The American Journal of Interdisciplinary Innovations and Research
ISSN 2642-7478 Volume 08 - 2026

Privacy-Preserving Processing of User Messages for LLM Services: Anonymization Methods, PII Leakage Assessment, and the "Confidentiality–Answer Quality" Trade-off

Andrei Shcherbinin

Social Discovery Group, Team Lead ML Engineer Tbilisi, Georgia

Received: 09 May 2026 | Received Revised Version: 26 May 2026 | Accepted: 12 June 2026 | Published: 25 June 2026

Volume 08 Issue 06 2026 | DOI: 10.37547/tajiir/Volume08Issue06-05

Abstract

This study provides a comprehensive analysis of architectural and algorithmic approaches aimed at maintaining data confidentiality during the operation of large-scale language models. Particular attention is paid to the identification and protection of personally identifiable information under conditions of continuous user interaction with intelligent systems. Evidence on security breaches from recent years is systematised, demonstrating a sharp increase in incidents associated with information leakage through generative models. The work examines in detail the hybrid RECAP methodology, combining deterministic algorithms with context-dependent prompts, as well as approaches grounded in differential privacy and machine "defocused" learning. The analysis further addresses the trade-off between the level of protection and the quality of generated answers, including the impact of anonymization on factual accuracy and on model capabilities, which are often described as cognitive. Based on the findings, recommendations are formulated for introducing adaptive routing strategies and multi-stage data cleansing into contemporary MLOps cycles.

Keywords: large language models, PII anonymization, differential privacy, machine learning, information security, ROC AUC, message anonymization, confidentiality–quality trade-off, RECAP, personal data protection.

Cite This Article: Shcherbinin, A. (2026). Privacy-Preserving Processing of User Messages for LLM Services: Anonymization Methods, PII Leakage Assessment, and the "Confidentiality–Answer Quality" Trade-off. The American Journal of Interdisciplinary Innovations and Research, 8(06), 80–87. https://doi.org/10.37547/tajiir/Volume08Issue06-05

Introduction

The 2024–2025 period was marked by a transition of generative artificial intelligence systems from experimental use to broad deployment within corporate architectures. This technological leap was accompanied by an unprecedented escalation of data protection risks. According to the Stanford AI Index Report 2025, the number of AI-related incidents increased by 56.4% within a single year, reaching 233 recorded cases in 2024 [1]. Such statistics indicate not merely quantitative growth but also a qualitative shift in the threat profile: the industry is gradually moving away from accidental algorithmic failures toward deliberate attacks designed to extract personal information [1, 2].

The economic consequences of leaks of personally identifiable information in the context of large language

model usage remain substantial. Despite a reduction in average global incident-remediation costs to USD 4.44 million in 2025, the United States recorded a historic maximum average cost per case—USD 10.22 million [3]. The situation is further intensified by tightening regulatory requirements, including the EU AI Act, alongside rising forensic analysis costs for complex AI systems [5, 16]. More than half of all recorded breaches in 2025 (53%) directly involved customer PII, including tax identifiers, address data, and biometric attributes [3, 4].

Public trust in corporations deploying AI continues to show a downward trajectory: from 50% in 2023 to 47% in 2024 [1]. Contemporary models exhibit the capacity to memorise and inadvertently reproduce sensitive information originating from training corpora, thereby amplifying leakage risk. Empirical findings indicate that adversaries can extract authentic personal data from closed models with probabilities as high as 48.9%, while the cost of obtaining a single PII record may be as low as USD 0.012 [6, 8]. Under such conditions, conventional perimeter defences appear insufficient, motivating the adoption of specialised real-time message-processing approaches [7, 10].

The aim of the study is the development and subsequent substantiation of a privacy-preserving approach to processing user messages in LLM services, combining PII anonymization, leakage probability assessment, and a quantitative analysis of the "confidentiality–answer quality" trade-off for integration into MLOps contours.

The working hypothesis is based on the assumption that hybrid multi-stage cleansing (deterministic templates + context-dependent LLM validation + consolidation), in combination with adaptive routing and, where necessary, unlearning, can materially reduce the risk of re-identification and PII extraction without statistically significant degradation of key utility metrics (e.g., ROC AUC and factuality) in applied scenarios.

Scientific novelty is concentrated in the proposal and conceptual justification of a unified framework for LLM services that jointly: (i) classifies and measures residual PII leakage risks (including indirect re-identification and "semantic laundering"); (ii) aligns these risks with answer-quality degradation (including a "factuality tax" under differential privacy); and (iii) defines an engineering-reproducible scheme for integrating these mechanisms into a secured gateway and an MLOps life cycle.

Materials and Methods

The empirical basis of the study is formed by a corpus of scientific publications, industry reports, and normative-methodological documents addressing the protection of personally identifiable information (PII) in the use of large language models (LLMs). The "materials" included: systematic surveys and primary works on PII leakage and data extraction from LLMs; practice-oriented guidance on PII protection (e.g., NIST recommendations); EU regulatory documents (including materials from the European Data Protection Board and the European Commission); and industry reports on incidents and breach costs (e.g., Verizon and IBM). Additionally, descriptions of engineering solutions and anonymization SDKs (e.g., Microsoft Presidio) were incorporated as representative examples of the applied privacy layer. This set of materials enabled alignment of academic results (methods and metrics) with regulatory requirements and with operational implementation practices in MLOps.

Source selection was both targeted and systematised. Included works met the following criteria: (1) an explicit connection to LLMs/generative models and PII risks; (2) the presence of a formalizable methodology (algorithm/protocol description) or measurable metrics (e.g., ESR, precision/recall/ROC AUC, factuality, privacy–utility); (3) reproducibility at the level of problem framing (PII definitions, attack scenarios, de-identification procedures); (4) relevance to the contemporary deployment context (priority to 2023–2026, while foundational standards and baseline definitions were permitted outside the window). Publications lacking a described methodology (purely position statements), duplicative preprints without added value, and materials that failed to distinguish leakage types (direct prompts vs. indirect re-identification/"semantic laundering") were excluded. To reduce bias introduced by "grey sources" (vendor reports), a balancing rule was applied: each industry thesis required scientific/regulatory support or an explicit limitation of applicability.

The core research strategy consisted of comparative analysis: all privacy-preserving processing methods were classified within a single taxonomy and compared

along consistent axes. Along the "mechanism" axis, the following were distinguished: deterministic approaches (templates/regular expressions/dictionaries), named entity recognition (NER) models, LLM-oriented anonymisation, hybrid schemes (including multi-stage pipelines of the RECAP class), differential privacy (DP) methods applied during training/fine-tuning, and the family of unlearning/forgetting methods. Along the "threat" axis, scenarios were fixed as: direct extraction (DirectQA), indirect re-identification via quasi-identifiers, attacks through paraphrasing/context manipulation, and exploitation of memorised training examples. Along the "engineering integration" axis, insertion points (gateway/preprocessing, postprocessing, training), latency requirements, policy controllability, and suitability for industrial MLOps governance were compared.

To standardise comparisons, an encoding scheme was applied: from each source, features were extracted (PII type, masking granularity, presence of contextual validation, support for ambiguity), evaluation parameters (datasets/benchmarks, attack protocols), and outcomes (privacy and utility metrics). The primary privacy metrics included Extraction Success Rate (ESR) and qualitative indicators of residual risk (especially for "semantically dense" PII categories). Utility was assessed using model-quality metrics under applied tasks, where ROC AUC served as the central metric for binary/multi-class message classification, while generative scenarios relied on factual accuracy/error frequency (accounting for the increase in "factuality tax" under DP described in the literature). Comparisons were conducted not as absolute numbers "in isolation," but as linked triples—"experimental conditions → metrics → applicability constraints"-to avoid invalid cross-study comparisons across divergent datasets and threat models.

Results and Discussion

A meaningful assessment of de-identification performance should be multidimensional and simultaneously reflect two tightly coupled criteria: first, the degree to which personal data are successfully concealed; second, the extent to which the model's applied utility—and the quality of the underlying task—are preserved. In production-grade message classification systems, one of the most widely used quality indicators remains the area under the ROC curve (ROC AUC), a pattern that is also consistent with real-world deployments of comparable solutions in operational environments [11, 12]. At the same time, utility measurement cannot be allowed to substitute for confidentiality evaluation: high predictive accuracy combined with the restoration of the original sensitive content effectively signals the failure of the de-identification regime.

In applied studies, the effectiveness of personal-data removal is often expressed via an extraction success rate, i.e., the share of cases in which an attacking procedure is able to recover redacted information. Analysis of the UnlearnPII test suite suggests that contemporary methods can fully block extraction in the "direct questioning" scenario, where the model is explicitly prompted for personal data. Nevertheless, residual vulnerability tends to persist for semantically rich categories of information that enable indirect re-identification through contextual inference, the triangulation of hints, and the use of background knowledge—even after explicit identifiers have been formally removed.

Methodologically, a correct de-identification evaluation implies separating threats into at least two classes: (i) "straightforward" extraction of removed fragments, and (ii) re-identification from a constellation of attributes. In the second class, quasi-identifiers (age, occupation, rare biographical events, geographic and temporal anchors) play a decisive role: individually, such attributes may not name a person, yet in combination they can sharply increase the probability of establishing identity. Accordingly, control procedures should include robustness tests against context-driven attacks that emulate the inference of personal data from plausible life narratives [9, 11].

Additional importance attaches to distinguishing de-identification goals in their legal sense: masking direct identifiers is not always equivalent to reaching a state in which the data subject is no longer "identifiable." Both European regulation (the General Data Protection Regulation) and the Russian approach to personal-data processing embed an evaluative notion of identifiability: risks are expected to be calibrated against "reasonably likely" means of identification, given available technologies and associated costs. Consequently, a technical extraction metric is more appropriately interpreted through the lens of residual risk rather than as a purely formal "removed/not removed" statement.

Finally, comparing confidentiality and utility requires a

transparent description of the trade-off between protection and task quality: stronger de-identification predictably entails the loss of some information content, especially in free-text processing, where meaningful signals are often interwoven with personal details. A practically defensible reporting format is a combined set of indicators: task-quality metrics (including ROC AUC) alongside residual disclosure metrics, complemented by an analysis of error types (which categories remain vulnerable, and why). Such a presentation supports not only method-to-method comparison, but also a reasoned argument about the sufficiency of the adopted controls for a specific data-processing scenario, where the "cost of error" is determined by both the likelihood and the severity of re-identification consequences [11, 20].

Table 1 provides a detailed description of the vulnerability of PII categories to indirect re-identification after anonymization.

Table 1. Vulnerability of PII categories to indirect re-identification after anonymization (compiled by the author based on [18]).

PII category	Residual leakage risk (ESR)	Degree of resistance to forgetting
First/last name	0.0% – 1.2%	High
Profession/place of employment	9.0%	Low
Diseases/diagnoses	6.7%	Medium
Document identifiers (TIN/SSN)	5.3%	High

A simple substitution of names with placeholder symbols does not, by itself, ensure adequate concealment of identity when the text retains indirect cues that enable reconstruction of the data subject. Such cues include, in particular, a uniquely recognisable job function within a narrow professional community, a combination of rare biographical circumstances, and information about an uncommon disease. In these situations, the de-identified text can remain analytically usable while preserving latent semantic linkages; when cross-referenced with external sources or exploited through contextual inference, these linkages create a basis for re-identification. The literature describes this phenomenon as "semantic laundering" of meanings [21, 22].

Operational practice at Social Discovery Group indicates that message de-identification does not necessarily degrade classification performance or user-perceived service quality. With an ROC AUC of >0.95, the system reliably distinguishes user intents even when names, addresses, and contact details are fully replaced with semantically unambiguous markers. In this setting, a high ROC AUC value implies stable class separation (e.g., "refund request" versus "service-quality complaint") at a low false-positive rate, which is especially important for ticket routing and queue prioritisation.

At the same time, research points to a so-called "price of factual accuracy" associated with differential privacy methods. The probability of factual errors and incorrect statements increases by 17–24% [23], typically attributed to the effect of privacy-induced noise on the formation and consolidation of associative links during fine-tuning. As a result, a model may preserve surface coherence while more frequently distorting details—dates, numbers, and causal relationships—which is consequential in applied scenarios that require heightened reliability.

Embedding de-identification systems into production pipelines for large language model–based services should account not only for current quality indicators but also for longer-run operational consequences, including ethical and organisational aspects of proportionality between confidentiality and functionality. In particular, quasi-identifiers play a significant role in re-identification risk assessment: attributes that are not direct identifiers yet, in combination, sharply increase the likelihood of establishing identity. Therefore, verification of de-identification effectiveness should include scenarios of contextual inference and cross-fragment linkage, rather than being limited to tests of direct extraction of removed fields.

From a practical standpoint, multidimensional validation is warranted, in which utility metrics (including ROC AUC) are assessed alongside residual disclosure indicators and a structured analysis of vulnerable information categories. Additional error-type control is also required: which specific cues persist after processing, in which message genres they occur more often, and which mechanisms enable identity recovery—verbatim reproduction, context-driven "guessing," or reconstruction from a constellation of rare characteristics. This approach reduces the risk of false assurances of safety based on a single indicator and provides an evidentiary basis for engineering and managerial decision-making.

Finally, correct integration of de-identification presupposes formalised risk governance procedures: documentation of the threat model, definition of acceptable residual-risk thresholds, regulation of access to raw data, and systematic monitoring of quality degradation over time. When the domain, user composition, or ticket topics shift, the structure of quasi-identifiers and the probability of re-identification also change, requiring periodic revision of substitution rules, marker vocabularies, and test scenarios. In this way, protection is maintained not as a formal "removal" of selected fields, but as a stable operating regime under model updates and the evolving practices of adversarial analysis.

A conceptual diagram of the real-time anonymization life cycle will be presented below in Figure 1.

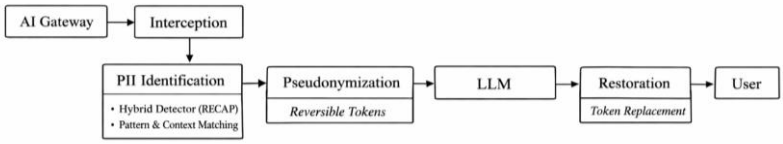

Fig.1. Conceptual diagram of the real-time anonymization life cycle (compiled by the author based on [9, 13, 14]).

Figure 2. "Confidentiality–Utility" trade-off curve for privacy-preserving message processing (compiled by the author based on [9, 17]).

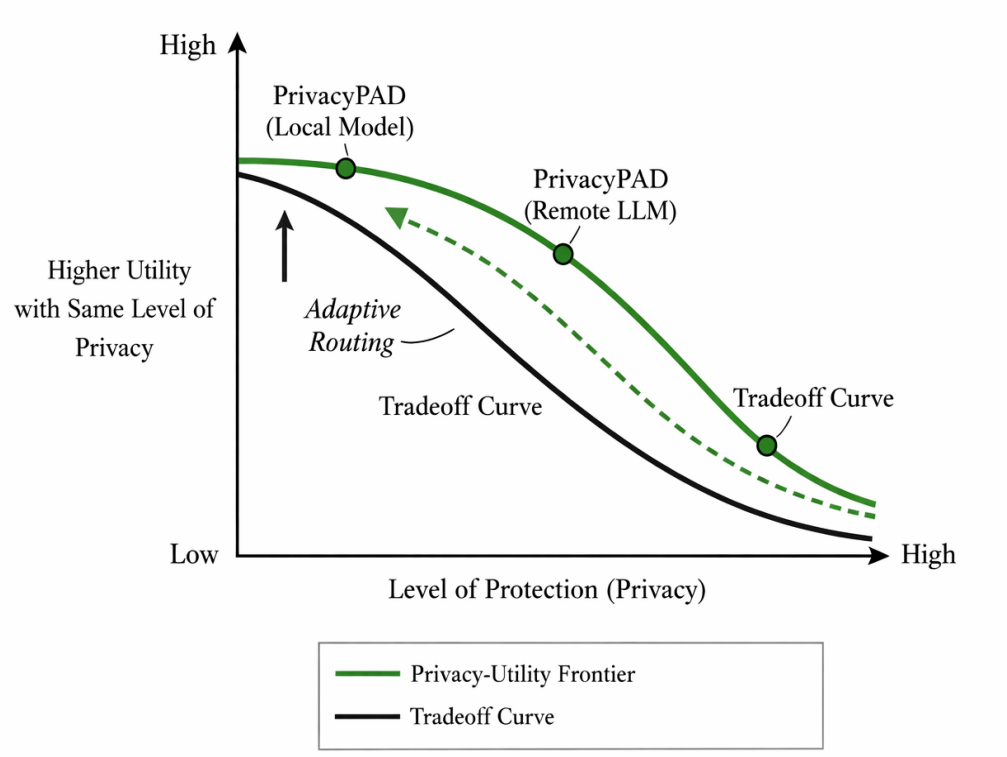

Fig.2. "Confidentiality–Utility" trade-off curve (compiled by the author based on [9, 17, 19, 24]).

High values of the area under the ROC curve matter not only as a classification-quality indicator, but also as an element of the overall abuse resilience of deployed systems. In countering prompt-injection attacks and task drift, the classifier's ability to reliably separate "clean" from manipulated text fragments at ROC AUC > 0.99 becomes a genuinely critical defensive boundary [15, 25]. Achieving this level of separability enables automated suppression of attempts to bypass de-identification mechanisms—most notably in cases where an adversary tries to trigger personal-data extraction through contextual manipulation and reframing of the request.

A correct problem formulation for detecting such impacts assumes that monitoring is not limited to superficial features of "suspiciousness" in the input, but includes functional signs of deviation from the system's permitted behaviour. Practically, this means detecting deliberate instruction insertion aimed at disabling safeguards, coercing the model into disclosing original information, or reconstructing identifying data through chains of probing follow-up questions. In that setting, ROC AUC does not represent an abstract notion of "accuracy." It captures the capacity to separate normal and adversarial patterns across a broad range of thresholds, which is central when choosing a response regime that simultaneously minimises missed attacks and unwarranted blocks of legitimate requests.

The reliability of a barrier classifier also directly affects the controllability of protection loops in production. When stable class separation is available, a multi-tier architecture becomes feasible: an initial filter flags prompt-injection and task-drift signals; then stronger de-identification modes are activated, context is shortened, permissible transformations are restricted, and logging is enabled for subsequent analysis. This arrangement increases the reproducibility of decisions and strengthens the evidentiary character of applied controls, because each trigger can be mapped back to formalised criteria rather than an informal "sense of risk."

At the same time, reliance on threshold-oriented ROC AUC targets should not replace a system-level security evaluation. Preventing de-identification bypass requires regularly refreshing adversarial example sets, testing against adaptive scenarios, and careful control of Type II errors in the operational sense—where excessive sensitivity leads to blocking good-faith messages and degrading service quality. ROC AUC, therefore, functions as a necessary but insufficient condition: it should be complemented by domain-shift robustness analysis, pressure-testing under context manipulation, and an estimate of residual re-identification risk where semantic linkages remain exploitable even after direct identifiers have been formally removed.

Conclusion

An analysis of privacy-preserving message-processing mechanisms for LLM services in 2024–2025 confirms that safeguarding data is a multidimensional engineering problem. The shift from static filters toward hybrid systems such as RECAP, together with the introduction of adaptive routing, enables organisations to reduce the risk of PII leakage while maintaining operational efficiency.

Achieving a level of 0.95 and higher indicates that the semantic transparency of text can be retained even under strict confidentiality requirements. At the same time, the identified "hallucinatory effect" of differential privacy—manifested as a 17–24% reduction in factuality—suggests that DP should be applied cautiously in scenarios requiring high-precision knowledge extraction.

Further industry evolution is likely to be associated with improvements in Shadow Unlearning and with the development of specialised small language models (SLMs) trained via adversarial distillation (SEAL). Such models can perform anonymisation locally, eliminating any transmission of sensitive data through external APIs. Under tightening regulations, including the EU AI Act, the deployment of multi-layer protection systems is expected to become a mandatory standard for building trustworthy artificial intelligence.

References

1. AI Index Steering Committee. (2025). AI Index Report 2025 | Stanford Institute for Human-Centered Artificial Intelligence (Stanford HAI). Retrieved from: https://aiindex.stanford.edu/report/ (date accessed: October 3, 2025).
2. Cheng, S., Li, Z., Meng, S., Ren, M., Xu, H., Hao, S., Yue, C., & Zhang, F. (2025). Understanding PII leakage in large language models: A systematic survey. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25) (pp. 10409–10417). https://doi.org/10.24963/ijcai.2025/1156
3. Verizon. (2025). Data Breach Investigations Report

(DBIR) 2025 | Verizon Business. Retrieved from: https://www.verizon.com/business/resources/reports/dbir/ (date accessed: October 6, 2025).

4. Cost of a Data Breach Report 2025: The AI Oversight Gap | Baker Donelson. (2025). Retrieved from: https://www.bakerdonelson.com/webfiles/Publications/20250822_Cost-of-a-Data-Breach-Report-2025.pdf (date accessed: October 9, 2025).
5. Regulation (EU) 2024/1689 (Artificial Intelligence Act) | EUR-Lex. (2024). Retrieved from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (date accessed: October 12, 2025).
6. AI Act | Shaping Europe's digital future | European Commission. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (date accessed: October 15, 2025).
7. IBM Report: 13% Of Organizations Reported Breaches Of AI Models Or Applications, 97% Of Which Reported Lacking Proper AI Access Controls | IBM Newsroom. (2025). Retrieved from: https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications%2C-97-of-which-reported-lacking-proper-ai-access-controls (date accessed: October 18, 2025).
8. Cheng, S., Meng, S., Xu, H., Zhang, H., Hao, S., Yue, C., Ma, W., Han, M., Zhang, F., & Li, Z. (2025). Effective PII extraction from LLMs through augmented few-shot learning. In Proceedings of the 34th USENIX Security Symposium (USENIX Security 25) (pp. 8155–8173). https://doi.org/10.5555/3766078.3766496
9. AI Privacy Risks & Mitigations – Large Language Models (LLMs) | European Data Protection Board (EDPB).(2025). Retrieved from: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf(date accessed: October 21, 2025).
10. Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models. In Proceedings of the 2023 IEEE Symposium on Security and Privacy (pp. 346–363). https://doi.org/10.1109/SP46215.2023.00028
11. Rajgarhia, H., Gupta, S., Shaik, A., Kumar, G. P., Santhoshraj, Y., Nishitha, S. N. T., & Mukherji, A. (2025). An evaluation study of hybrid methods for multilingual PII detection. arXiv. https://doi.org/10.48550/arXiv.2510.07551
12. Manzanares-Salor, B., & Sánchez, D. (2025). A comparative analysis, enhancement and evaluation of text anonymization with pre-trained large language models. Expert Systems with Applications, 297, 129474. https://doi.org/10.1016/j.eswa.2025.129474
13. McCallister, E., Grance, T., & Scarfone, K. (2010). Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) (NIST SP 800-122) | NIST. Retrieved from: https://csrc.nist.gov/pubs/sp/800/122/final (date accessed: November 3, 2025).
14. Presidio: Data Protection and De-identification SDK | Microsoft. Retrieved from: https://microsoft.github.io/presidio/text_anonymization/ (date accessed: November 7, 2025).
15. Edemacu, K., & Wu, X. (2025). Privacy preserving prompt engineering: A survey. ACM Computing Surveys, 57(10). https://doi.org/10.1145/3729219
16. Ji, W., & Ying, Z. (2026). An LLM-powered framework for privacy-preserving and scalable labor market analysis. Mathematics, 14(1), 53. https://doi.org/10.3390/math14010053
17. Abbasi, W., Mori, P., & Saracino, A. (2025). Trading-off privacy, utility, and explainability in deep learning-based image data analysis. IEEE Transactions on Dependable and Secure Computing. https://doi.org/10.1109/TDSC.2024.3400608
18. Parii, D., van Osch, T., & Sun, C. (2025). Machine unlearning of personally identifiable information in large language models. In Proceedings of the Natural Legal Language Processing Workshop 2025 (pp. 54–67). https://doi.org/10.18653/v1/2025.nllp-1.6
19. Yao, J., Chien, E., Du, M., Niu, X., Wang, T., Cheng, Z., & Yue, X. (2024). Machine unlearning of pre-trained large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8403–8419). https://doi.org/10.18653/v1/2024.acl-long.457
20. Cutler, E., Levonian, Z., & Christie, S. T. (2025). Detecting student intent for chat-based intelligent tutoring systems. arXiv. https://doi.org/10.48550/arXiv.2502.15096
21. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

22. Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. ACM Computing Surveys, 57(6), Article 152, 1–39. https://doi.org/10.1145/3712001

23. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models | European Data Protection Board (EDPB). (2024). Retrieved from: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf (date accessed: November 12, 2025).

24. Hui, Z., Dong, Y. R., Sivapiromrat, S., Shareghi, E., & Collier, N. (2025). PrivacyPAD: A reinforcement learning framework for dynamic privacy-aware delegation. arXiv. https://doi.org/10.48550/arXiv.2510.16054

25. Abdelnabi, S., Fay, A., Cherubin, G., Salem, A., Fritz, M., & Paverd, A. (2024). Are you still on track!? Catching LLM task drift with activations. arXiv. https://doi.org/10.48550/arXiv.2406.00799