of Attacks on AI Systems

AISI Japan AI Safety Institute

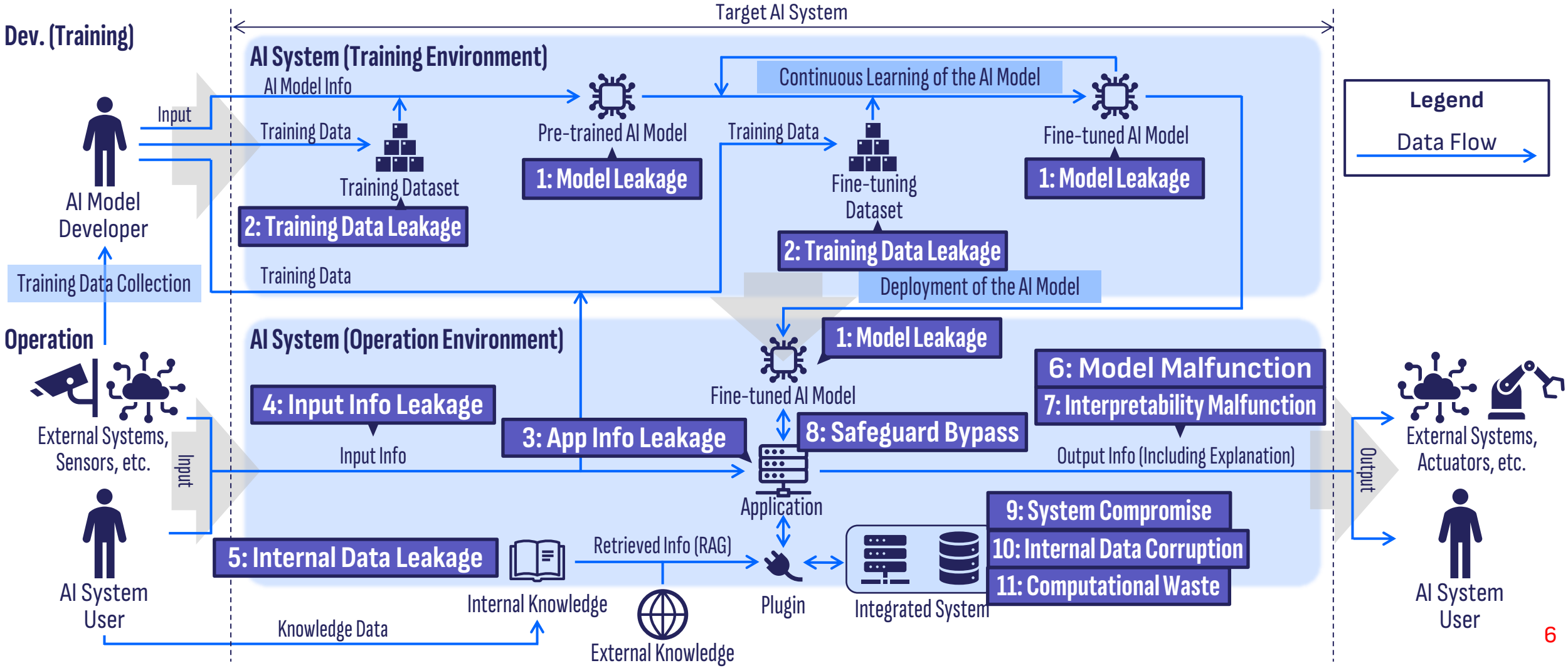
- The following figure illustrates attacks on the assumed AI system (with the scope described earlier).
- Training data-related information gathering and other attacks can be further classified into multiple types.



Overview of Impacts of Attacks on AI Systems

AISI Japan AI Safety Institute

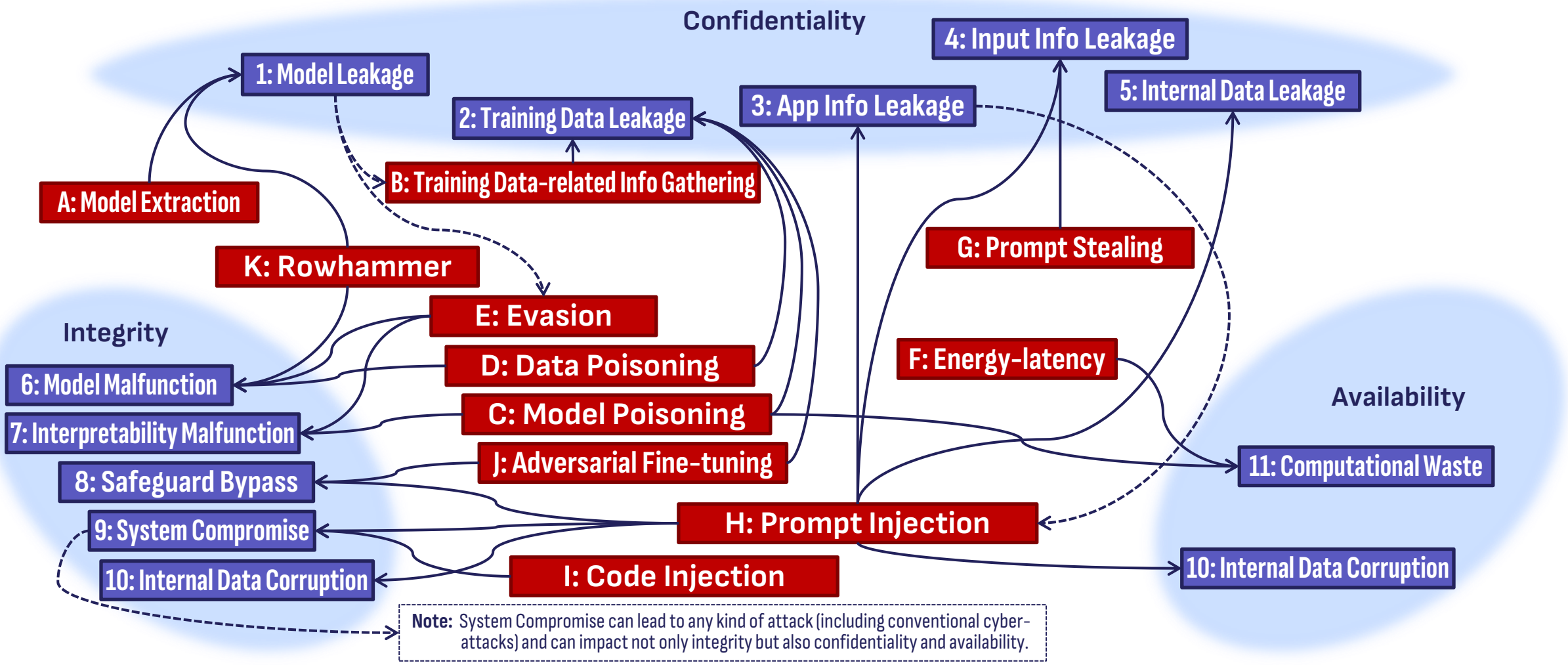
- The following figure illustrates the impacts of attacks described in the previous slide.
- In some cases, the same type of impact may occur at multiple points, such as model leakage.



Attacks and Their Impacts on AI Systems: Overview

AISI Japan AI Safety Institute

- The relationship between the attacks and their impacts described earlier is as follows.
- The dashed line indicates the potential for use in an attack at the arrow's endpoint.



Note: System Compromise can lead to any kind of attack (including conventional cyber-attacks) and can impact not only integrity but also confidentiality and availability.

Attacks and Their Impacts on AI Systems: Detailed List

AISI Japan AI Safety Institute

Attack		Impact	Attacker's Capability	Modality : Target AI[Literature] (This is summary from literature and does not limit modality or target)
A	Model Extraction	Model Leakage	Massive queries to the model	Image: NN[OSF19a,OSF19b,JSM+19,CBB+18], LR[TZJ+16] Tabular: RR[WG18], DT[TZJ+16], LR,NN,SVM[TZJ+16,WG18]
B	Membership Inference (Training Data-related Information Gathering)	Training Data Leakage	Input to the model	Image, Tabular: LR,DT[YGF+18], NN[AYM+19,YGF+18] Text: NN[AYM+19]
			Massive queries to the model	Image, Tabular: NN[SSS+17,LBW+18] Text: LM[LSS+23]
			Access to the model's internal info	Image, Tabular: NN[NSH19]
	Attribute Inference	Training Data Leakage	Input to the model	Image, Tabular: LR, DT[YGF+18]
			Access to the model's internal info	Image, Text: NN[SS20]
	Property Inference	Training Data Leakage	Access to the model's internal info	Image, Tabular: NN[GWY+18] Audio, Network Traffic: NN, SVM, HMM, DT[AMS+15]
	Model Inversion	Training Data Leakage	Massive queries to the model	Image, Tabular: DT, NN[FJR15] Text: Transformer[ZHK22]
	Data Extraction	Training Data Leakage	Massive queries to the model	Text: LSTM, RNN[CLE+19], LM[CTW+21,LSS+23]
C	Model Poisoning	Interpretability Malfunction	Model tampering	Tabular: NN, EM[SHJ+20]
		Training Data Leakage	Training program tampering	Image: NN[SRS17] Text: SVM, LR[SRS17]
		Computational Waste	Model tampering	Image: NN[CDB+23]
D	Data Poisoning	Model Malfunction	Training data injection	Image: NN[Car21] Text: LM[Sch19]
		Training Data Leakage	Training data injection, Massive queries to the model	Tabular: LR, NN[MGC22] Text: LR[MGC22]
E	Evasion	Model Malfunction	Massive queries to the model	Image: NN, LR, SVM, DT, kNN[PMG+17] Text: LM[BSA+22]
			Access to the model's internal info	Image: NN[SZS+14,GSS15] Text: LSTM[ERL+18]
			Input to the model	Text: Transformer[GSJ+21], LSTM, DA[WFK+19]
		Interpretability Malfunction	Massive queries to the model	Image: NN[DAA+19]
F	Energy-latency	Computational Waste	Input to the model	Image, Text: NN[SZB+21]
			Massive queries to the model	Text: LM[BSA+22]
G	Prompt Stealing	Input Info Leakage	Access to the model's output	Text to Image: Diffusion[SQB+24]
H	Prompt Injection	Safeguard Bypass	Input to the model	Text: LM[LDL+24,SCB+24,ZWC+23,WHS23,CRD+24,MZK+24,PHS+22,WFK+19]
			Access to the model's internal info	Text, Image: NN[CNC+23]
		App Info Leakage	Input to the model via system API	Text: LM[ZCI24]
		System Compromise	AI system-referenced info poisoning	Text, Image to Text: LM[GAM+23]
		Internal Data Leakage/Corruption	Input to the model via system API	Text to Tabular: LM[PEC+25]
		Input Info Leakage	User-referenced info poisoning	Text: A Specific Chat Service[Sam23]
I	Code Injection	System Compromise	Malicious model distribution	Any: Any[ZWZ+24]
J	Adversarial Fine-tuning	Safeguard Bypass	Access to the model's internal info	Text: LM[YYC+24]
		Training Data Leakage	Access to fine-tuning functionality	Text: LM[CTZ+24]
K	Rowhammer	Model Malfunction	Access to physical memory	Image: NN[LWX+24]
			Access to model's internal info, physical memory	Image, Text: LM, Transformer[NMF+24]
		Model Leakage	Access to physical memory	Image: NN[RCY+22]

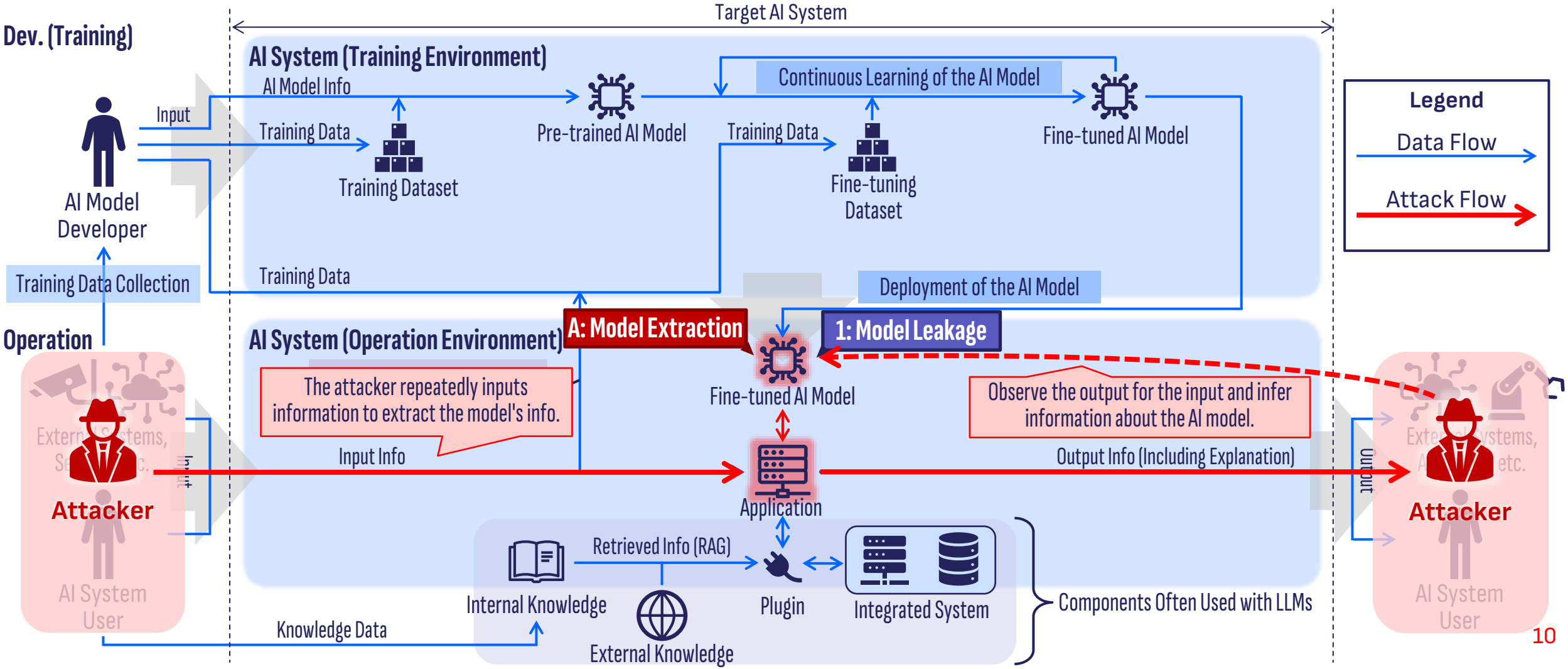
NN: Neural Network, LR: Logistic Regression, RR: Ridge Regression, SVM: Support Vector Machine, DT: Decision Tree, kNN: Nearest Neighbor, HMM: Hidden Markov Model, LM: Language Model, LSTM: Long-short Term Memory, RNN: Recurrent Neural Network, EM: Ensemble Method, DA: Decomposable Attention

Overview of Each Attack

Attack A: Model Extraction

AISI Japan AI Safety Institute

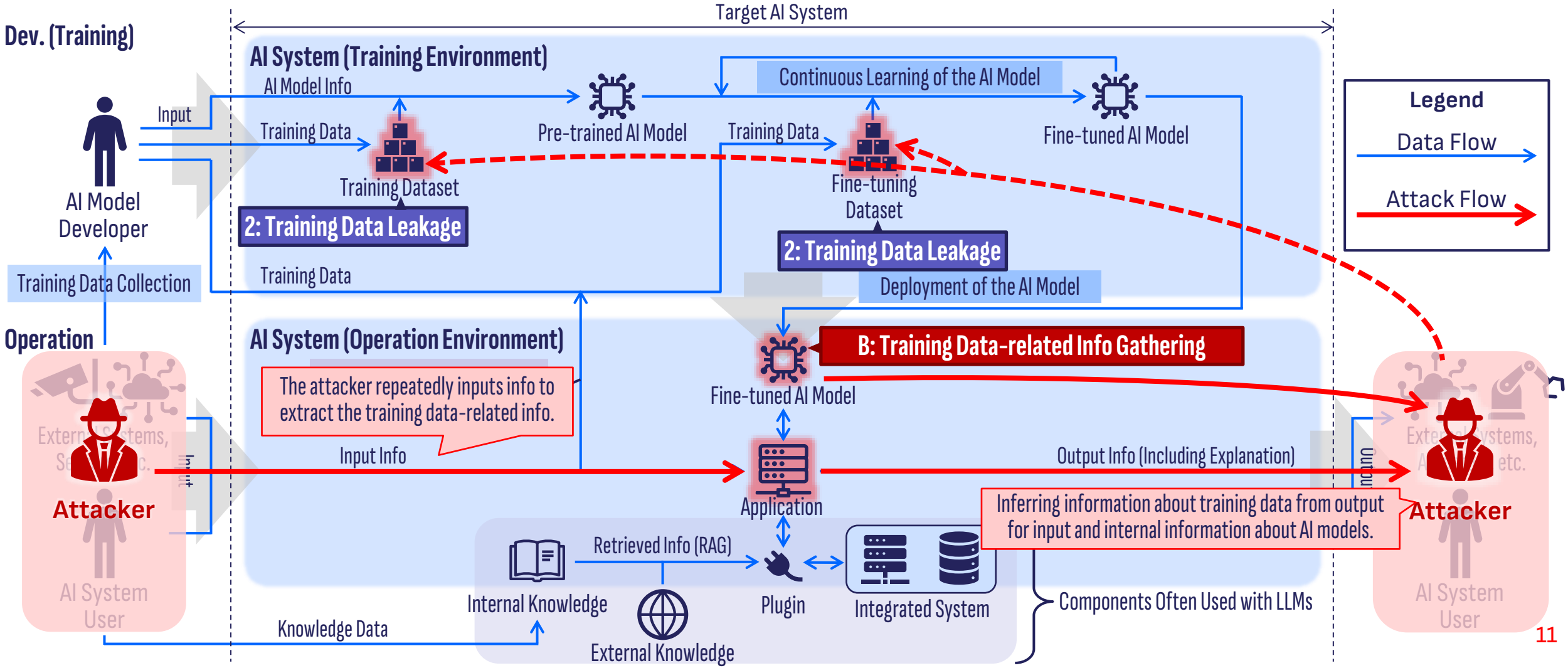
- By observing the output produced in response to given input data, this attack can cause the leakage of information about the AI model without requiring direct access to it.



Attack B: Training Data-related Info Gathering

AISI Japan AI Safety Institute

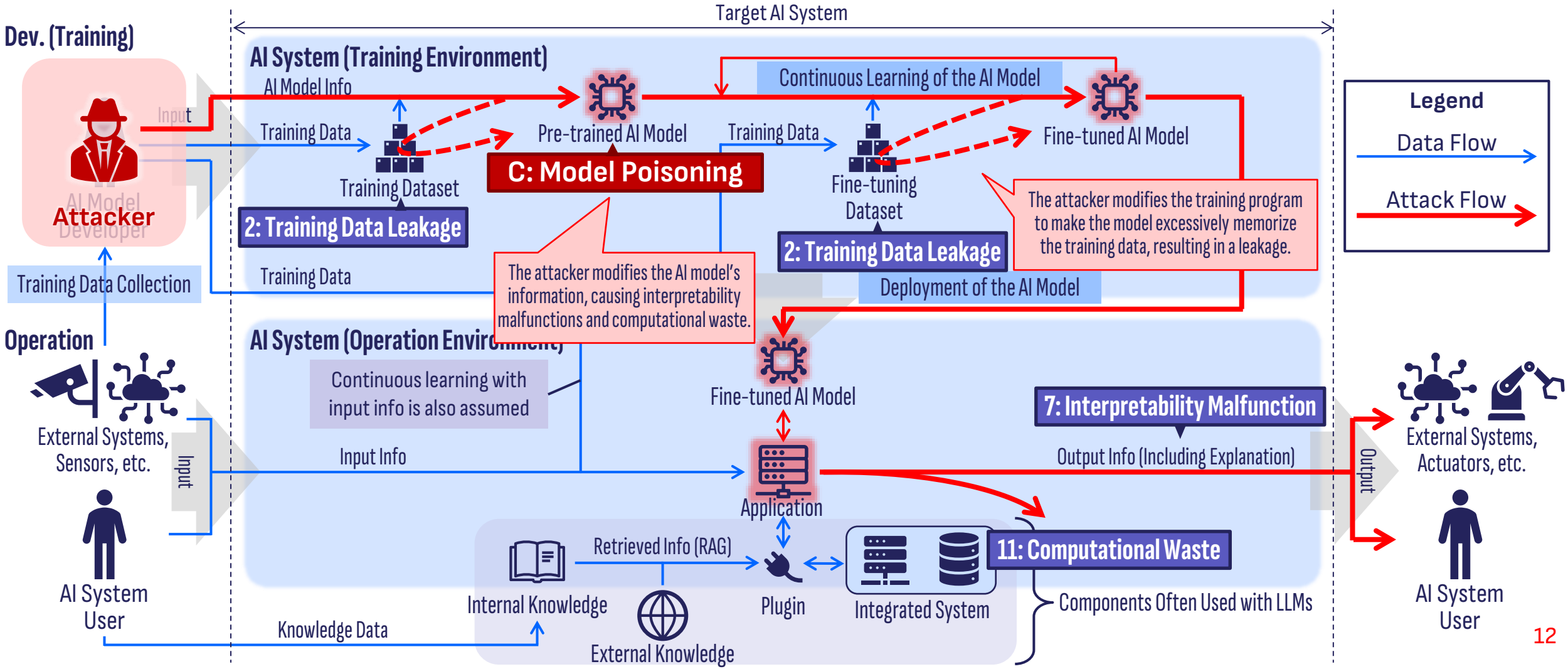
- Information about the training dataset is gathered from the AI model's internal details and the correspondence between its inputs and outputs. There are variations of the attack depending on the type of information targeted.



Attack C: Model Poisoning

AISI Japan AI Safety Institute

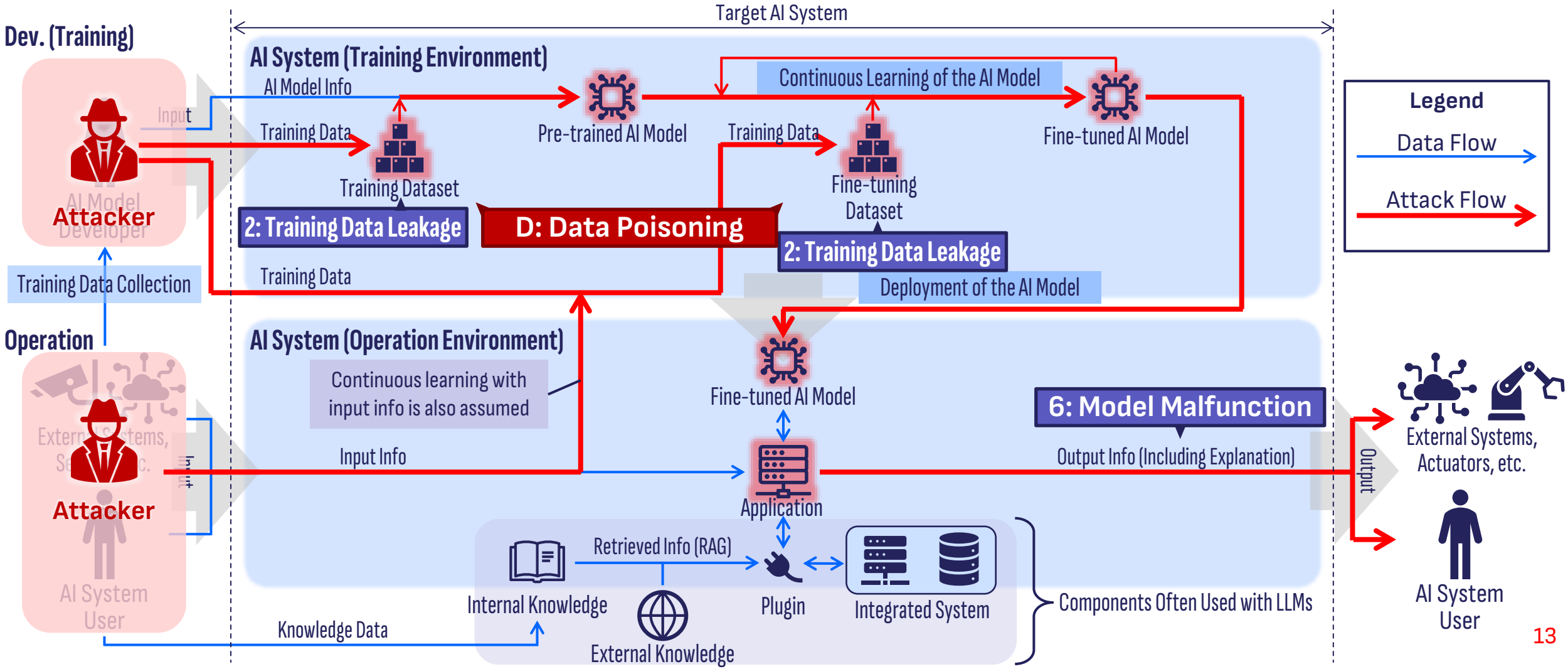
- By modifying the information or training programs of the AI model, this attack can cause interpretability malfunctions, computational waste, and training data leakage when the model is in operation.



Attack D: Data Poisoning

AISI Japan AI Safety Institute

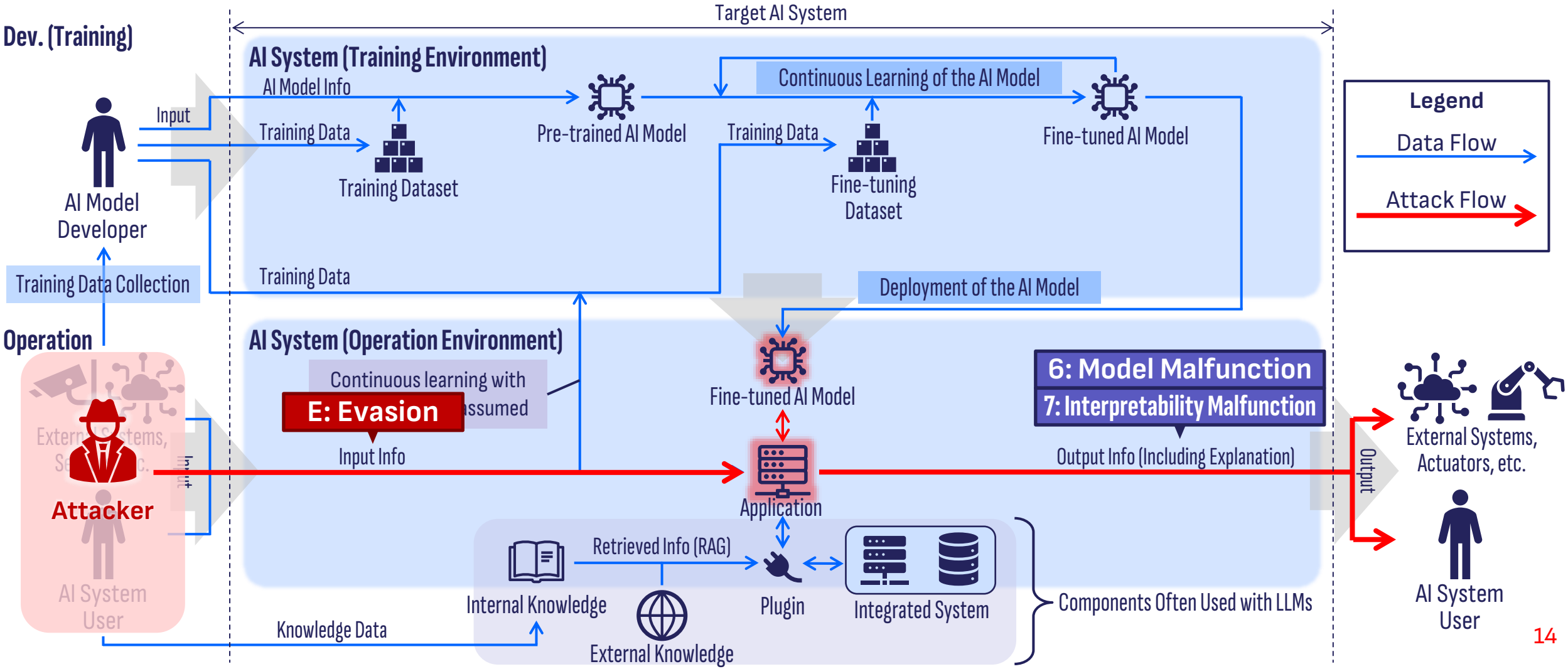
- By inserting special data into the training dataset, this attack can cause the model to malfunction when processing input information during operation or the leakage of information about the remaining training data.



Attack E: Evasion

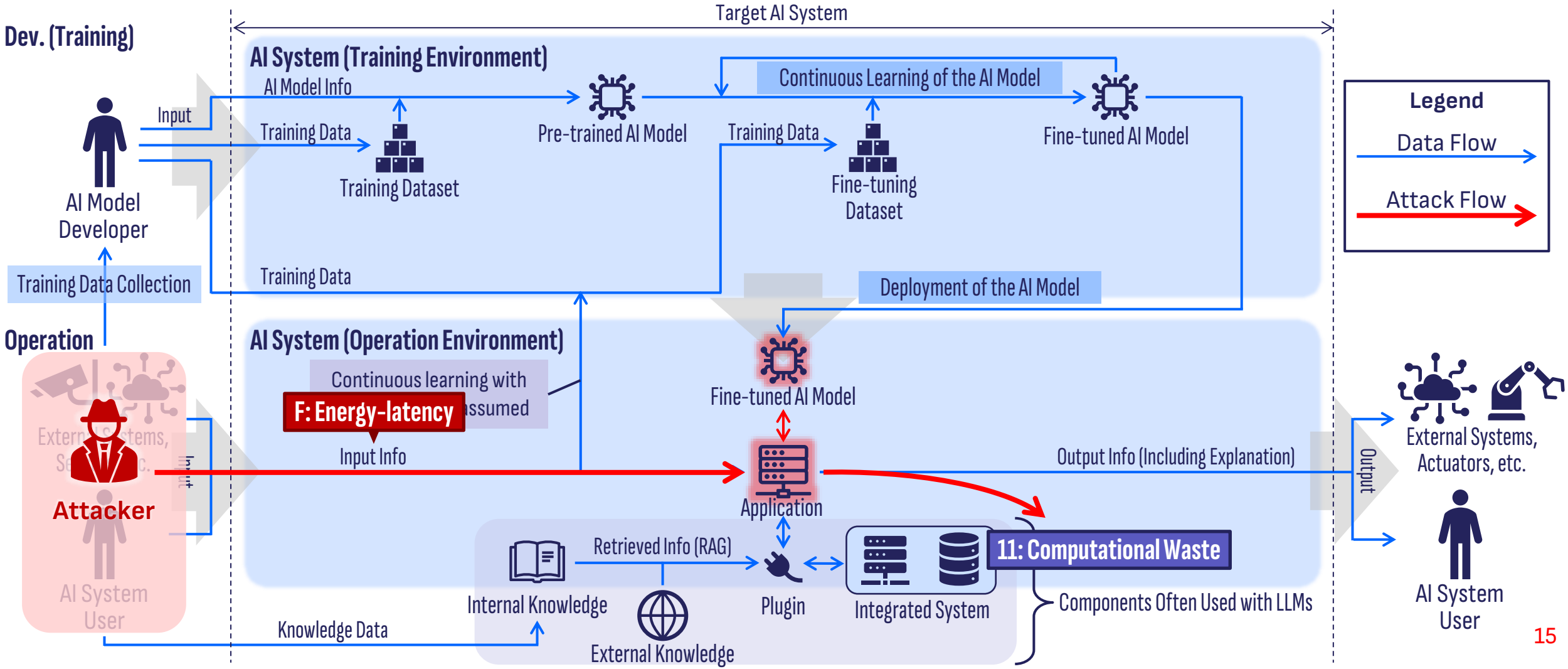
AISI Japan AI Safety Institute

- By providing the AI system with specially crafted inputs called adversarial examples during operation, this attack can cause malfunctions in both its output and interpretability.



Attack F: Energy-latency

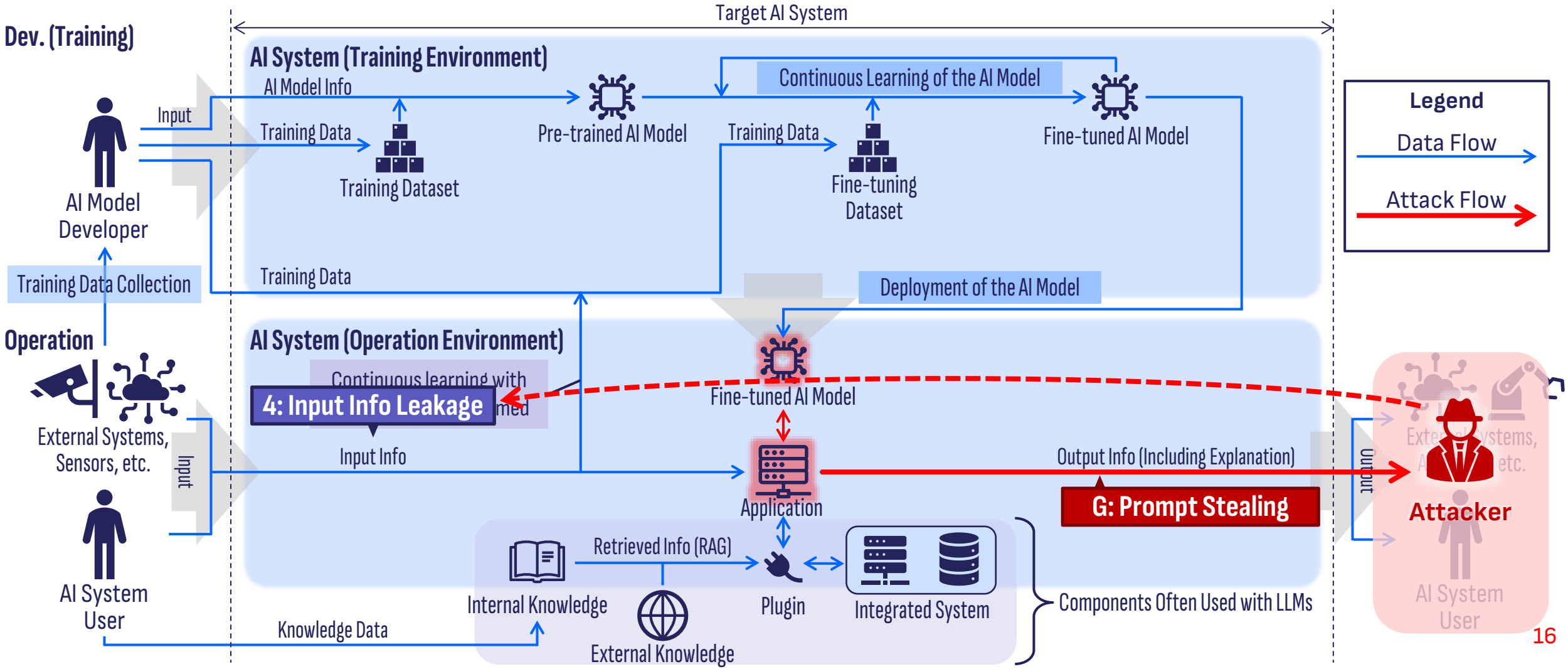
- By providing the AI system with specially crafted inputs called sponge examples during operation, this attack can waste computing resources, such as by increasing response time and energy consumption.



Attack G: Prompt Stealing

AISI Japan AI Safety Institute

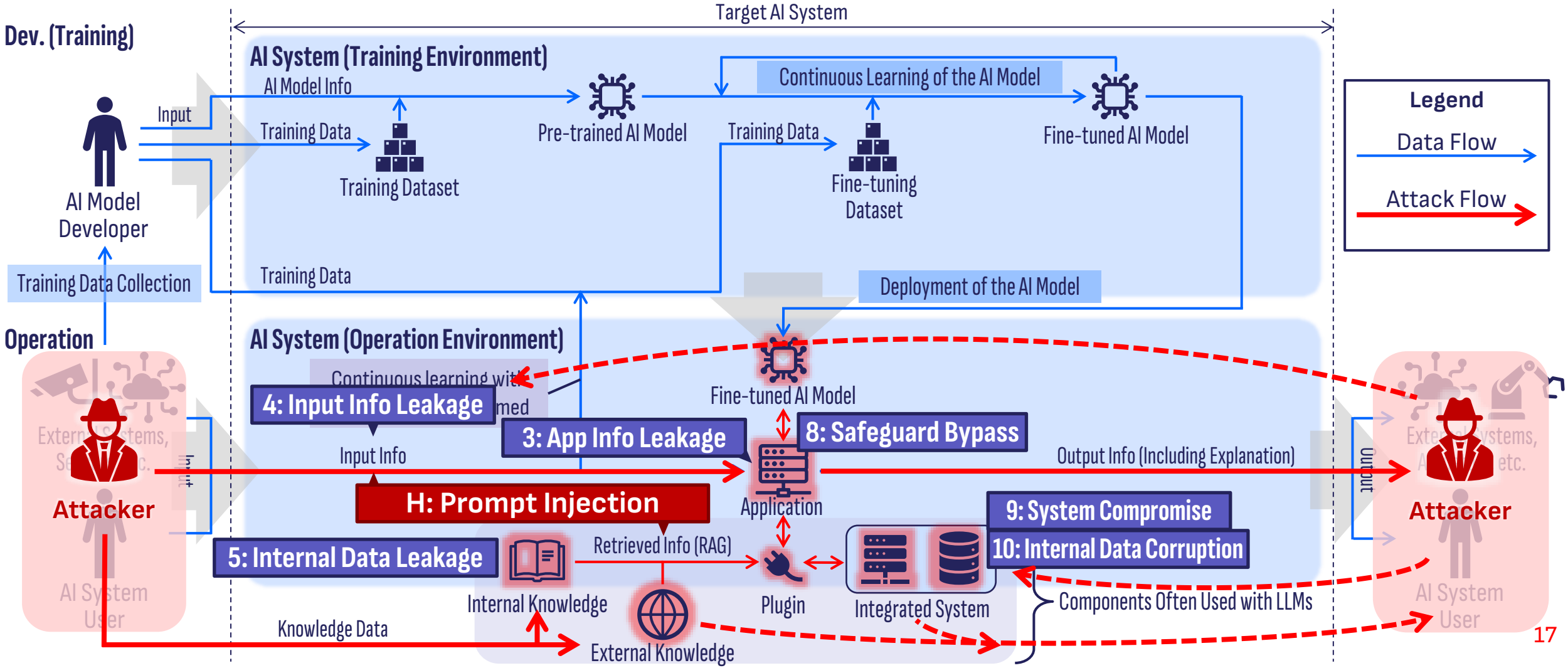
- By inferring the original prompt from outputs such as images generated by the AI model, this attack can leak input information that constitutes prompt engineering know-how.



Attack H: Prompt Injection

AISI Japan AI Safety Institute

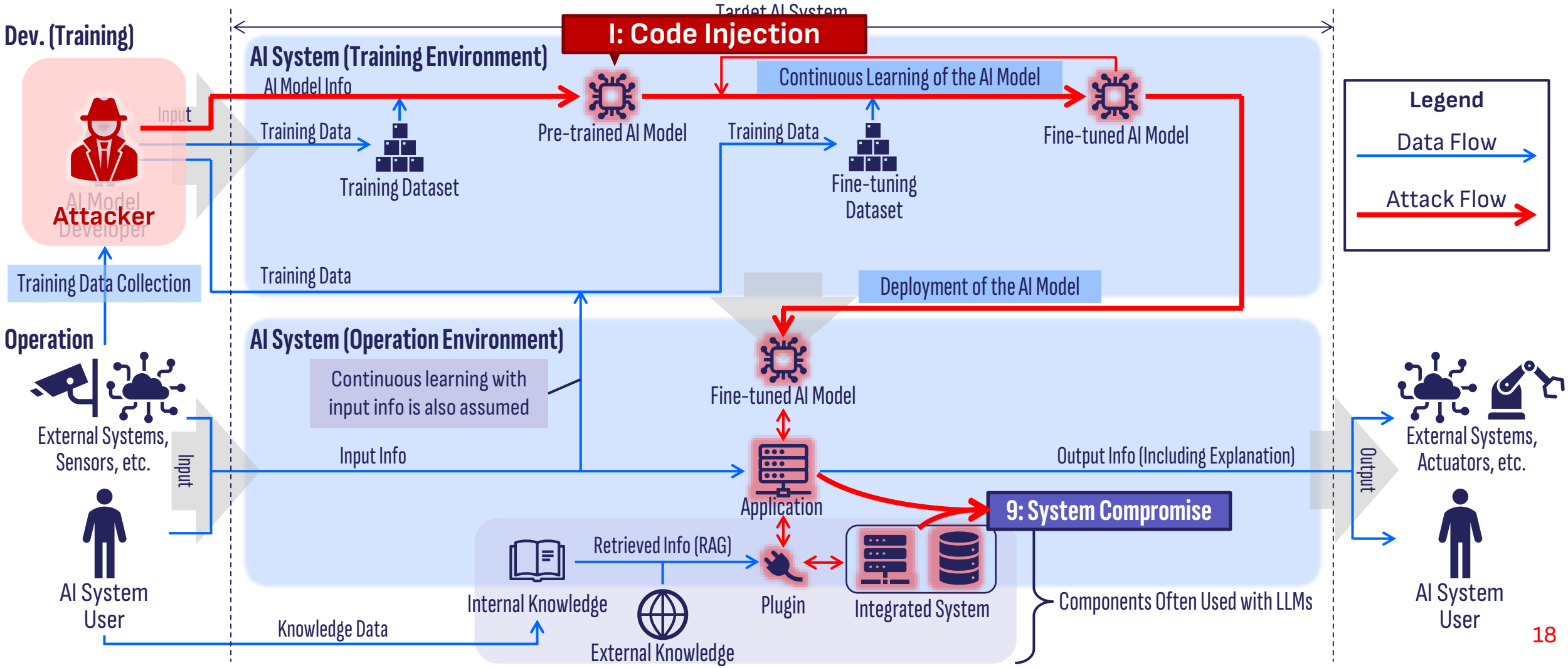
- By directly inputting adversarial instructions into the model or indirectly injecting them through its information sources, this attack can cause various forms of information leakage, safeguard bypass, and system compromise.



Attack I: Code Injection

AISI Japan AI Safety Institute

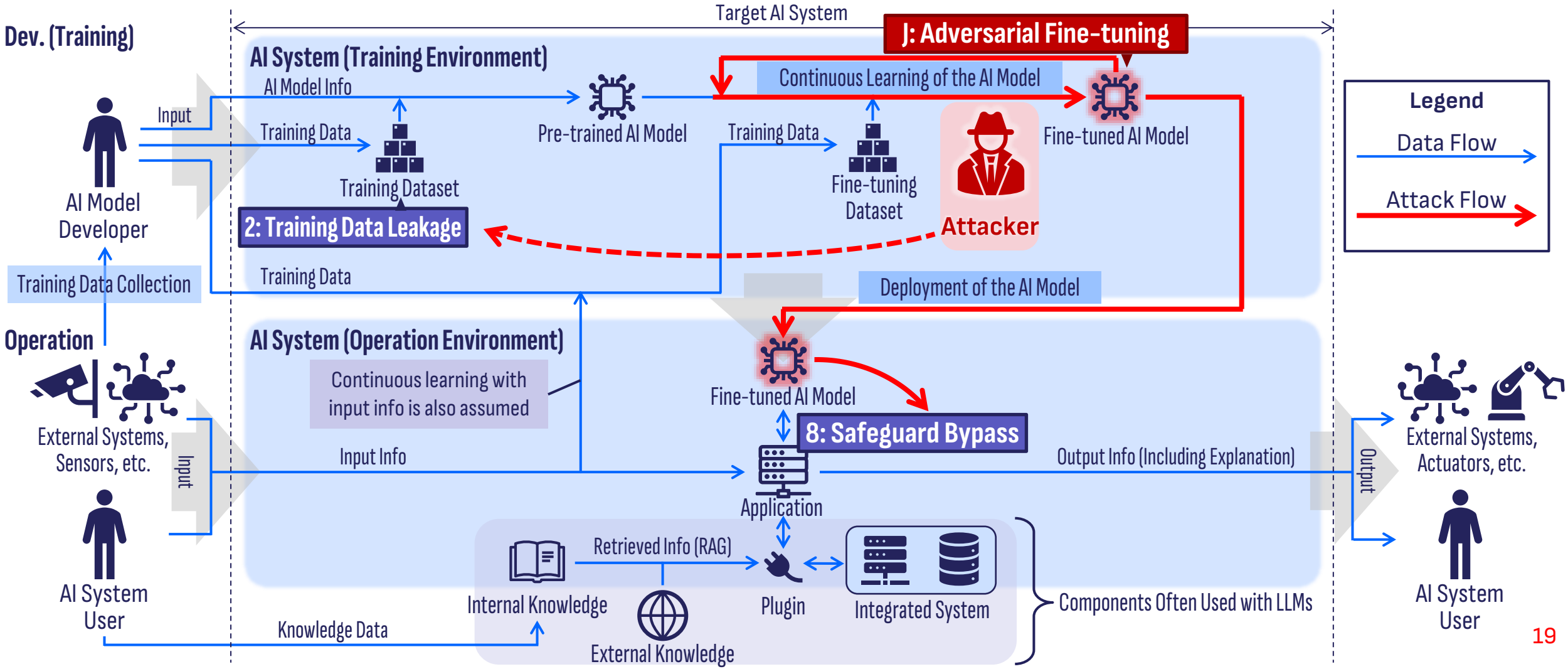
- Embedding executable code within the AI model itself and ensuring that the embedded code is executed when the AI model is invoked can lead to system compromise.



Attack J: Adversarial Fine-tuning

AISI Japan AI Safety Institute

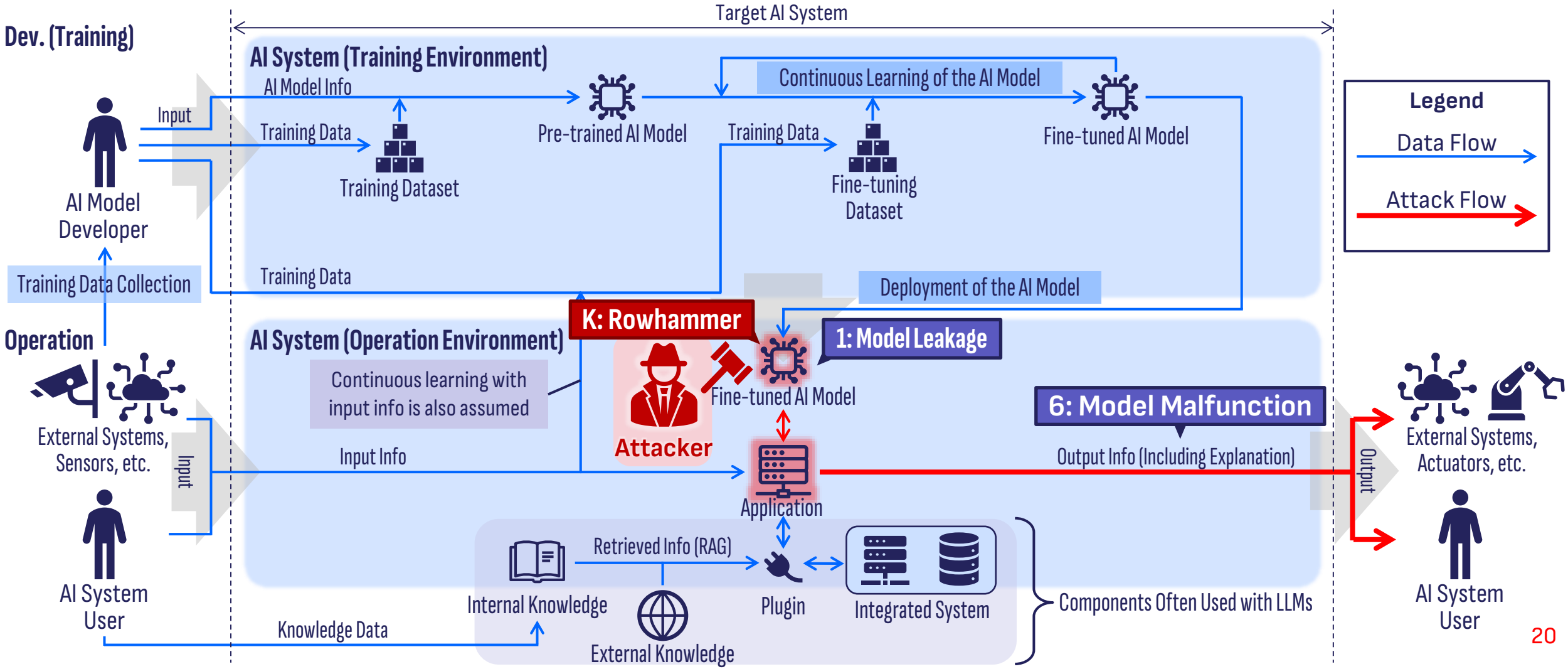
- By performing specific fine-tuning on the target pre-trained AI model, the AI model is manipulated to bypass safeguards and leak its pre-trained data.



Attack K: Rowhammer

AISI Japan AI Safety Institute

- An attacker who shares physical memory with the target AI model induces bit flips in the model's memory through memory cell interference, causing model leakage or malfunction.



References (Part 1 of 4)

AISI Japan AI Safety Institute

[AMS+15] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, et al. "Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers". In: Int. J. Secur. Netw. 10.3 (Sept. 2015), pp. 137–150. doi: 10.1504/IJSN.2015.071829.

[AYM+19] Salem Ahmed, Zhang Yang, Humbert Mathias, Berrang Pascal, Fritz Mario, et al. "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models". In: Proceedings 2019 Network and Distributed System Security Symposium (2019). doi: 10.14722/ndss.2019.23119.

[BSA+22] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. "Bad Characters: Imperceptible NLP Attacks". In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1987–2004. doi: 10.1109/SP46214.2022.9833641.

[Car21] Nicholas Carlini. "Poisoning the Unlabeled Dataset of Semi-Supervised Learning". In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Aug. 2021, pp. 1577–1592. url: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

[CBB+18] Jacson Rodrigues Correia-Silva, Rodrigo F. Berriel, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos. "Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data". In: 2018 International Joint Conference on Neural Networks (IJCNN). 2018, pp. 1–8. doi: 10.1109/IJCNN.2018.8489592.

[CDB+23] Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Energy-Latency Attacks via Sponge Poisoning. 2023. arXiv: 2203.08147 [cs.CR].

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks". In: 28th USENIX Security Symposium (USENIX Security 19). Santa Clara, CA: USENIX Association, Aug. 2019, pp. 267–284. url: https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

[CNC+23] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, et al. "Are aligned neural networks adversarially aligned?" In: Advances in Neural Information Processing Systems. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 61478–61500. url: https://proceedings.neurips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.

[CRD+24] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, et al. Jailbreaking Black Box Large Language Models in Twenty Queries. 2024. arXiv: 2310.08419 [cs.LG].

[CTW+21] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, et al. "Extracting Training Data from Large Language Models". In: 30th USENIX Security Symposium (USENIX Security 21). USENIX Association, Aug. 2021, pp. 2633–2650. url: https://www.usenix.org/conference/ usenixsecurity21/presentation/carlini-extracting.

[CTZ+24] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, et al. "The Janus Interface: How Fine-Tuning in Large Language Models Amplifies the Privacy Risks". In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. CCS '24. Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 1285–1299. doi: 10.1145/3658644.3690325.

[DAA+19] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, et al. "Explanations can be manipulated and geometry is to blame". In: Advances in Neural Information Processing Systems. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, et al. Vol. 32. Curran Associates, Inc., 2019. url: https://proceedings.neurips.cc/paper/2019/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html.

References (Part 2 of 4)

AISI Japan AI Safety Institute

[ERL+18] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. "HotFlip: White-Box Adversarial Examples for Text Classification". In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia, July 2018, pp. 31–36. doi: 10.18653/v1/P18-2006.

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. CCS '15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. doi: 10.1145/2810103.2813677.

[GAM+23] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, et al. "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection". In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. AISec '23. Copenhagen, Denmark: Association for Computing Machinery, 2023, pp. 79–90. doi: 10.1145/3605764. 3623985.

[GSJ+21] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. "Gradient-based Adversarial Attacks against Text Transformers". In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 5747–5757. doi: 10.18653/v1/2021. emnlp-main.464.

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. Ed. by Yoshua Bengio and Yann LeCun. 2015. arXiv: 1412.6572 [stat.ML].

[GWY+18] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations". In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. CCS '18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 619–633. doi: 10.1145/3243734.3243834.

[JSM+19] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. "PRADA: Protecting Against DNN Model Stealing Attacks". In: 2019 IEEE European Symposium on Security and Privacy (EuroS&P). 2019, pp. 512–527. doi: 10.1109/EuroSP.2019.00044.

[LBW+18] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, et al. Understanding Membership Inferences on Well-Generalized Learning Models. 2018. arXiv: 1802.04889 [cs.CR].

[LDL+24] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, et al. Prompt Injection attack against LLM-integrated Applications. 2024. arXiv: 2306.05499 [cs.CR].

[LSS+23] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, et al. "Analyzing Leakage of Personally Identifiable Information in Language Models". In: 2023 IEEE Symposium on Security and Privacy (SP). 2023, pp. 346–363. doi: 10.1109/SP46215.2023.10179300.

[LWX+24] Shaofeng Li, Xinyu Wang, Minhui Xue, Haojin Zhu, Zhi Zhang, et al. "Yes, One-Bit-Flip Matters! Universal DNN Model Inference Depletion with Runtime Code Fault Injection". In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1315–1330. url: https://www.usenix.org/conference/usenixsecurity24/presentation/li-shaofeng.

[MGC22] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. "Property Inference from Poisoning". In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1120–1137. doi: 10.1109/SP46214.2022.9833623.

[MZK+24] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, et al. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. 2024. arXiv: 2312.02119 [cs.LG]

[NMF+24] Najmeh Nazari, Hosein Mohammadi Makrani, Chongzhou Fang, Hossein Sayadi, Setareh Rafatirad, et al. "Forget and Rewire: Enhancing the Resilience of Transformer-based Models against Bit-Flip Attacks". In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1349–1366. url: https://www.usenix.org/conference/usenixsecurity24/presentation/nazari.

References (Part 3 of 4)

AISI Japan AI Safety Institute

[NSH19] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning”. In: 2019 IEEE Symposium on Security and Privacy (SP). 2019, pp. 739–753. doi: 10.1109/SP.2019.00065.

[OSF19a] Seong Joon Oh, Bernt Schiele, and Mario Fritz. “Towards Reverse-Engineering Black-Box Neural Networks”. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Ed. by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. Cham: Springer International Publishing, 2019, pp. 121–144. doi: 10.1007/978-3-030-28954-6_7.

[OSF19b] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Knockoff Nets: Stealing Functionality of BlackBox Models”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019, pp. 4949–4958. doi: 10.1109/CVPR.2019.00509.

[PEC+25] Rodrigo Pedro, Miguel E. Coimbra, Daniel Castro, Paulo Carreira, and Nuno Santos. “Prompt-to-SQL Injections in LLM-Integrated Web Applications: Risks and Defenses”. In: 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE). Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 76–88. doi: 10.1109/ICSE55347.2025.00007.

[PHS+22] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, et al. “Red Teaming Language Models with Language Models”. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Association for Computational Linguistics, Dec. 2022, pp. 3419–3448. doi: 10.18653/v1/2022.emnlp-main.225.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, et al. “Practical Black-Box Attacks against Machine Learning”. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ASIA CCS ’17. Abu Dhabi, United Arab Emirates: Association for Computing Machinery, 2017, pp. 506–519. doi: 10.1145/3052973.3053009.

[RCY+22] Adnan Siraj Rakin, Md Hafizul Islam Chowdhuryy, Fan Yao, and Deliang Fan. “DeepSteal: Advanced Model Extractions Leveraging Efficient Weight Stealing in Memories”. In: 2022 IEEE Symposium on Security and Privacy (SP). 2022, pp. 1157–1174. doi: 10.1109/SP46214.2022.9833743.

[Sam23] Roman Samoilenko. “New Prompt Injection Attack on ChatGPT Web Version. Markdown Images can Steal Your Chat Data.” In: System Weakness (Mar. 2023). url: https://systemweakness.com/newprompt-injection-attack-on-chatgpt-web-version-ef717492c5c2.

[SCB+24] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. “"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models”. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. CCS ’24. Salt Lake City, UT, USA, 2024, pp. 1671–1685. doi: 10.1145/3658644.3670388.

[Sch19] Oscar Schwartz. “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation”. In: IEEE Spectrum (Nov. 2019). Updated: Jan. 2024. url: https://spectrum.ieee.org/in- 2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[SHJ+20] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods”. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 180–186. doi: 10.1145/3375627.3375830.

[SQB+24] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. “Prompt Stealing Attacks Against Text-to-Image Generation Models”. In: 33rd USENIX Security Symposium (USENIX Security 24). Philadelphia, PA: USENIX Association, Aug. 2024, pp. 5823–5840. url: https://www.usenix.org/conference/ usenixsecurity24/presentation/shen-xinyue.

[SRS17] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. “Machine Learning Models that Remember Too Much”. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. CCS ’17. Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 587–601. doi: 10.1145/3133956.3134077.

[SS20] Congzheng Song and Vitaly Shmatikov. “Overlearning Reveals Sensitive Attributes”. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. url: https://openreview.net/forum?id=SJeNz04tDS.

References (Part 4 of 4)

AISI Japan AI Safety Institute

[SSS+17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models". In: 2017 IEEE Symposium on Security and Privacy (SP). 2017, pp. 3–18. doi: 10.1109/SP.2017.41.

[SZB+21] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, et al. "Sponge Examples: Energy-Latency Attacks on Neural Networks". In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P). 2021, pp. 212–231. doi: 10.1109/EuroSP51992.2021.00024.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, et al. "Intriguing properties of neural networks". In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. Ed. by Yoshua Bengio and Yann LeCun. 2014. arXiv: 1312.6199 [cs.CV].

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. "Stealing Machine Learning Models via Prediction APIs". In: 25th USENIX Security Symposium (USENIX Security 16). Austin, TX: USENIX Association, Aug. 2016, pp. 601–618. url: https://www.usenix.org/ conference/usenixsecurity16/technical-sessions/presentation/tramer.

[WFK+19] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. "Universal Adversarial Triggers for Attacking and Analyzing NLP". In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China, Nov. 2019, pp. 2153–2162. doi: 10.18653/v1/D19-1221.

[WG18] Binghui Wang and Neil Zhenqiang Gong. "Stealing Hyperparameters in Machine Learning". In: 2018 IEEE Symposium on Security and Privacy (SP). 2018, pp. 36–52. doi: 10.1109/SP.2018.00038.

[WHS23] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. "Jailbroken: How Does LLM Safety Training Fail?" In: Advances in Neural Information Processing Systems. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. Vol. 36. Curran Associates, Inc., 2023, pp. 80079–80110. url: https : //proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790- Abstract-Conference.html.

[YGF+18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting". In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). 2018, pp. 268–282. doi: 10.1109/CSF.2018.00027.

[YYC+24] Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, et al. "On the Vulnerability of Safety Alignment in Open-Access LLMs". In: Findings of the Association for Computational Linguistics: ACL 2024. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9236–9260. doi: 10.18653/v1/2024.findings-acl.549.

[ZCI24] Yiming Zhang, Nicholas Carlini, and Daphne Ippolito. "Effective Prompt Extraction from Language Models". In: First Conference on Language Modeling. 2024. url: https://openreview.net/forum?id=0o95CVdNuz.

[ZHK22] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. 2022. arXiv: 2209.10505 [cs.CL].

[ZWC+23] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. 2023. arXiv: 2307.15043 [cs.CL].

[ZWZ+24] Jian Zhao, Shenao Wang, Yanjie Zhao, Xinyi Hou, Kailong Wang, et al. "Models Are Codes: Towards Measuring Malicious Code Poisoning Attacks on Pre-trained Model Hubs". In: Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering. ASE '24. Sacramento, CA, USA: Association for Computing Machinery, 2024, pp. 2087–2098. doi: 10.1145/3691620.3695271

AISI

Japan AI Safety Institute

Japan AI Safety Institute
“Known Attacks and Their Impacts on AI Systems” (Mar. 2025)
https://aisi.go.jp/effort/effort_security/known_attacks_and_impacts/