AIシステムに対する既知の攻撃と影響

2025年 3月

AISI Japan AI Safety Institute

概要

AISI Japan AI Safety Institute

- 本資料は、**AIシステムに対する特有の攻撃とその影響**を俯瞰する。
- AIシステムの利活用が急速に進む中、**AIの特性を悪用した攻撃**が次々と発表されている。これらの攻撃は、学習データの漏洩や、AIモデルの異常な出力といった影響の原因となる。
- 実社会では、これらの影響が人々の生活に深く関係してくる。（以下、例）
  - 医療診断でのAIシステムの学習データの漏洩はプライバシー侵害になる。
  - 自動運転でのAIシステムの異常な出力は人命に関わる事故につながる。
- このようにAIシステムに特有のセキュリティ対策は今や必要不可欠である。
- 本資料は、学術論文等で発表された、AIシステムに特有の攻撃と影響を俯瞰し、対策を検討するための参考となる情報を提供する。

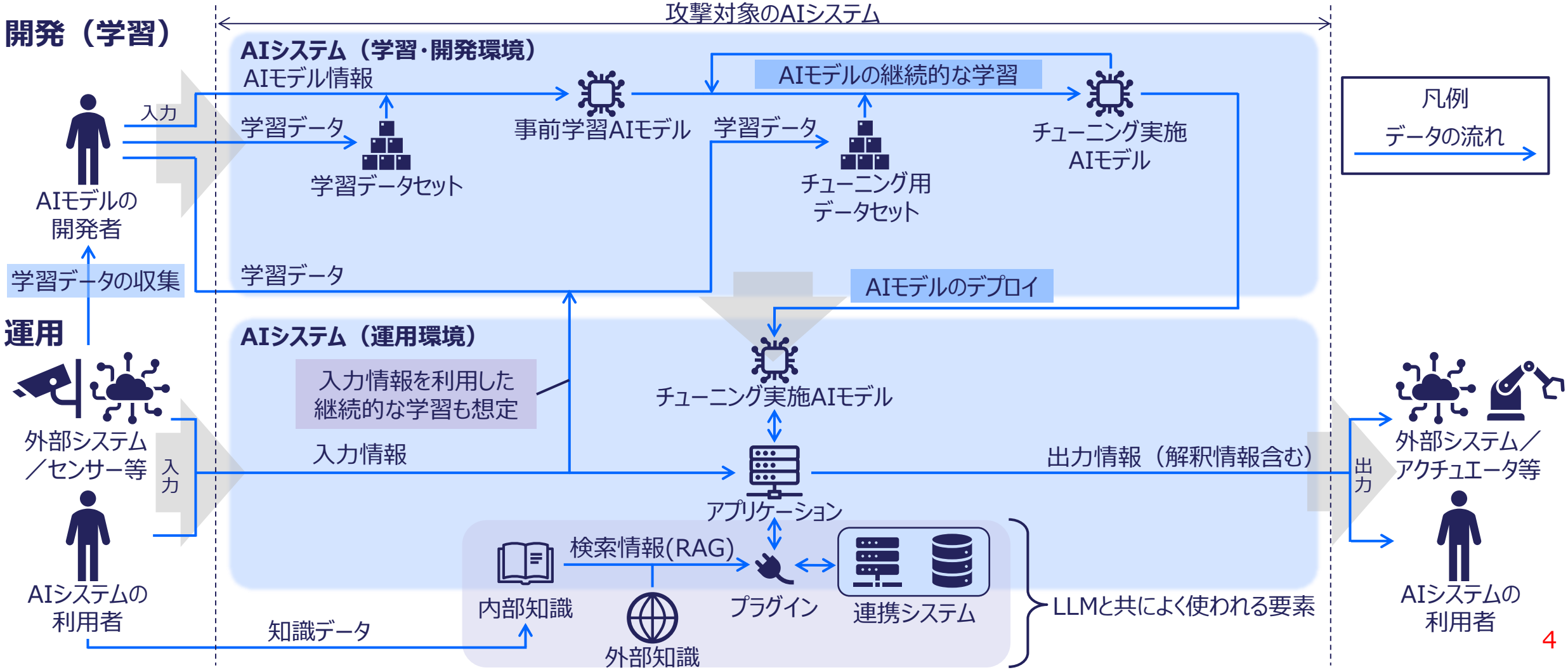
本資料のスコープ

AISI Japan AI Safety Institute

- 従来のサイバーセキュリティ攻撃への対策が必要であることは前提とする。
- 本資料では、既知の攻撃の中でも**システムがAIを含むことで留意すべき、AIモデルの学習もしくは推論の過程への介入や、AIモデルの入出力の利用・改ざんを伴う攻撃**を扱う。
  - 扱う例
    - 学習時：データポイズニング攻撃（学習の過程への介入）
    - 推論時：モデル抽出攻撃（AIモデルの出力の利用）
  - 扱わない例
    - ネットワーク機器の脆弱性を突いた不正アクセスによるAIモデルの窃取
      ※攻撃対象はAIだが、対象とするシステムが持つAIモデルの学習・推論への介入や入出力の利用・改ざんを伴う攻撃ではない

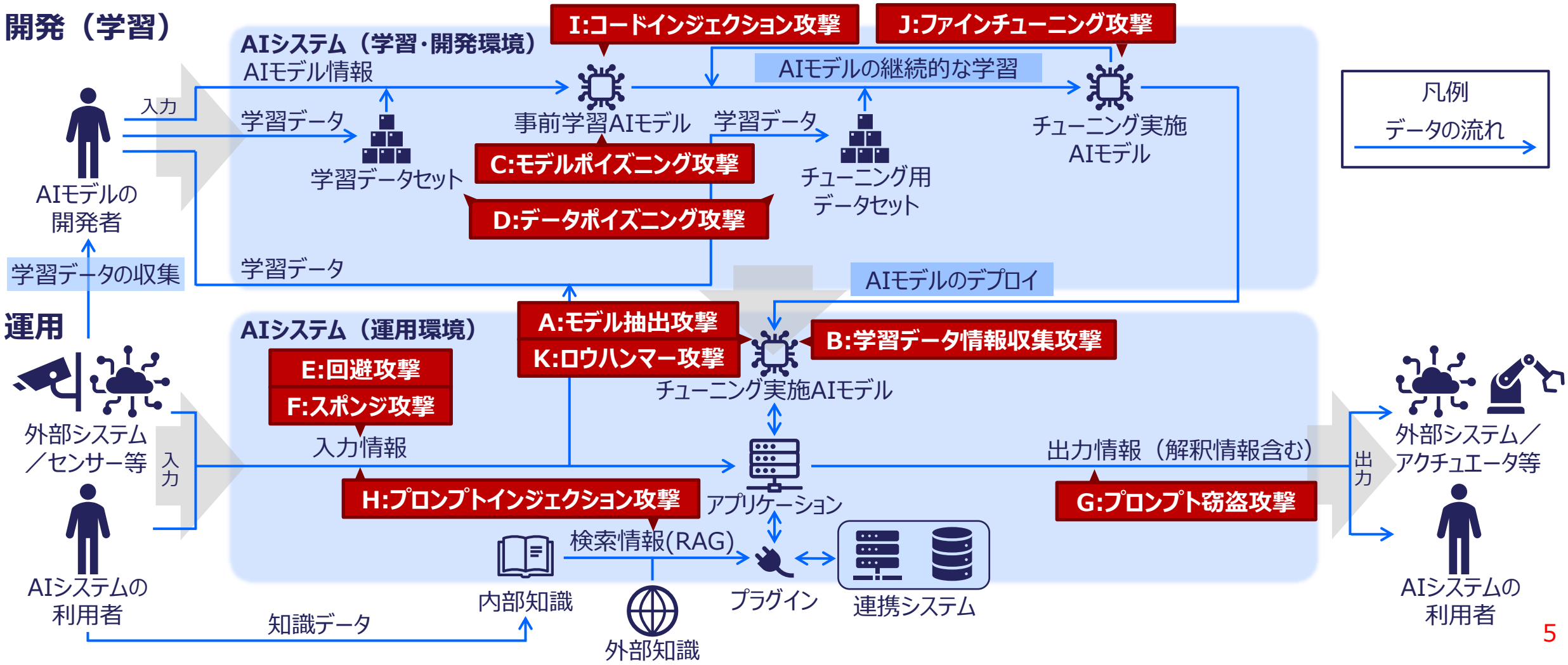
想定するAIシステム

- 開発（学習）と運用の２種類の環境で構成されるAIシステムを想定する。
- LLMと共によく使われるRAG等を明記したが、AIモデルはLLMに限定しない。



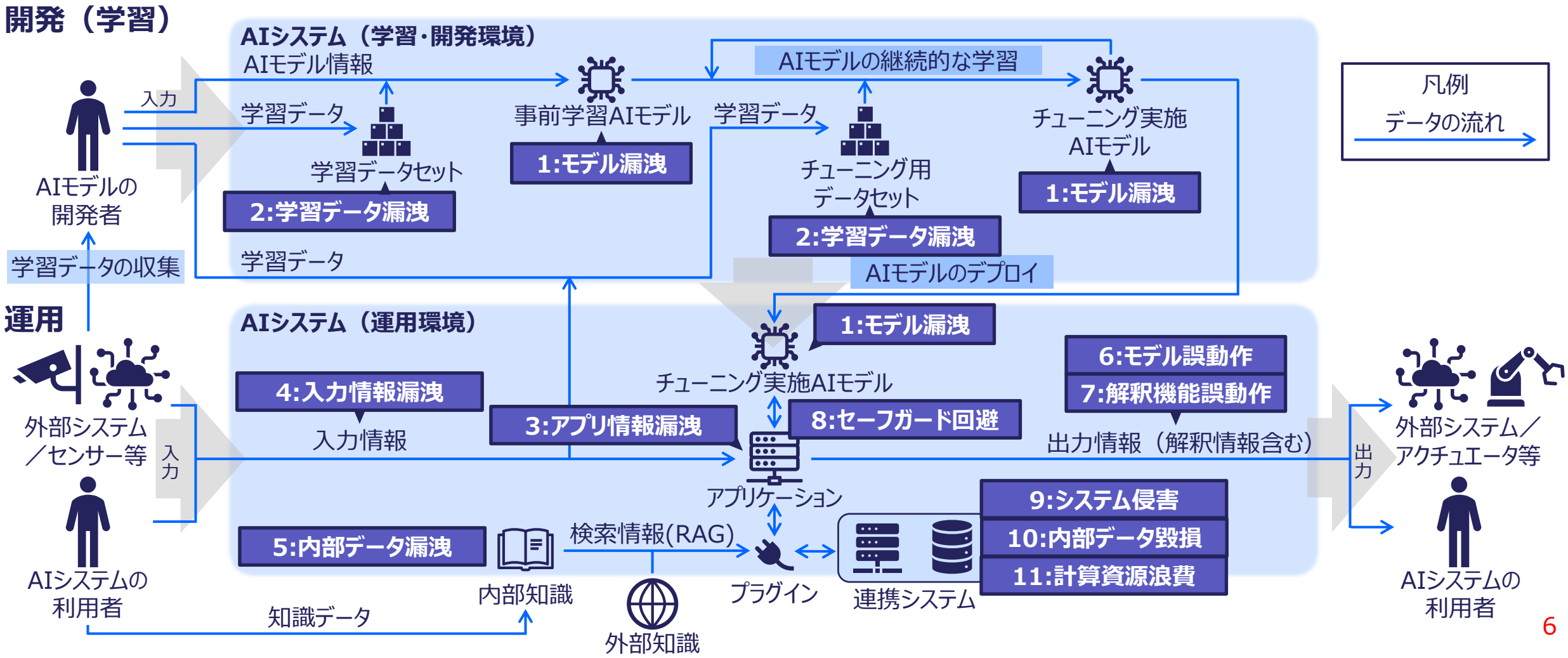
AIシステムに対する攻撃の俯瞰

- 想定するAIシステムへの攻撃（スコープは前述の通り）を下図に示す。
- 学習データ情報収集攻撃等の攻撃は、さらに複数の種類に分類できる。



AIシステムに対する攻撃による影響の俯瞰

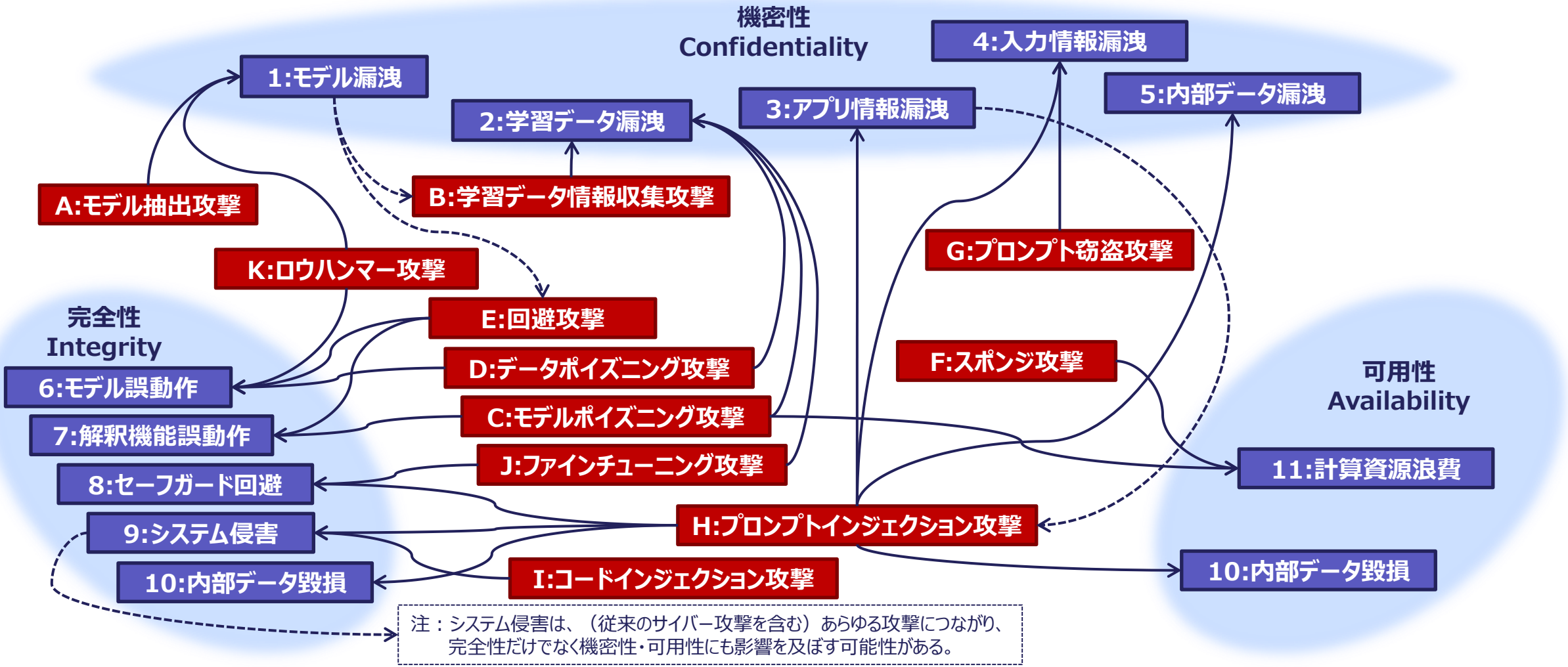
- 前スライドに示した攻撃による影響を下図に示す。
- モデル漏洩のように同じ種類の影響が複数の箇所に生じる場合もある。



AIシステムへの攻撃とその影響（概観）

AISI Japan AI Safety Institute

- 前スライドに示した攻撃とその影響の関係は下図の通り。
- 破線は矢印の先の攻撃に利用される可能性を示す。



AIシステムへの攻撃とその影響（詳細一覧）

AISI Japan AI Safety Institute

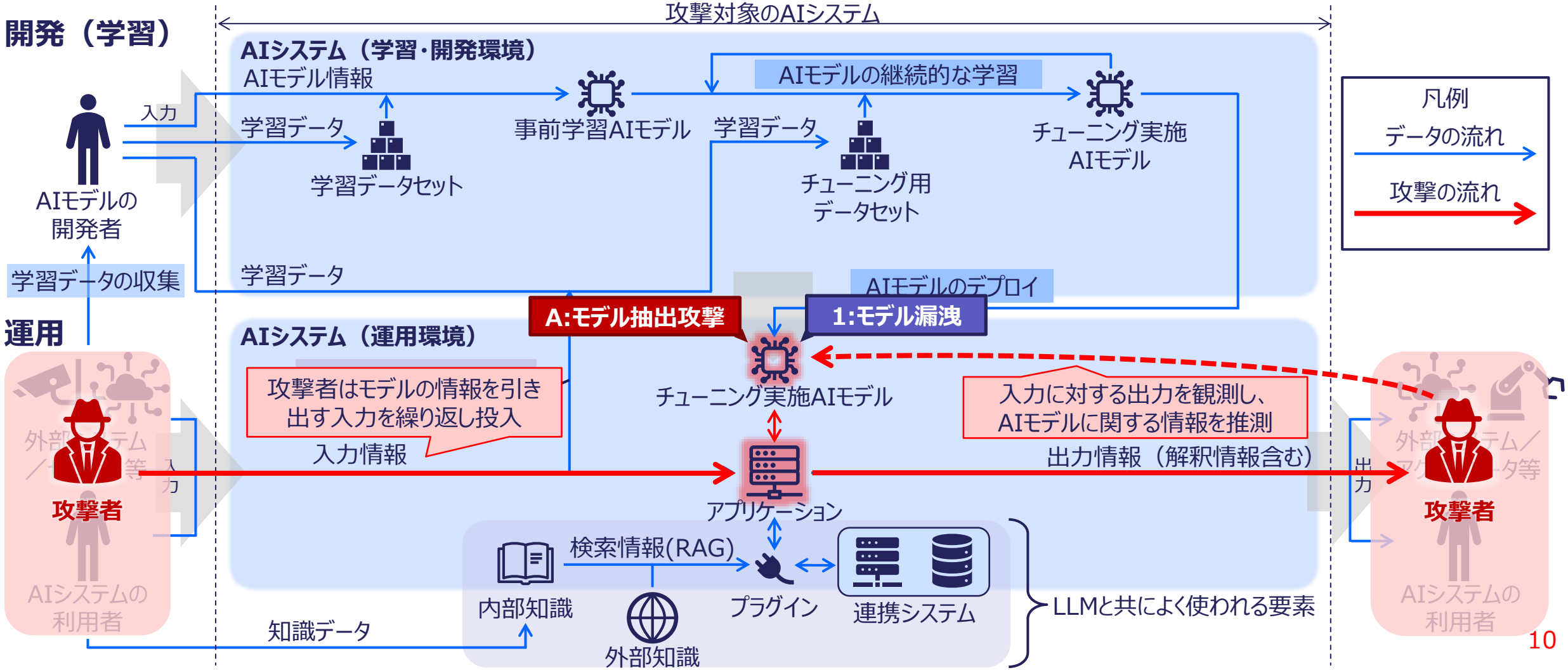
攻撃		影響	攻撃者の能力（攻撃前提）	データ形式：対象AI[文献]（文献からの要約でありデータ形式や対象を限定するものではない）
A モデル抽出		モデル漏洩	モデルへの大量のクエリー	画像：NN[OSF19a,OSF19b,JSM+19,CBB+18], LR[TZJ+16] 表：RR[WG18], DT[TZJ+16], LR, NN, SVM[TZJ+16,WG18]
B 学習データ情報収集	メンバーシップ推論	学習データ漏洩	モデルへの入力	画像・表：LR,DT[YGF+18], NN[AYM+19,YGF+18] 文章：NN[AYM+19]
			モデルへの大量のクエリー	画像・表：NN[SSS+17,LBW+18] 文章：LM[LSS+23]
			モデルの内部情報へのアクセス	画像・表：NN[NSH19]
	属性推論	学習データ漏洩	モデルへの入力	画像・表：LR, DT[YGF+18]
			モデルの内部情報へのアクセス	画像・文章：NN[SS20]
	プロパティ推論	学習データ漏洩	モデルの内部情報へのアクセス	画像・表：NN[GWY+18] 音声・通信：NN, SVM, HMM, DT[AMS+15]
	モデルインバージョン	学習データ漏洩	モデルへの大量のクエリー	画像・表：DT, NN[FJR15] 文章：Transformer[ZHK22]
	データ抽出	学習データ漏洩	モデルへの大量のクエリー	文章：LSTM, RNN[CLE+19], LM[CTW+21,LSS+23]
C モデルポイズニング		解釈機能誤動作	モデルの改変	表：NN, EM[SHJ+20]
		学習データ漏洩	学習プログラムの改変	画像：NN[SRS17] 文章：SVM, LR[SRS17]
		計算資源浪費	モデルの改変	画像：NN[CDB+23]
D データポイズニング		モデル誤動作	学習データの挿入	画像：NN[Car21] 文章：LM[Sch19]
		学習データ漏洩	学習データの挿入・モデルへの大量のクエリー	表：LR,NN[MGC22] 文章：LR[MGC22]
E 回避		モデル誤動作	モデルへの大量のクエリー	画像：NN, LR, SVM, DT, kNN[PMG+17] 文章：LM[BSA+22]
			モデルの内部情報へのアクセス	画像：NN[SZS+14,GSS15] 文章：LSTM[ERL+18]
			モデルへの入力	文章：Transformer[GSJ+21], LSTM, DA[WFK+19]
		解釈機能誤動作	モデルへの大量のクエリー	画像：NN[DAA+19]
F スポンジ		計算資源浪費	モデルへの入力	画像・文章：NN[SZB+21]
			モデルへの大量のクエリー	文章：LM[BSA+22]
G プロンプト窃盗		入力情報漏洩	モデルの出力へのアクセス	文章からの画像生成：Diffusion[SQB+24]
H プロンプトインジェクション		セーフガード回避	モデルへの入力	文章：LM[LDL+24,SCB+24,ZWC+23,WHS23,CRD+24,MZK+24,PHS+22,WFK+19]
			モデルの内部情報へのアクセス	文章・画像：NN[CNC+23]
		アプリ情報漏洩	システムAPI経由でのモデルへの入力	文章：LM[ZCI24]
		システム侵害	AIシステムが引用する情報の汚染	文章・画像からの文章生成：LM[GAM+23]
		内部データ漏洩/毀損	システムAPI経由でのモデルへの入力	文章から表の操作：LM[PEC+25]
		入力情報漏洩	ユーザーが引用する情報の汚染	文章：特定のチャットサービス[Sam23]
I コードインジェクション		システム侵害	悪性モデルの配布	任意：任意[ZWZ+24]
J ファインチューニング		セーフガード回避	モデルの内部情報へのアクセス	文章：LM[YYC+24]
		学習データ漏洩	ファインチューニング機能へのアクセス	文章：LM[CTZ+24]
K ロウハンマー		モデル誤動作	物理メモリへのアクセス	画像：NN[LWX+24]
			モデルの内部情報・物理メモリへのアクセス	画像・文章：LM, Transformer[NMF+24]
		モデル漏洩	物理メモリへのアクセス	画像：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

各攻撃の概要

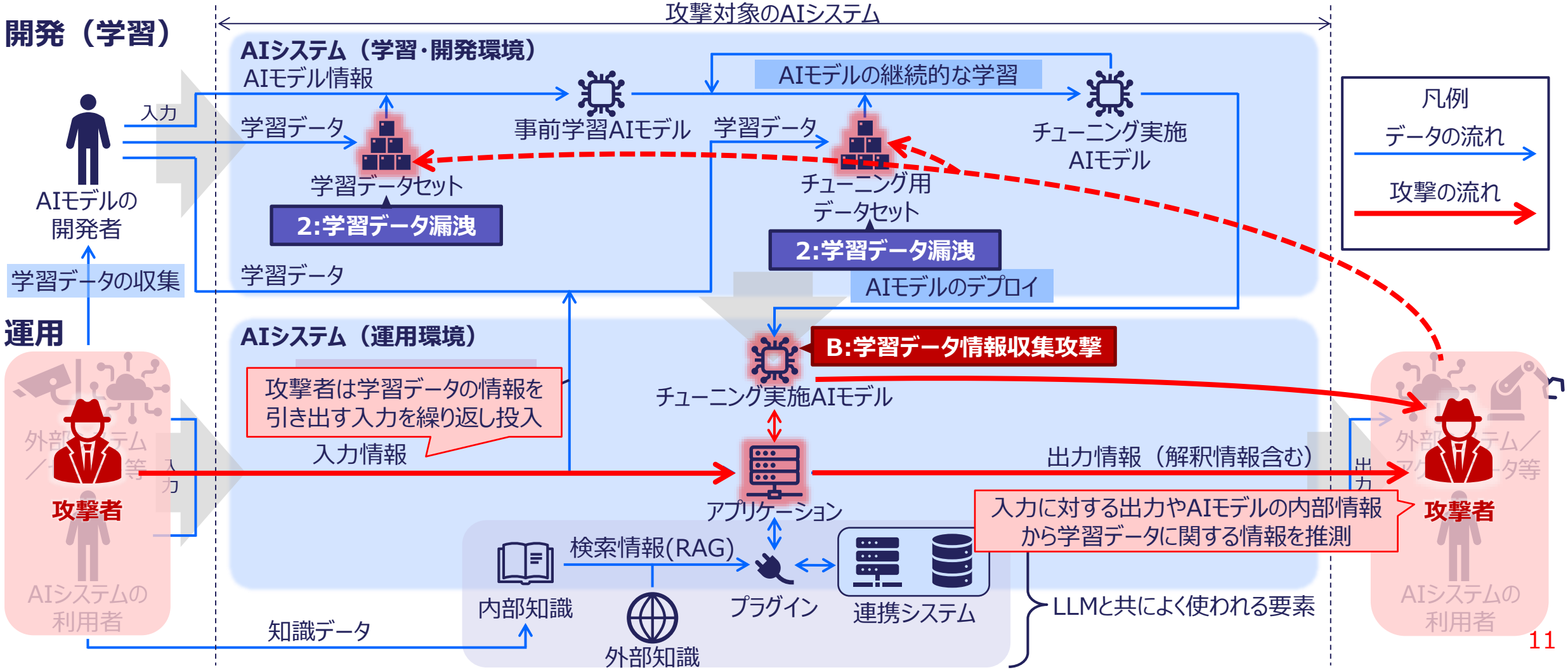
攻撃A:モデル抽出攻撃

- 入力情報に対する出力を観察することで、AIモデルに直接アクセスすることなく、AIモデルに関する情報の漏洩を引き起こす。



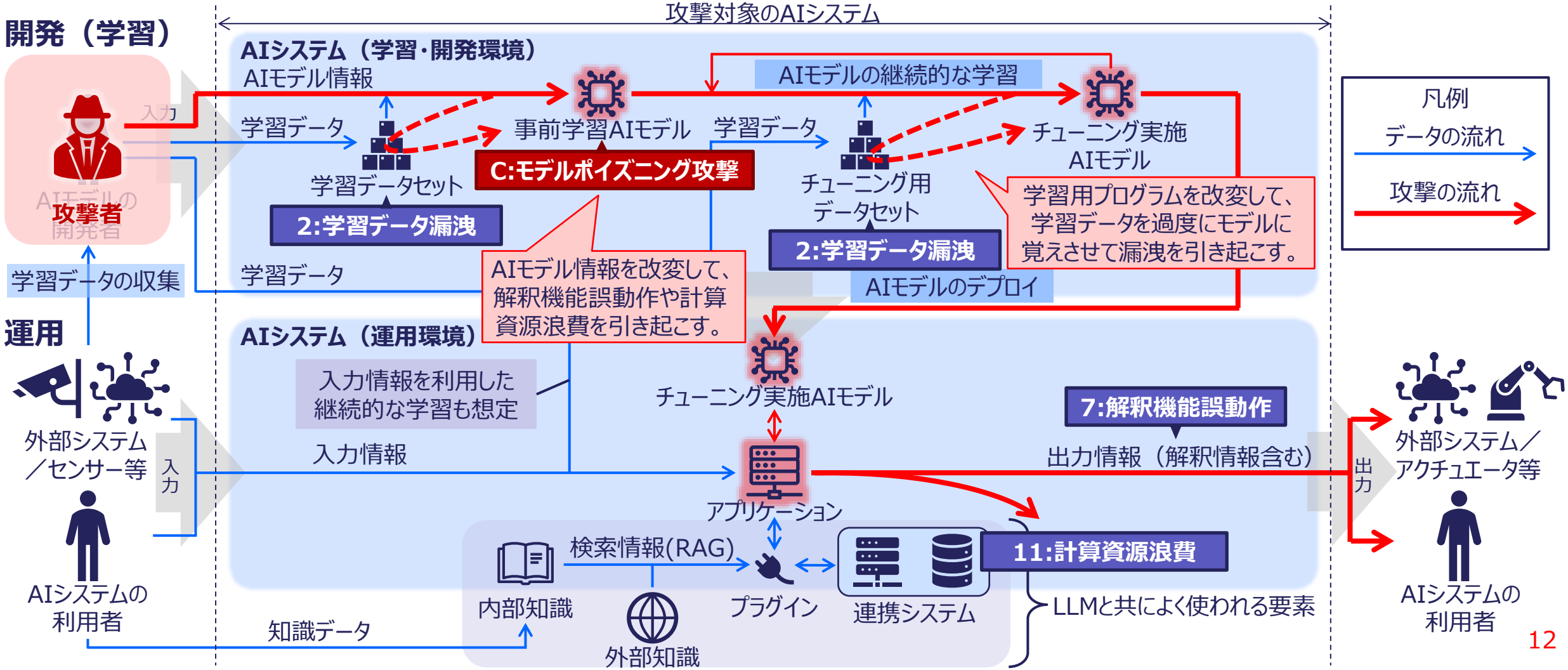
攻撃B:学習データ情報収集攻撃

- AIモデルの内部情報や入出力情報の対応から、学習データセットに関する情報を収集する。標的となる情報の種類により攻撃のバリエーションがある。



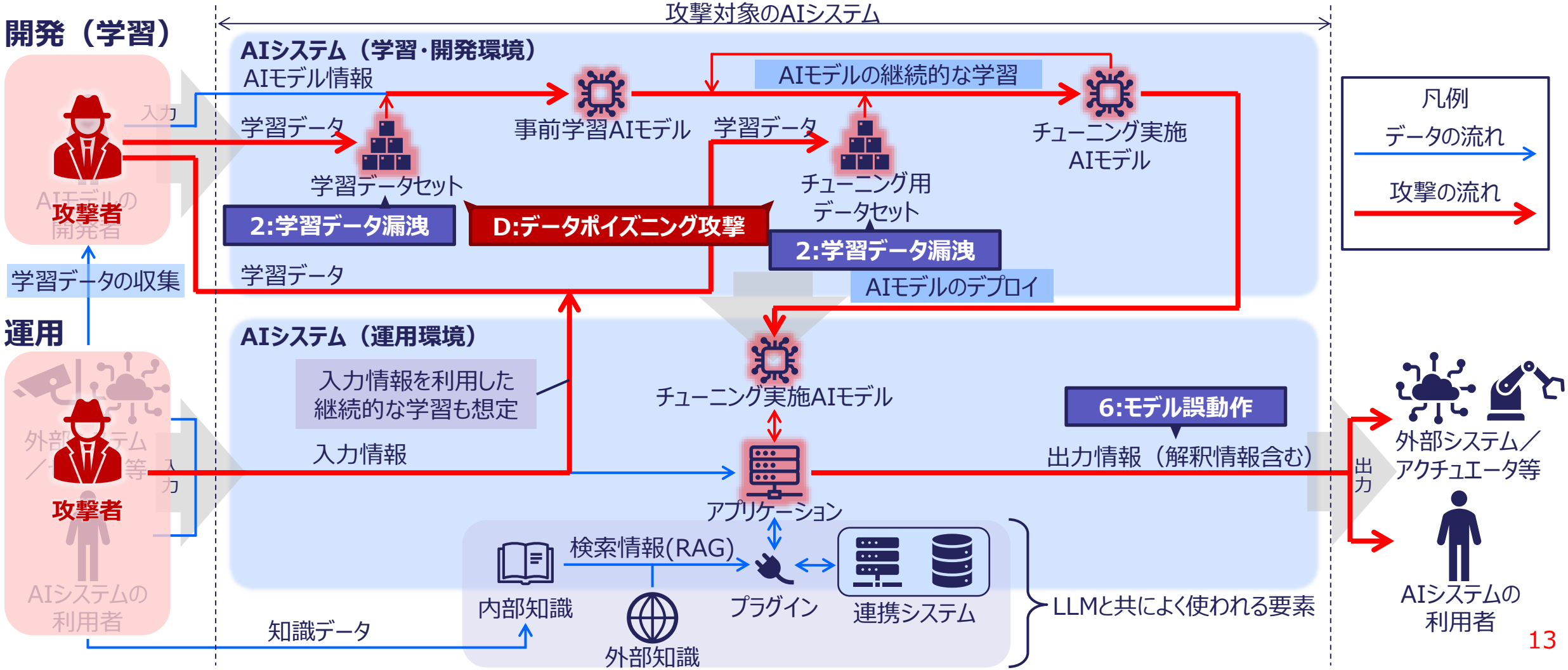
攻撃C:モデルポイズニング攻撃

- AIモデルの情報や学習用プログラムを改変することで、AIモデルの運用時に解釈機能誤動作や計算資源浪費、学習データ漏洩を引き起こす。



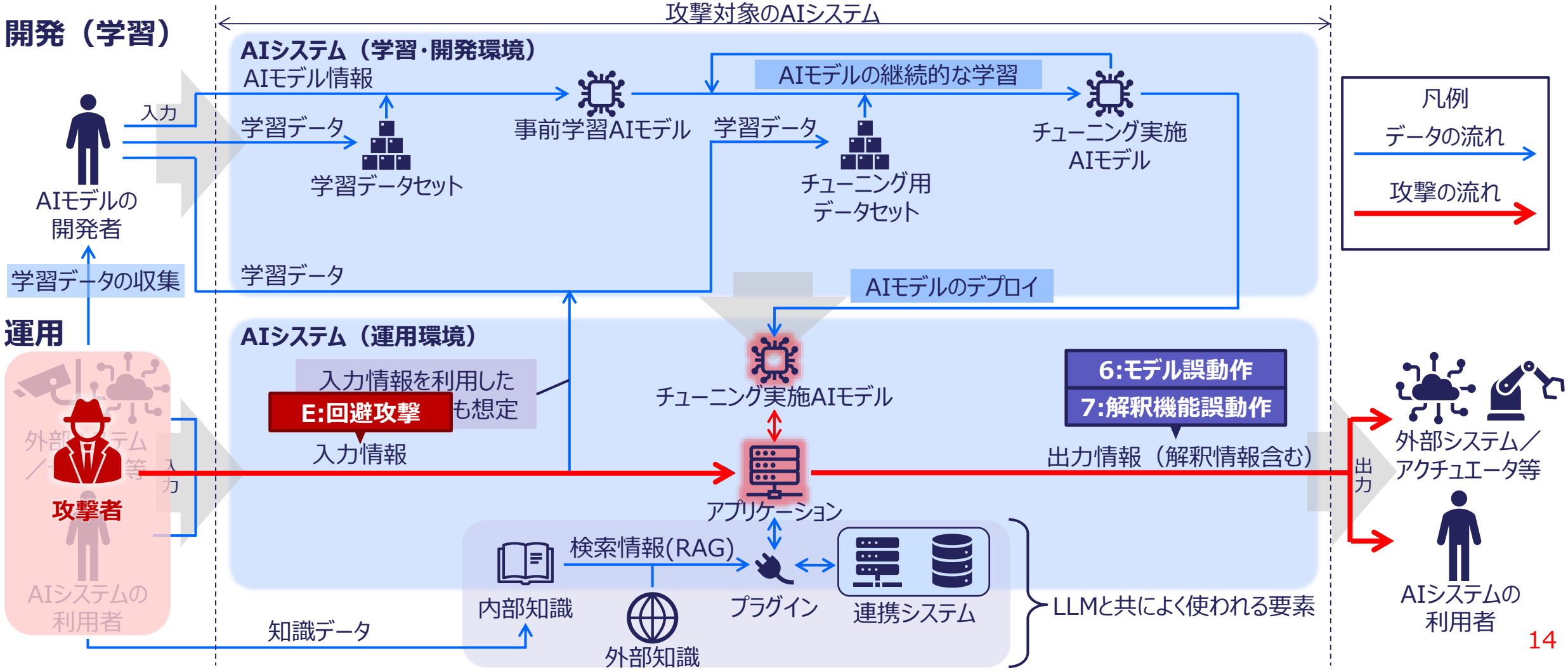
攻撃D:データポイズニング攻撃

- 学習データセットに特殊なデータを挿入することで、残りの学習データに関する情報の漏洩や運用時の入力情報に対するモデルの誤動作を引き起こす。



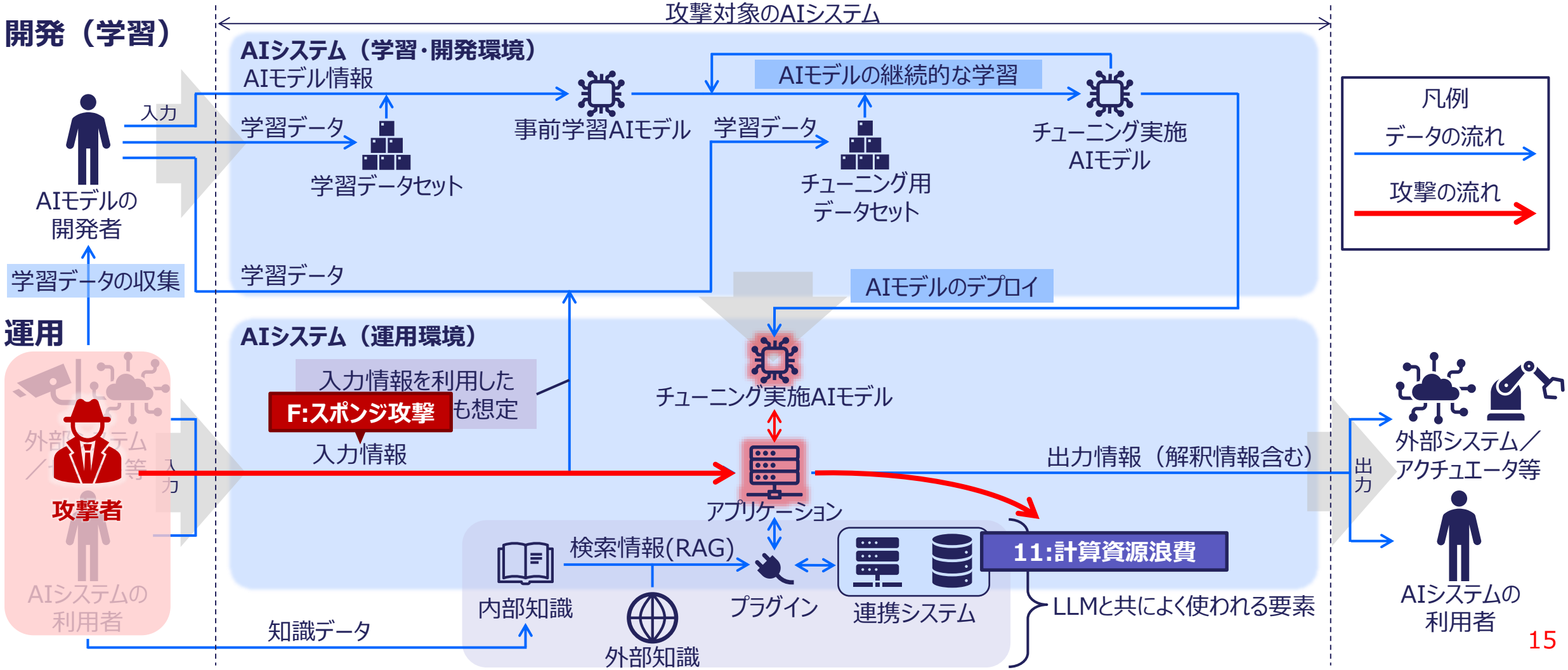
攻撃E:回避攻撃

- 運用時に敵対的サンプルと呼ばれる特殊な入力をAIシステムに与えることで、AIシステムの出力におけるモデル誤動作や解釈機能誤動作を引き起こす。



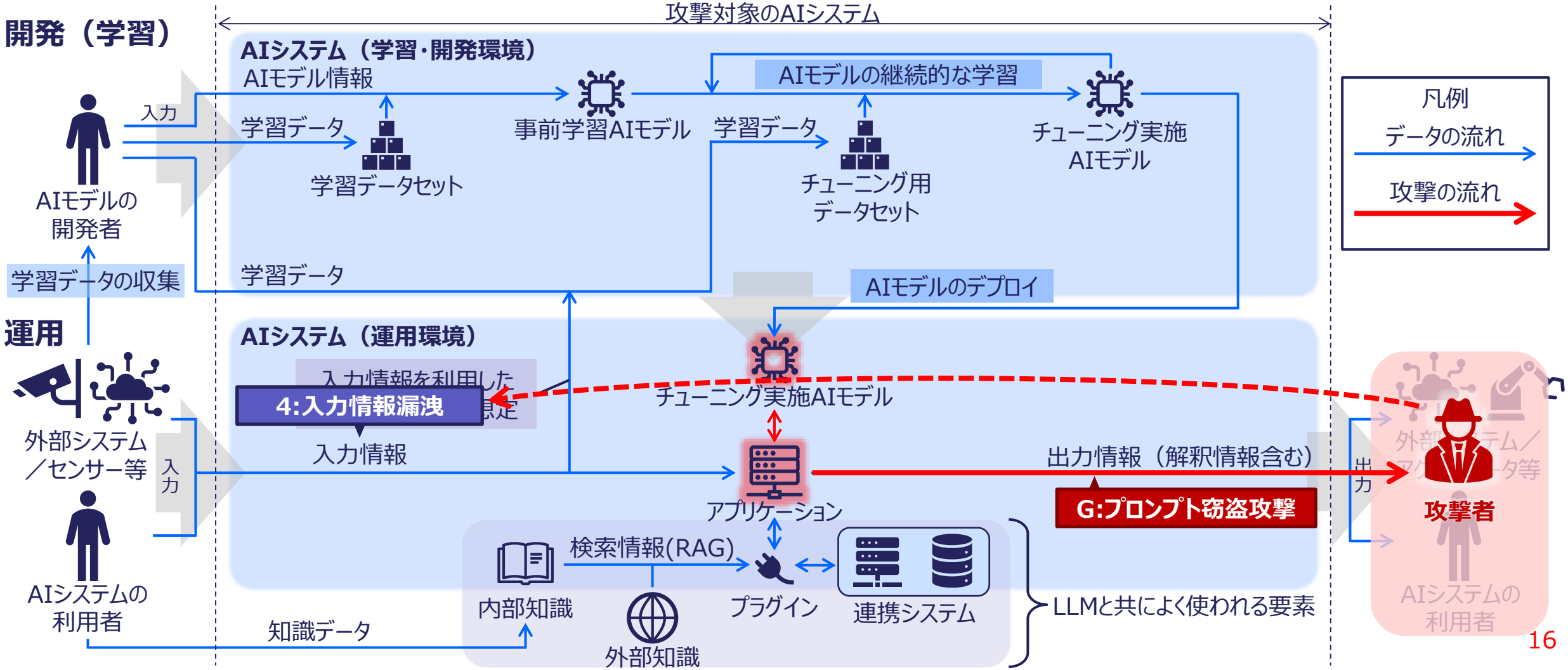
攻撃F:スポンジ攻撃

- 運用時にスポンジ・データと呼ばれる特殊な入力をAIシステムに与えることで、応答時間やエネルギー消費増大といった計算資源浪費を引き起こす。



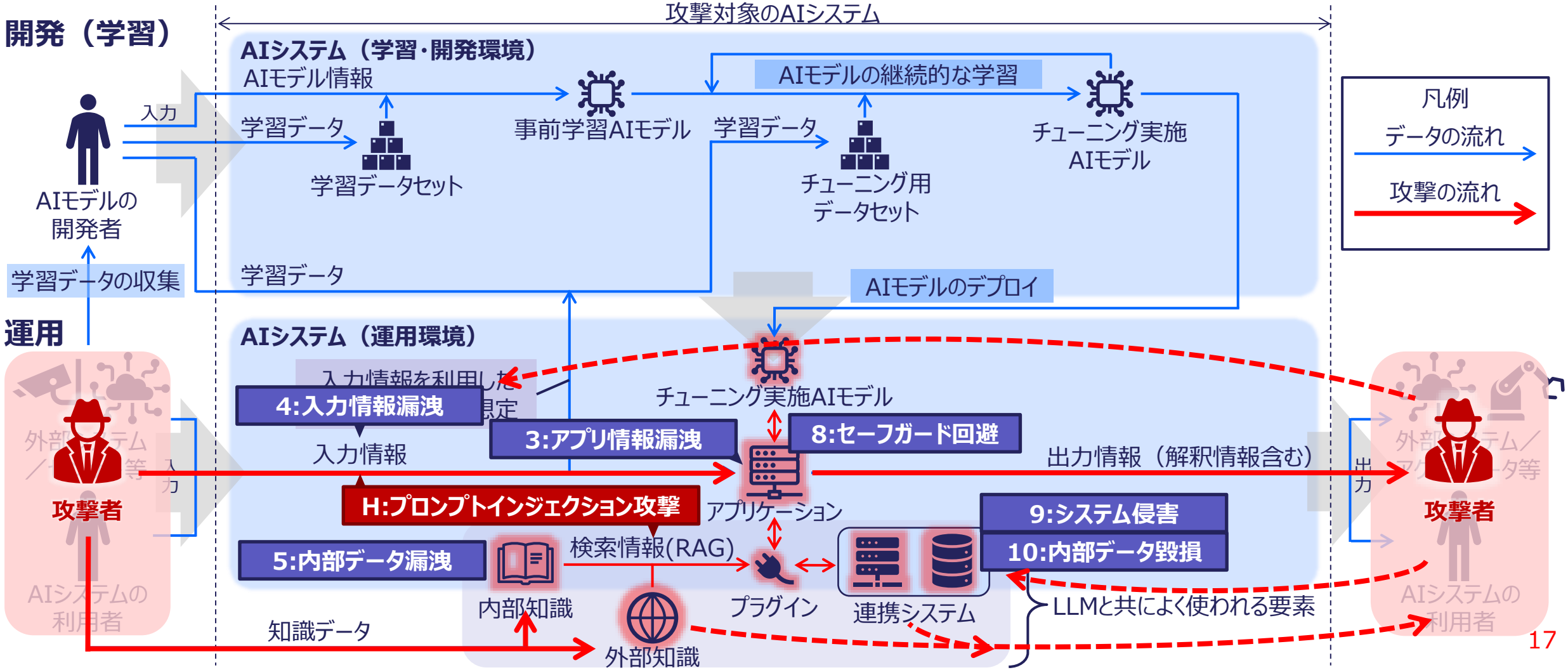
攻撃G:プロンプト窃盗攻撃

- AIモデルが生成した画像などの出力から、元となったプロンプトを推測することで、プロンプトエンジニアリングのノウハウとなる入力情報の漏洩を引き起こす。



攻撃H:プロンプトインジェクション攻撃

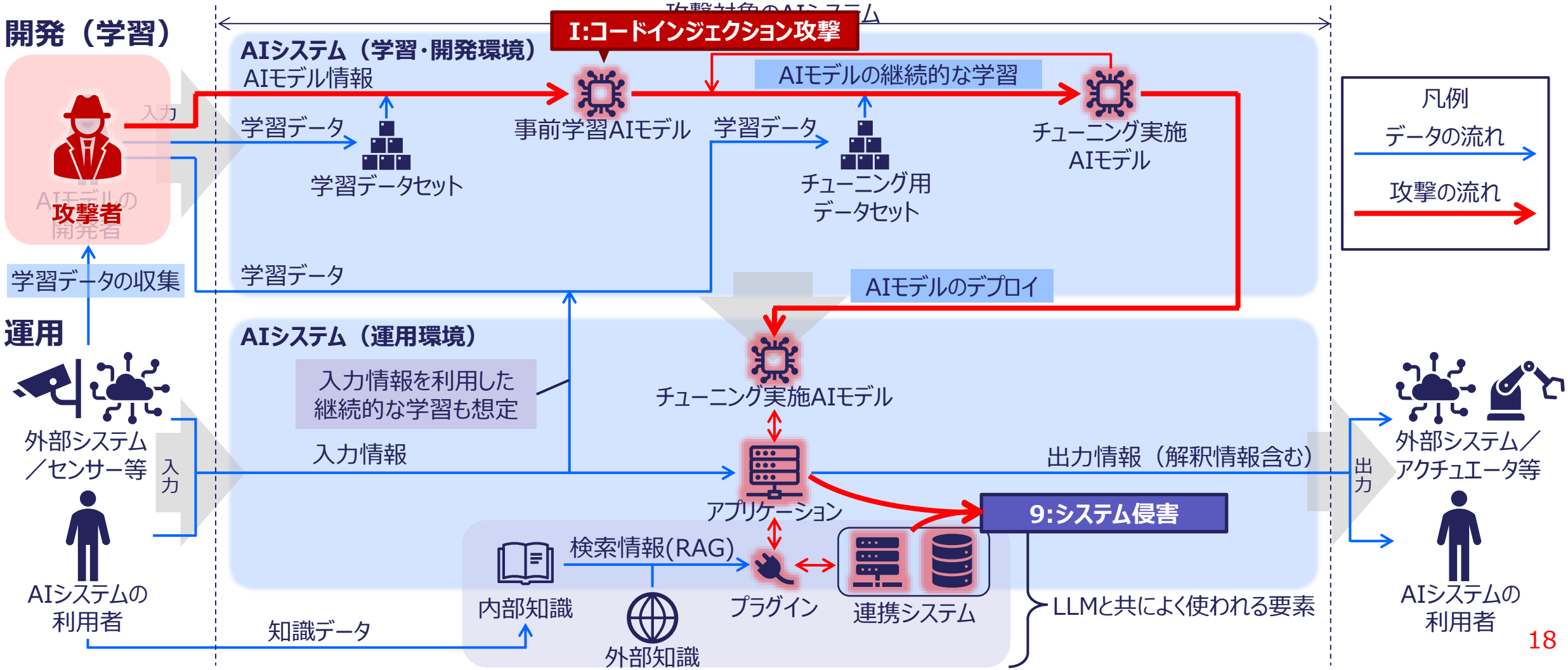
- モデルに直接、またはモデルが参照する情報源から間接的に、敵対的な指示を入力することで、各種情報漏洩やセーフガード回避、システム侵害を引き起こす。



攻撃I:コードインジェクション攻撃

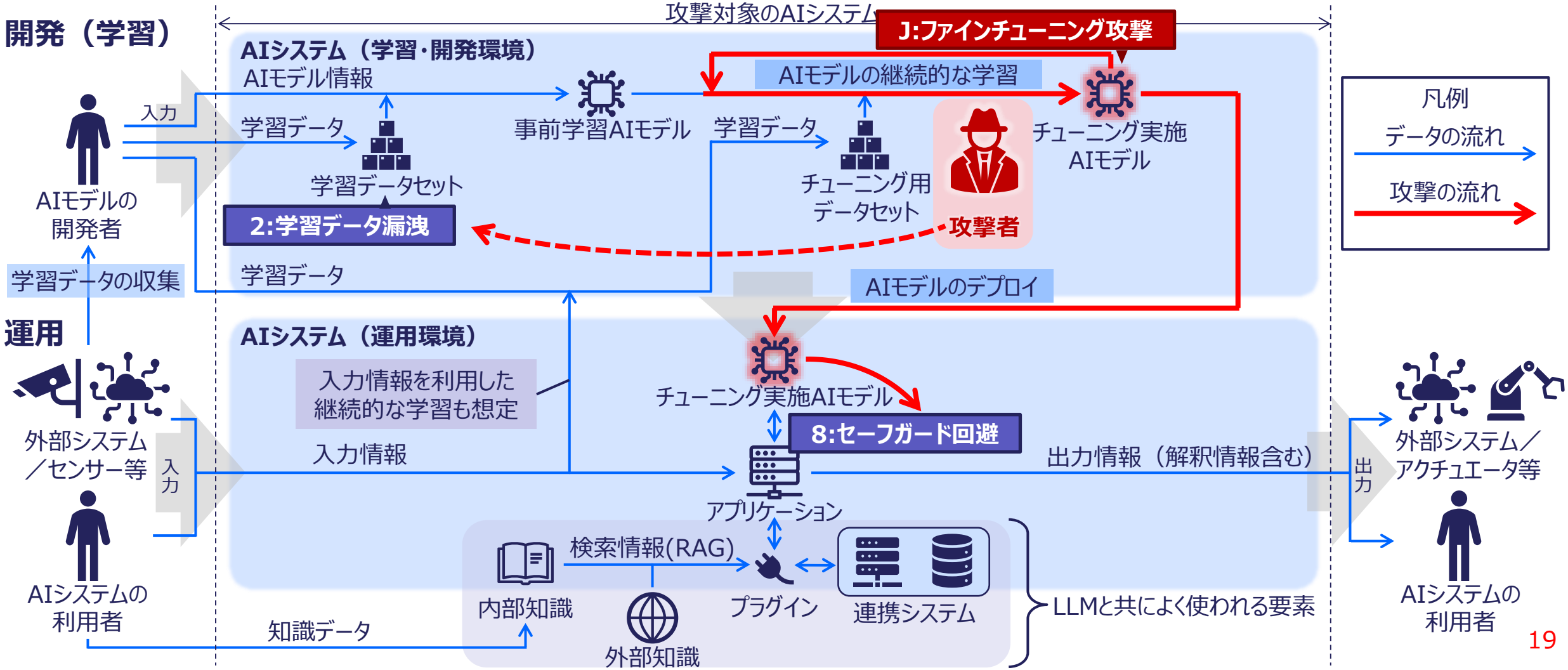
AISI Japan AI Safety Institute

- AIモデル自体に実行可能なコードを埋め込み、AIモデルを呼び出した際に埋め込んだコードが実行されるようにしておくことで、システム侵害を引き起こす。



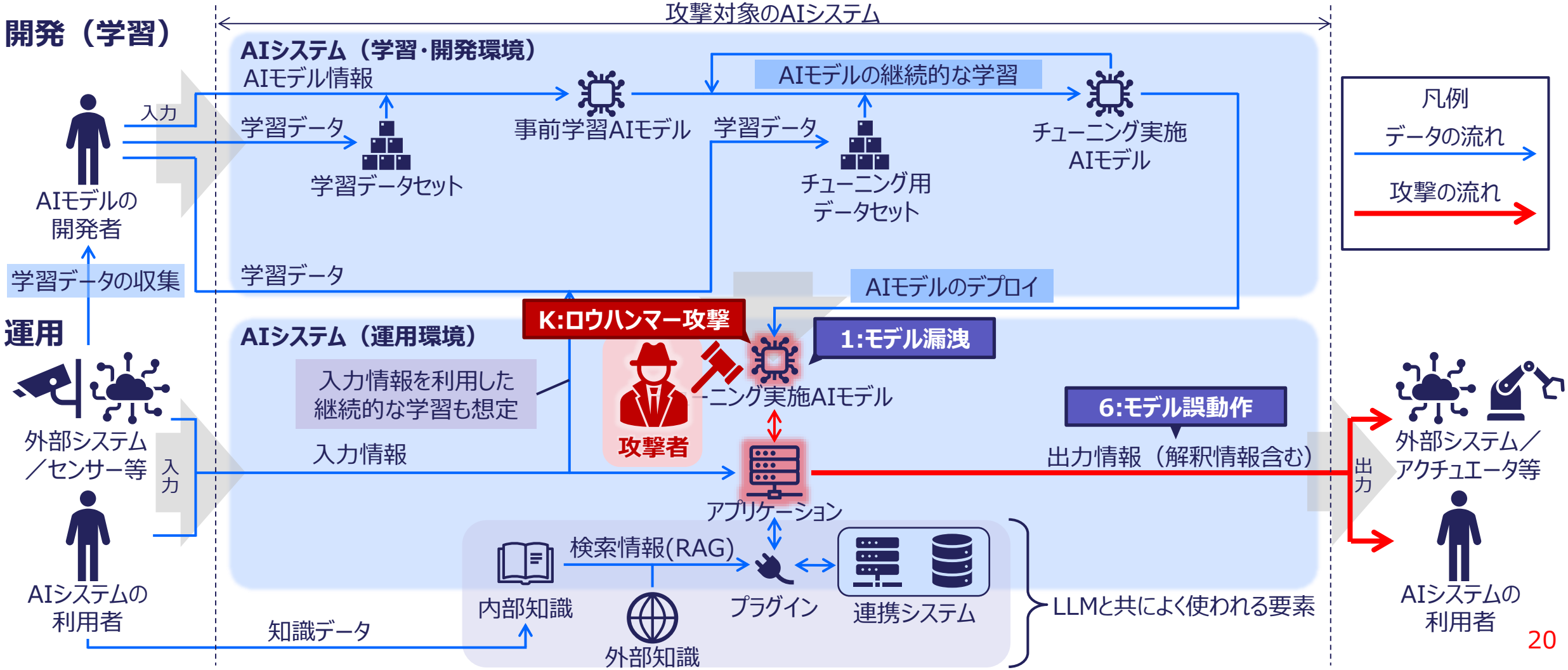
攻撃J:ファインチューニング攻撃

- 標的の事前学習AIモデルに対して特定のファインチューニングを実施することで、AIモデルがセーフガードを回避し、事前の学習データを漏洩するように仕向ける。



攻撃K:ロウハンマー攻撃

- 標的のAIモデルと物理メモリを共有する攻撃者が、AIモデルのメモリのビットをメモリセル間の干渉により反転させることでモデルの漏洩や誤動作を引き起こす。



参考文献 (1/4)

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.

参考文献 (2/4)

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.

参考文献 (3/4)

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.

参考文献 (4/4)

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

AIセーフティ・インスティテュート
『AIシステムに対する既知の攻撃と影響』（2025年3月）
https://aisi.go.jp/effort/effort_security/known_attacks_and_impacts/