

ZBORNİK RADOVA
COLLECTION OF SCIENTIFIC PAPERS



Zbornik radova Mediji i ekonomija 2026
Collection of scientific papers Media and economy 2026

Izdavač /Publisher:



Banjaluka



Banjaluka College

Za izdavača /For Publisher :

Danilo Novaković

Urednik zbornika / Editor in Chief:

prof. dr Svetlana Dušanić Gačić

Recenzenti:

prof. dr Jasna Čošabić, Austrija

prof. dr Inga Koryagina, Rusija

prof. dr Ana Anufrijeva, Srbija

prof. dr Željko Petrović, BiH

prof. dr Bojana Ostojić, Srbija

prof. dr Mihajlo Manić, Srbija

prof. dr Marija Boban, Hrvatska

prof. dr Nenad Novaković, BiH

prof. dr Mladen Miroslavljević, BiH

prof. dr Ljiljana Stojanović Bijelić, BiH

prof. dr Slavica Išaretović, BiH

prof. dr Zorana Agić, BiH

prof. dr Mirjana Milovanović, BiH

prof. dr Lidija Stefanovska, Sjeverna Makedonija

prof. dr Dragor Zarevski, Sjeverna Makedonija

doc. dr Malči Grivec, Slovenija

prof. dr Vjara Najdenova, Bugarska

prof. dr Tatjana Georginova, Sjeverna Makedonija

doc. dr Tamara Jovović, Crna Gora

Lektor /Lector:

Višnja Tatić - srpski jezik, Vesna Đurović- engleski jezik

ISSN 2566-4697 (Print)

ISSN 2744-1822 (Online)

ISBN

Copyright ©  2026

ZBORNİK RADOVA
COLLECTION OF SCIENTIFIC PAPERS

Međunarodna naučna konferencija
Mediji i ekonomija
International Scientific Conference
Media and Economy

septembar, 2026

ISSN 2566-4697 (Print)

ISSN 2744-1822 (Online)



Banjaluka, 2026.

Sadržaj

MEDIJI

UTICAJ INFLUENSER MARKETINGA NA KUPOVNE ODLUKE PRIPADNIKA GENERACIJE Z U CRNOJ GORI <i>Nikolina Pejović</i>	11
DATA LEAKAGE AND MACHINE UNLEARNING IN GENERATIVE AI: A LITERATURE SURVEY ON EXTRACTION VULNERABILITIES AND PRIVACY-PRESERVING COUNTERMEASURES <i>Dušan Rajčević, Ana Veljić, Marko Marković</i>	29
WHO BELONGS TO EUROPE? A MEDIA DISCOURSE ANALYSIS OF CITIZENSHIP, MOBILITY AND THE ECONOMIC PRIVILEGE OF THE EUROPEAN PASSPORT <i>Adem Olovčić, Dženana Telalagić, Ajla Aljović, Jasmin Jašarević</i>	43
FROM DIGITAL ENGAGEMENT TO SUSTAINABLE CHOICE: AN INTEGRATIVE MODEL OF MARKETING COMMUNICATION WITH GENERATION Z <i>Mirjana Milovanović</i>	57
MEDIA DIPLOMACY IN A MULTIPOLAR WORLD: ANALYZING THE EFFICACY AND FUTURE TRAJECTORY OF RUSSIAN STATE EVENT COVERAGE <i>Koryagina Inga</i>	69
OD DIGITALNIH KOMPETENCIJA DO ZDRAVSTVENE PISMENOSTI: ULOGA MEDIJSKE PISMENOSTI U KRITIČKOJ PROCENI ZDRAVSTVENIH INFORMACIJA NA DRUŠTVENIM MREŽAMA <i>Olja Arsenijević, Marija Lugonjić</i>	77
ULOGA CIVILA U IZVJEŠTAVANJU IZ SAVREMENIH ORUŽANIH SUKOBA <i>Dejan Kovačević</i>	95
MEDIA AND ECONOMICS: ADVERTISING OPTIMIZATION IN THE DIGITAL AGE <i>Drago Pupavac</i>	109
TRANSFORMATION OF THE ROLE OF THE PROFESSOR THROUGH THE DEVELOPMENT OF DIGITAL TOOLS AND THE APPLICATION OF ARTIFICIAL INTELLIGENCE <i>Maja Vojinović, Tatjana Davidov, Slavica Anđelić</i>	119
NOVINARSKA ETIKA OD PROFESIONALNE ODGOVORNOSTI DO IZAZOVA VJEŠTAČKE INTELIGENCIJE <i>Svetlana Dušanić Gačić</i>	129
STORYTELLING U SAVREMENOJ POSLOVNOJ KOMUNIKACIJI: KOMUNIKOLOŠKA PERSPEKTIVA DIGITALNE TRANSFORMACIJE <i>Selmira Đulić, Marko Milić</i>	135

AFRIČKI SOCIJALIZAM – PUT IZMEĐU KOLONIJALIZMA, KAPITALIZMA, SOCIJALIZMA I DIKTATURA <i>Željko Petrović</i>	147
EKONOMIJA PAŽNJE U SAVREMENOM PREDUZETNIŠTVU I ULOGA MEDIJA U KREIRANJU KONKURENTSKE PREDNOSTI PREDUZEĆA <i>Lena Komatina, Ana Anufrijević</i>	155
EKONOMSKA KRIZA, MIGRACIJE I MEDIJI: RODNA PERSPEKTIVA SAVREMENIH MIGRACIJA <i>Aleksa Jakica</i>	169
UTICAJ DRUŠTVENIH MEDIJA NA PONAŠANJE POTROŠAČA U DIGITALNOM OKRUŽENJU <i>Aida Jahić</i>	177
FOREIGN LANGUAGE COMPETENCE AS A PREREQUISITE FOR INTERNATIONAL ACADEMIC MOBILITY AND RESEARCH COLLABORATION <i>Tatjana Dragičević Radičević, Ivana Lešević</i>	191
TRANSFORMATION OF THE TOURISM INDUSTRY AND CONTEMPORARY MEDIA TRENDS <i>Aleksandar Lukić, Denis Stojkanović</i>	203
ИНТЕРФЕРЕНЦИЈА ОТ УНГАРСКИЯ ЕЗИК (УЕ) ВЪРХУ ЕЗИКА НА ЮЖНОСЛАВЯНСКИТЕ МАЛЦИНСТВА В УНГАРИЯ – ГЛОТОФАГИЯ <i>Вяра Найденова</i>	211
МЕДИЈИ КАО ФАКТОР АФИРМАЦИЈЕ СТОЧАРСТВА: АНАЛИЗА КРЕТАЊА УВОЗА И ИЗВОЗА ГРЛА СТОКЕ У СРБИЈИ (2019-2024) <i>Ђокић Кристина</i>	221
EVALUATING THE EXECUTION SAFETY AND SIMD OPTIMIZATION QUALITY OF LLM-GENERATED MODERN C++ CODE <i>Dušan Rajčević, Ivona Brajević, Goran Jocić</i>	237
MONETIZACIJA POVJERENJA U DIGITALNIM MEDIJIMA: PSIHOLOŠKI MEHANIZMI KOJI DIGITALNI UTJECAJ PRETVARAJU U EKONOMSKU VRIJEDNOST <i>Minel Duvnjak</i>	253
POSLOVNA IZVRSNOST U MEDIJSKIM ORGANIZACIJAMA: IZAZOVI I MODELI <i>Tatjana Janovac, Marijana Vukčević</i>	263
PROFESIONALNA ORIJENTACIJA I KARIJERNO VOĐENJE <i>Sanda Milošević</i>	271
VJEŠTAČKA INTELIGENCIJA U OBRAZOVANJU: JESMO LI BILI SPREMNI? <i>Zorana Agić, Slavica Išaretović</i>	279
HOW PERSONALIZATION IN MARKETING IMPROVES CUSTOMER EXPERIENCE: A MIXED METHOD ANALYSIS WITH A CASE STUDY OF NETFLIX <i>Esma Nur Cerinan Otović</i>	291
SPORTS BROADCASTING RIGHTS IN THE DIGITAL ERA: LEGAL AND ECONOMIC CHALLENGES <i>Danijela Glušac</i>	299
DIGITALNI MEDIJI U PRODAJNIM STRATEGIJAMA MALIH PROIZVOĐAČA ISTOČNE HRVATSKE <i>Zvonimir Filipović</i>	307

ETIKA U SAVREMENOM NOVINARSTVU POD PRITISKOM DIGITALNOG DOBA: PROFESIONALNI STANDARDI, KLIKABILNOST, DEZINFORMACIJE I VJEŠTAČKA INTELIGENCIJA <i>Dubravka Gajić</i>	315
--	-----

EKONOMIJA

UTICAJ POSLOVNE KOMUNIKACIJE NA POSLOVNE PERFORMANSE: PRIMER ANALIZE STAVOVA ZAPOSLENIH U REPUBLICI SRBIJI <i>Oliver Momčilović, Adriana Radosavac</i>	325
DOES SOCIALLY RESPONSIBLE INVESTMENT AFFECT LABOR RELATIONS? <i>Bojana Radović, Jelena Ignjatović, Lidija Weis</i>	337
IMPACTS OF UNPLANNED URBANIZATION ON MICROCLIMATIC TRANSFORMATION AND URBAN HEAT ISLAND INTENSIFICATION IN MOUNTAINOUS SETTLEMENTS: A CASE STUDY OF MURREE AND ADJACENT AREAS <i>Kamran Khan</i>	347
PODZEMNI RESURSI I ODRŽIVOST - KONFLIKT IZMEĐU EKSPLOATACIJE MINERALNIH SIROVINA I ZAŠTITE IZVORIŠTA VODA <i>Kasim Bajramović, Irhad Bajramović, Irfan Bajramović</i>	357
TRAFFIC NOISE AND ECOLOGICAL BALANCE: FROM TECHNICAL EMISSIONS TO ENVIRONMENTAL DAMAGE <i>Predrag Pravdić, Zorana Milosavljević, Darko Manjenčić, Violeta Đorđević</i>	375
MARKETINŠKI PROJEKTI ZASNOVANI NA BIHEJVIORALNOM PRISTUPU <i>Bojana Ostojić, Milena Cvjetković, Drago Cvijanović</i>	389
UTICAJ NEFINANSIJSKIH PRAKSI NA FINANSIJSKE PERFORMANSE AVIO-KOMPANIJA <i>Miloš Milošević, Zorica Davidović, Nataša Manojlović</i>	399
ZLATNI STANDARD XIX VIJEKA: IZMEĐU MONETARNE STABILNOSTI, IMPERIJALNE MOĆI I KRIZE POVJERENJA <i>Stevan Gajević</i>	407
LUXURY MARKETING IN THE GLOBAL MARKET AND THE NEED FOR FINANCIAL RESOURCES FOR BETTER LUXURY MARKETING DEVELOPMENT <i>Emanuela Esmerova, Aleksandar Kostadinovski</i>	423
UTICAJ MERA MONETARNE POLITIKE NA EKONOMIJU ŠVEDSKE U PERIODU FINANSIJSKE KRIZE <i>Nemanja Budimir</i>	435
DIGITAL TWINS AS DATA-DRIVEN SUPPORT FOR PERFORMANCE MANAGEMENT IN THE EFQM MODEL: A COMPARATIVE AND THEORETICAL ANALYSIS <i>Denčić Jelena, Petrović Tijana</i>	455
GREEN TRANSITION – ECONOMIC CHALLENGES AND MANAGERIAL STRATEGIES FOR SUSTAINABLE DEVELOPMENT <i>Denis Stojkanović, Aleksandar Lukić, Ratko Garić</i>	465
DINAMIKA PRINOSA NA OBVEZNICE U BOSNI I HERCEGOVINI: KOMPARATIVNA ANALIZA REGIONALNIH I EVROPSKIH KRETANJA U PERIODU 2015–2025 <i>Luka Tulić</i>	477

EXPONENTIAL HUMAN RESOURCE MANAGEMENT AS A CATALYST FOR AGILE ORGANIZATIONS <i>Jelena Lukić Nikolić</i>	497
DIGITALNA TRANSFORMACIJA POSLOVANJA PORESKE UPRAVE: EFIKASNOST, DIGITALNE USLUGE I SAVREMENI IZAZOVI <i>Petar Trajković</i>	513
ANALIZA FINANSIJSKIH IZVJEŠTAJA KAO ALAT ZA UNAPREĐENJE POSLOVNOG ODLUČIVANJA <i>Džejlana Čejvan, Aida Mehić</i>	527
УПРАВЉАЊЕ ВОДНИМ РЕСУРСИМА У РЕПУБЛИЦИ СРБИЈИ <i>Анђела Миленковић Ицић</i>	545
ULAGANJA U OBRAZOVANJE I PRODUKTIVNOST U BOSNI I HERCEGOVINI <i>Mila Praštalo, Ema Trninić</i>	555
ANALIZA LIKVIDNOSTI LIZING KOMPANIJA U REPUBLICI SRBIJI <i>Ivana Medved, Biljana Đuričić, Momčilo Vasiljević, Marijana Milanović</i>	575
ULOGA EKOLOŠKOG UPRAVLJAČKOG RAČUNOVODSTVA U ZELENOM TRANZICIJI PREDUZEĆA <i>Nedeljko Babić</i>	583
KONCEPTUALNI MODEL ODRŽIVOG RAZVOJA RURALNIH DESTINACIJA I TURISTIČKIH PROIZVODA PRIMENOM AHP METODE <i>Predrag Todorov</i>	595
WELLNESS TURIZAM KAO FAKTOR ODRŽIVOG RAZVOJA I KONKURENTNOSTI TURISTIČKIH DESTINACIJA <i>Milan Novović, Nataša Marjanovi, Živana Krejić</i>	607
TRANSPARENTNOST JAVNIH INFORMACIJA KAO DETERMINANTA ODLUKE O KUPOVINI ŽIVOTNOG OSIGURANJA <i>Anufrijević Ana, Dašić Goran, Aničić Anđelka</i>	621
PREGLED EKONOMSKIH POSLEDICA PANDEMIJE COVID-19 I RUSKO-UKRAJINSKOG ORUŽANOG KONFLIKTA NA EKONOMIJU EVROPSKE UNIJE <i>Pavle Jeremić, Vuk Mirčetić, Pavle Brzaković</i>	635
UPRAVLJANJE NOVČANIM TOKOVIMA U SAVREMENOM POSLOVANJU <i>Marko Milić, Darko Pašalić</i>	645
DIGITALNA TRANSFORMACIJA I NOVO LICE LIDERSTVA <i>Marijana Vukčević, Tatjana Janovac</i>	657
LINUX ECOSYSTEM SECURITY AND THE RISKS OF USER REPOSITORIES: A CASE STUDY OF THE “ATOMIC ARCH” ATTACK ON THE ARCH USER REPOSITORY <i>Aleksandar Šijan, Luka Ilić, Dejan Viduka</i>	665
DOSTOP DO FINANCIRANJA MALIH IN SREDNJE VELIKIH PODJETIJ: PRIMERJALNA ANALIZA IZBRANIH DRŽAV <i>Malči Grivec</i>	673
GENERACIJSKI KONFLIKTI IN NJIHOV VPLIV NA NASLEDSTVO V DRUŽINSKIH PODJETJIH <i>Neja Kosec, Anita Blažič</i>	693
KROŽNO GOSPODARSTVO IN OKOLJSKI PRITISKI: PRIMERJALNA ANALIZA SLOVENIJE IN EVROPSKE UNIJE <i>Kaja Miklavčič, Malči Grivec</i>	705

DATA LEAKAGE AND MACHINE UNLEARNING IN GENERATIVE AI: A LITERATURE SURVEY ON EXTRACTION VULNERABILITIES AND PRIVACY-PRESERVING COUNTERMEASURES

Dušan Rajčević¹

Ana Veljić²

Marko Marković³

Keywords: *Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Data Leakage, Threat Taxonomy, Differential Privacy, Machine Unlearning*

Abstract – *Bringing Large Language Models (LLMs) into business processes through fine-tuning or Retrieval-Augmented Generation (RAG) can pose serious risks, particularly when it comes to protecting sensitive information and Personally Identifiable Information (PII). In this paper, we conduct a thorough review of existing literature to identify and analyze the various ways data can leak in these systems. By examining peer-reviewed research on key security threats, we create a framework to categorize privacy risks throughout the AI lifecycle. We also look at current protective measures such as differential privacy and machine unlearning. Our goal is to provide a clear threat classification, assess how effective existing defenses are, and offer guidance to security architects working on privacy-focused generative AI solutions. Using a structured review of 35 coded primary publications through August 24, 2026, we distinguish parameter-mediated disclosure from model training and retrieval-mediated disclosure from RAG datastores. The evidence indicates that no single countermeasure spans the full lifecycle: differential privacy provides the clearest ex ante formal protection, whereas current LLM unlearning remains primarily an approximate post-hoc remedy that requires attack-based verification.*

CURENJE PODATAKA I MAŠINSKO ZABORAVLJANJE U GENERATIVNOJ VJEŠTAČKOJ INTELIGENCIJI: PREGLED LITERATURE O RANJIVOSTIMA EKSTRAKCIJE PODATAKA I MJERAMA ZA OČUVANJE PRIVATNOSTI

Ključne riječi: *veliki jezički modeli (LLM), generisanje potpomognuto pretraživanjem (RAG), curenje podataka, taksonomija prijetnji, diferencijalna privatnost, mašinsko zaboravljanje*

¹ Faculty for Applied Management, Economics and Finance, Belgrade, Serbia

² Faculty of Economics and Engineering Management, Novi Sad, Serbia

³ Faculty of Economics and Engineering Management, Novi Sad, Serbia

Sažetak: Integracija velikih jezičkih modela (LLM) u poslovne procese putem dodatnog prilagođavanja modela (fine-tuning) ili generisanja potpomognutog pretraživanjem (Retrieval-Augmented Generation – RAG) može predstavljati ozbiljne rizike, naročito kada je riječ o zaštiti osjetljivih informacija i podataka koji omogućavaju identifikaciju pojedinca (Personally Identifiable Information – PII). U ovom radu sprovodi se sveobuhvatan pregled postojeće literature s ciljem identifikovanja i analize različitih načina na koje može doći do curenja podataka u ovim sistemima. Analizom recenziranih istraživanja o ključnim bezbjednosnim prijetnjama formira se okvir za kategorizaciju rizika po privatnost tokom cjelokupnog životnog ciklusa sistema vještačke inteligencije. Takođe se razmatraju postojeće mjere zaštite, kao što su diferencijalna privatnost i mašinsko zaboravljanje. Cilj rada je da pruži jasnu klasifikaciju prijetnji, procijeni efikasnost postojećih mehanizama zaštite i ponudi smjernice bezbjednosnim arhitektama koji rade na razvoju generativnih AI rješenja usmjerenih na očuvanje privatnosti. Na osnovu strukturiranog pregleda 35 kodiranih primarnih publikacija, zaključno sa 24. avgustom 2026. godine, u radu se pravi razlika između otkrivanja podataka posredovanog parametrima modela, koje proizlazi iz obuke modela, i otkrivanja podataka posredovanog pretraživanjem iz RAG skladišta podataka. Rezultati pokazuju da nijedna pojedinačna protivmjera ne obuhvata cjelokupan životni ciklus sistema: diferencijalna privatnost pruža najjasniju formalnu zaštitu *ex ante*, dok trenutno dostupne metode mašinskog zaboravljanja kod LLM-ova uglavnom predstavljaju približnu *post hoc* mjeru koja zahtijeva provjeru zasnovanu na napadima.

1. Introduction

Generative artificial intelligence systems are increasingly combining large language models (LLMs) with organization-specific data. Two implementation patterns are particularly common. Fine-tuning modifies model parameters using domain data, while search-assisted generation (RAG) maintains a searchable non-parametric database and feeds the found fragments to the model at runtime (Lewis et al., 2020). Both patterns improve task relevance, but both also create privacy risks when underlying material contains personally identifiable information (PII), confidential records, proprietary text, or other data that should not be accessible to unauthorized users.

The risk is not just hypothetical. Language models can remember rare sequences from training and later repeat them under appropriate prompts (Carlini et al., 2019, 2021). Memory is enhanced by duplication, model capacity, and available query context (Carlini et al., 2023; Kandpal et al., 2022; Tirumala et al., 2022). Black-box attacks reconstructed PII from fine-tuned models, including cases where conventional data cleansing or sentence-level differential privacy (DP) reduced but did not eliminate detection (Lukas et al., 2023). Recent research shows that alignment is not a complete barrier. Attacks on production models recovered thousands of memorized training examples after bypassing protection mechanisms (Nasr et al., 2025). The privacy risks are not limited to cases where the model repeats the training data verbatim. Membership and provenance attacks can allow an adversary to determine whether a particular record or user text was included in the training data, even if the model never reproduces that text exactly (Mattern et al., 2023; Shokri et al., 2017; Song & Shmatikov, 2019).

RAG modifies but does not eliminate this problem. Its external memory facilitates knowledge updating and can reduce dependence on parametric memory, but the retrieved context itself becomes a questionable privacy boundary. Zeng et al. (2024) found that RAG can mitigate some data leakage from model training, but at the same time introduces leakage from the private search database. Later attacks extracted the contents of the RAG via

instruction tracing, query injection, graph searching, or seemingly benign requests (Liu et al., 2026; Qi et al., 2025; Wang et al., 2026). Indirect query injection and database poisoning further demonstrate that unreliable found content can manipulate generation and weaken the assumed separation between data and instructions (Greshake et al., 2023; Zou et al., 2025).

Defense measures research is divided into several approaches. Differential privacy provides a formal method to limit the influence of information about an individual on a model or its outputs (Abadi et al., 2016; Dwork et al., 2006). However, stronger privacy protection usually requires more noise during training, which may reduce the accuracy of the model or its general utility. Selective DP narrows protection to policy-defined sensitive tokens, improving utility but creating a dependency on the quality of the sensitivity policy itself (Shi, Cui, et al., 2022; Shi, Shea, et al., 2022). Unlearning addresses a different point in the life cycle. It tries to remove the influence of post-training data. Although exact removal is possible in limited settings (Bourtoule et al., 2021; Cao & Yang, 2015; Guo et al., 2020), recent LLM tests have shown that approximate forgetting can suppress recall without reliably removing privacy leaks or preserving model utility (Maini et al., 2024; Shi et al., 2025).

This literature leaves a practical gap. Leak studies often examine a single attack surface, while privacy decisions include data acquisition, training, search, generation, and deletion. At the same time, preventive controls and subsequent forgetting are sometimes considered as substitutes even though they provide different guarantees and operate at different stages. This review therefore asks four questions:

- RQ1.** What leakage mechanisms and adversarial pathways have been demonstrated through the training, fine-tuning, RAG and post-deployment phases?
- RQ2.** What system and data factors consistently increase extraction risk, and how do parameter-mediated leaks differ from search-mediated ones?
- RQ3.** What protections do data curation, differential privacy, RAG-specific privacy controls and machine forgetting provide, and where do their limitations remain?
- RQ4.** What implementation-oriented control strategy emerges from the evidence for privacy-focused generative AI systems?

The contribution is a lifecycle-based taxonomy of threats that separates parameter-mediated detection from search-mediated detection, a comparative synthesis of preventive and deletion-based measures, and a set of architectural implications based on reported attacks and defenses.

2. Methods

Review design and scope

The paper used a structured literature review design rather than a PRISMA-style systematic review or statistical meta-analysis. Searches were completed in July and August 2026. The review included proceedings of the USENIX Security Conference, records from the IEEE Symposium on Security and Privacy, ACM Digital Library, ACL Anthology, Proceedings of Machine Learning Research, NeurIPS Proceedings, ICLR/OpenReview, and

arXiv for version discovery and tracking. When there was a peer-reviewed paper, the published version of the conference paper or findings was used, rather than the preprint.

Four query families were used:

1. Language model I memorization, training data extraction, leakage, PII, membership inference, or provenance
2. Retrieval-augmented generation (RAG) and privacy, leakage, extraction, prompt injection, or poisoning
3. Differential privacy or selective differential privacy and language model or RAG
4. Machine unlearning or data deletion and language model, generative model, or privacy Backward reference tracking was used for basic methods, while forward/version tracking was used to find peer-reviewed successors to recent preprints.

Eligibility and coding

Eligible studies were primary publications in the English language that either demonstrated or measured a privacy-relevant leakage mechanism, evaluated preventive privacy controls, or proposed/evaluated data erasure or machine forgetting. Basic studies on differential privacy and forgetting are retained even when they predate modern LLMs, as they define the guarantees that later work relies on. Studies were excluded when they addressed general AI safety without a data confidentiality mechanism, when they were reviews rather than primary evidence, when they duplicated a later peer-reviewed version, or when they appeared after the review's time cut-off. The basic RAG work of Lewis et al. (2020) is cited for system definition, but not included in offense/defense counts.

The coded evidence set contained 35 primary publications:

- 17 focused primarily on leakage, extraction or threat measurement,
- nine on preventive privacy mechanisms,
- nine to forget or erase.

Twenty-one of the 35 coded studies (60.0%) were published from 2022 to the review cutoff in 2026. For each study, the synthesis recorded the life cycle stage, protected resource, adversary approach, attack or defense mechanism, reported evaluation metric, assurance type, and stated utility tradeoff. Quantitative values were included in the results only when they were explicitly reported in the original study. Because the attack targets, datasets, model families, and metrics differ significantly, no pooled effect size was calculated.

The search interfaces used in the various sources do not allow a stable, common export of all ranked candidates, and this review did not retain a PRISMA-compatible log of initial filtering. Consequently, the number of candidates in the filtering phase is not reported. This decision avoids inventing a precision not supported by the review record, but also limits reproducibility compared to a pre-registered systematic review.

3. Results

RQ1: Leakage mechanisms across the generative-AI lifecycle

The literature shows that leakage is a life-cycle property, not just a single model failure. Table 1 summarizes six recurring attack surfaces. During training, sensitive text can be encoded in parameters and later retrieved through generation, membership inference, or provenance testing. During fine-tuning, domain-specific PII may be particularly exposed because the data is narrower and often more sensitive. RAG adds another level: the database and found context can be directly extracted even when the underlying model has never been trained on those records. Database poisoning and indirect query injection further turn the search layer into an integrity problem that can facilitate inadvertent discovery. Finally, deletion requests create a verification problem, as a model that has seemingly been “forgotten” may still retain extractive traces.

Lifecycle stage	Primary asset	Demonstrated mechanism	Adversary surface	Representative evidence
Pre-training / base training	Training sequences	Memorization, verbatim extraction, membership/provenance inference	Black-box generation or score access	Carlini et al., 2019, 2021, 2023; Shokri et al., 2017; Song & Shmatikov, 2019
Fine-tuning	Domain data and PII	PII extraction, inference, reconstruction	Model API	Lukas et al., 2023; Mattern et al., 2023
RAG ingestion / index	External documents	Indirect prompt injection; malicious datastore entries	Ability to influence retrieved content	Greshake et al., 2023; Zou et al., 2025
RAG retrieval / context	Private datastore text	Instruction-following extraction and regurgitation	Repeated RAG queries	Zeng et al., 2024; Qi et al., 2025
Graph / stealth RAG querying	Entities, relationships, implicit knowledge	Structured extraction; benign-query exploration	Black-box RAG access	Liu et al., 2026; Wang et al., 2026; Zhang et al., 2026
Deletion / unlearning	Previously learned sensitive data	Residual memorization or privacy leakage after approximate forgetting	Post-unlearning probing	Maini et al., 2024; Shi et al., 2025; Ashuach et al., 2025

Table 1. Lifecycle threat taxonomy for data leakage in generative AI

Note. Authors' synthesis of the cited primary studies. The categories describe the principal exposure mechanism; individual studies can span more than one lifecycle stage.

RQ2: Risk amplifiers and two leakage planes

In parameter-mediated attacks, memory is not uniform. Duplication is a powerful and repeatedly confirmed amplification factor. Kandpal et al. (2022) report that, on average, a sequence that appeared ten times in the training data was generated about 1,000 times more often than a sequence that appeared once. Carlini et al. (2023) further identify log-linear relationships between memorization and model capacity, number of duplications, and query context length, while Tirumala et al. (2022) find that larger language models memorize examples faster during training. These findings support data deduplication as a useful risk mitigation measure, but do not imply that it is impossible to extract unique sensitive records.

RAG shows different boosters, because private data does not need to be stored in parameters. Extraction can occur when the retriever repeatedly inserts protected records into the model context, and the model following the instructions is instructed to reproduce them. Qi et al. (2025) demonstrated almost perfect attack success on 25 customized GPTs with a maximum of two queries and reported a verbatim extraction rate of 41% from a book of 77,000 words and 3% from a corpus of 1,569,000 words using 100 generated queries. LeakDojo later benchmarked six attacks across 14 LLMs and four datasets and found that stronger instruction monitoring correlated with higher leaks, while improvements in RAG fidelity could also increase leaks (Zhang et al., 2026). Graph RAG changes the form of exposure. Liu et al. (2026) observed less raw text leakage in some environments, but greater vulnerability to structured entity and relation extraction. Wang et al. (2026) showed that malicious queries are not necessary, as their benign query attack (“IKEA attack”) significantly outperformed basic extraction methods under existing defenses.

These results motivate the two-plane model. Parameter-mediated leakage originates from information encoded during training or fine-tuning. And a search-mediated leak originates from external memory that is injected into the context at runtime.

Planes can coexist in a single application and converge on the same output interface, but their causes differ. Removing a document from the RAG index can eliminate future retrieval of that document without changing model parameters; the same deletion does not remove the copy already learned during training. Conversely, parameter-level privacy controls do not prevent an authorized retriever from providing an overly sensitive document to the generator.

Study	Setting	Reported result	Interpretive use in this survey
Kandpal et al. (2022)	Training duplication	10 training occurrences $\rightarrow$ $\sim 1,000\times$ higher average regeneration than one occurrence	Duplication strongly amplifies parametric leakage.
Lukas et al. (2023)	PII leakage from fine-tuned GPT-2	New attacks extracted up to $10\times$ more PII than prior attacks; sentence-level DP still leaked $\sim 3\%$ of PII sequences	Scrubbing and sentence-level DP reduced but did not eliminate PII disclosure in the tested setting.
Yu et al. (2022)	DP fine-tuning, MNLI	87.8% accuracy with RoBERTa-Large at $\epsilon=6.7$ versus 90.2% non-private	Formal privacy can retain substantial utility, but a measurable utility gap remained.

Study	Setting	Reported result	Interpretive use in this survey
Qi et al. (2025)	Production RAG extraction	Near-perfect success on 25 customized GPTs with ≤ 2 queries; 41% and 3% verbatim extraction in two larger-corpus tests	RAG datastores can be directly extractable at meaningful scale.
Zou et al. (2025)	RAG datastore poisoning	90% attack success with five malicious texts per target question in a database containing millions of texts	Small poisoned inserts can control retrieval-grounded output in the evaluated systems.
Shi et al. (2025)	LLM unlearning benchmark	Eight algorithms on 7B models; only one avoided severe privacy leakage	Suppressing recall did not reliably imply privacy-safe unlearning.
Wang et al. (2026)	Benign-query RAG extraction	IKEA surpassed baselines by $>80\%$ in extraction efficiency and $>90\%$ in attack success rate	Input/output maliciousness filters alone do not cover stealth extraction.

Table 2. Representative quantitative findings reported in the primary literature
Values are reproduced from the cited studies and are not directly comparable because datasets, models, threat models and metrics differ. ϵ denotes the differential-privacy budget reported by Yu et al. (2022).

RQ3: Effectiveness and limits of privacy-preserving countermeasures

Preventive controls have the strongest evidence when they act before sensitive data becomes widely exposed. Deduplication reduces storage-based extraction, and careful removal of PII data reduces risk, but neither of these methods is a formal guarantee of privacy (Kandpal et al., 2022; Lukas et al., 2023). Differential privacy (DP) provides the clearest formal protection because it limits how much the presence or absence of a protected record can change the distribution of outputs (Dwork et al., 2006). DP-SGD extends this principle to deep learning (Abadi et al., 2016), while user-level and public overtraining approaches show that cost-to-utility depends heavily on data volume, model initialization, and privacy granularity (Kerrigan et al., 2020; McMahan et al., 2018). Yu et al. (2022) showed that large pretrained models can be fine-tuned with useful task performance under explicit privacy budgets.

The margin of protection is important. Privacy regularization can improve empirical trade-offs between privacy and utility, but does not provide the same worst-case guarantee as DP (Mireshghallah et al., 2021). Selective DP protects policy-defined sensitive tokens and can preserve more utility, but missed sensitive tokens weaken coverage; the work Just Fine-tune Twice explicitly studies this imperfect policy problem (Shi, Cui, et al., 2022; Shi, Shea, et al., 2022). For RAG, Mori et al. (2026) proposed DP-SynRAG, which creates a multiple differentially private synthetic database and thus avoids repeated query time noise and privacy overhead for each subsequent RAG query. Its results are promising, but confirm effectiveness for the evaluated design, not a universal solution for all RAG leakage paths.

Machine forgetting (unlearning) provides a follow-up response when data must be removed after learning. Basic works formalized efficient deletion through retraining and certified removal for limited model classes (Cao & Yang, 2015; Ginart et al., 2019; Guo et al., 2020), while SISA training limits retraining to affected partitions (Bourtole et al., 2021). Scrubbing deep networks attempts to make the model weights less recognizable compared to a model trained without forgotten data (Golatkar et al., 2020). However, for LLMs, the current evidence remains largely approximate. TOFU found that the underlying methods evaluated did not achieve efficient forgetting against benchmark criteria (Maini et al., 2024), while MUSE found severe limitations in privacy, utility, scale, and sequential requirements across eight algorithms (Shi et al., 2025). Targeted methods like REVS and selective token forgetting improve the privacy–utility tradeoff in evaluated settings (Ashuach et al., 2025; Wan et al., 2025), but none establish a general guarantee that a large model is distributionally equivalent to retraining without deleted data.

Control class	Lifecycle position	Evidence / guarantee	Principal limitation	Representative sources
Deduplication and PII curation	Before training	Empirically reduces regeneration and direct exposure	No formal guarantee; detectors and curation can miss sensitive data	Kandpal et al., 2022; Lukas et al., 2023
Full differential privacy	Training / fine-tuning	Formal privacy under stated adjacency and (ϵ, δ) budget	Utility and compute cost; guarantee depends on privacy unit and accounting	Dwork et al., 2006; Abadi et al., 2016; Yu et al., 2022
Selective differential privacy	Fine-tuning	Formal protection for policy-defined sensitive portions	Coverage depends on sensitivity policy and missed-token handling	Shi, Cui, et al., 2022; Shi, Shea, et al., 2022
Private RAG datastore construction	Before / during retrieval	DP-SynRAG provides a reusable DP synthetic datastore in the evaluated framework	Does not cover every retrieval, prompt-injection, or access-control failure	Mori et al., 2026
Exact / certified / partitioned unlearning	Post-training	Strong guarantees in restricted settings; retraining can be localized	Not directly scalable to monolithic frontier LLM training	Cao & Yang, 2015; Guo et al., 2020; Bourtole et al., 2021
Approximate LLM unlearning / editing	Post-training	Can reduce target recall while preserving more utility in some tests	Residual privacy leakage, collateral utility loss and weak equivalence-to-retraining evidence	Maini et al., 2024; Shi et al., 2025; Ashuach et al., 2025; Wan et al., 2025

Table 3. Countermeasure classes and evidence boundaries

Note. Authors' synthesis. A "formal" guarantee applies only under the assumptions and privacy unit defined by the cited method; it should not be generalized to system components outside that threat model.

RQ4: Architecture implications for privacy-focused generative AI

The evidence favors multi-layered defenses over reliance on a single privacy mechanism. At data entry, organizations should minimize collection, maintain data provenance, remove duplicate text, and classify or redact sensitive fields before training or indexing. When sensitive data must influence model parameters, differential privacy (DP) should be considered where the privacy unit and utility trade-off fit in the application. For RAG, the database should remain a separate security boundary: access should be segmented by user and tenant, entry of untrusted data should be controlled, and search should expose only the context necessary for the task. Defenses against query injection should treat found text as potentially adversarial, as an integrity failure can become a confidentiality failure (Greshake et al., 2023; Zou et al., 2025).

At the output/API layer, the reviewed attacks imply the need for extraction-oriented red-teaming, query volume monitoring, and checks for PII or unusually literal context reproduction. These are architectural implications of the observed attack mechanisms, not universal controls with a single measurable efficiency. Deletion workflows should preserve enough data provenance to determine whether the targeted information exists only in the RAG database, in the fine-tuning data, or in both. Removing a RAG record is preferable to parametric forgetting when the model has never learned that record; when parameters are affected, approximate forgetting should be validated with extraction, membership, or reconstruction tests, and not judged solely on rejection behavior.

4. Discussion***Interpretation of the research questions***

RQ1 is answered through a life cycle taxonomy in which privacy exposure occurs before, during, and after generation. Training data extraction and membership inference work through the model's parametric memory, while RAG creates a non-parametric database that can be searched, poisoned, or indirectly trained. This distinction is important because the same output can reveal data for completely different reasons. A rejection layer that masks the memorized text does not provide the underlying RAG base, and the private RAG base does not remove retroactively memorized training records.

RQ2 identifies recurrent risk factors but not a single universal predictor. Duplication, capacity, and query context are associated with stronger parametric storage. In RAG, extraction increases with factors such as instruction-following behavior, search exposure, repeated queries, and structured search design (Liu et al., 2026; Qi et al., 2025; Zhang et al., 2026). Studies use different models and metrics, so these relationships cannot be converted into a common numerical risk score. They are more appropriately used as engineering indicators to focus privacy testing.

RQ3 shows that differential privacy and machine forgetting are complementary, not interchangeable. DP is preventive and provides a mathematical guarantee defined by the privacy budget and the ratio of adjacent datasets. Forgetting is remedial and raises the question of whether the influence of selected data can be removed after training. Accurate retraining remains the clearest reference behavior, while current forgetting methods in LLMs often optimize proxies such as reduced recall, rejection, or targeted parameter

modification. The gap between behavioral forgetting and privacy removal is crucial: MUSE shows that the method can reduce memorization and still leave a serious privacy leak (Shi et al., 2025).

RQ4 therefore supports a layered architecture. The strongest design is one that prevents unnecessary sensitive data from entering model parameters, keeps search memory separate and access-controlled, limits context exposure, tests the output interface against extraction, and maintains a delete path with independent verification. This interpretation is also consistent with mixed evidence on RAG: RAG may reduce some training data leakage because information does not have to be “embedded” in parameters, but at the same time creates a more directly searchable database (Zeng et al., 2024).

Practical implications for security architects

For a corporate implementation, the first architectural question should be where sensitive information must reside. If information can remain external, RAG offers operational advantages because records can be updated or removed without retraining; however, search authorization and context leakage then become key security requirements. If sensitive information must be learned through fine-tuning, data minimization, deduplication, and explicit differential privacy decisioning should be performed prior to training. The deletion service should record whether the subject’s data has reached the training set, search base, cached queries, or multiple components. This origin determines whether deletion from the index is sufficient or whether retraining/forgetting is also required.

Limitations

This review has four main limitations

1. This is a structured review, not a pre-registered systematic review, and does not provide a PRISMA number of candidates at the filtering stage.
2. The evidence base is changing rapidly, particularly for RAG safety and forgetting in LLMs; results published after August 24, 2026 are not included.
3. Attack success, privacy budgets, extraction rates, and forgetting metrics are defined differently across studies, making a defensible pooled effect estimate impossible.
4. Many experiments use specific open models, synthetic forgetting sets, or controlled RAG configurations. The resulting mechanisms are informative, but absolute attack or defense rates cannot be assumed to be invariant when transferred to another model, data distribution, or access control regime.

5. Conclusion

The reviewed literature shows that privacy leakage in generative artificial intelligence cannot be reduced to memorization alone. Parameter-mediated leakage occurs when training or fine-tuning data can be recovered through model behavior. Retrieval-mediated leakage (RAG) occurs when external memory is repeatedly exposed to the generator and becomes extractive through direct, indirect, structured, or seemingly benign queries. These two planes require different controls, even though they meet at the same exit to the user.

Preventive measures have the strongest evidence when implemented early. Deduplication and curation of PII data reduce exposure but provide no formal assurance. Differential Privacy (DP) offers the clearest formal protection within a defined privacy and budget unit, including new RAG-specific approaches such as differentially private synthetic databases. Machine forgetting (unlearning) remains necessary for subsequent erasure, but current methods in LLMs should be treated as approximate, unless equivalence with retraining or a comparable formal guarantee is established.

In practice, privacy-focused generative AI should combine:

- data minimization and origin tracking,
- DP to the appropriate extent,
- search layer isolation,
- extraction testing,
- and erasure workflows that verify forgetting, rather than inferring it based on opt-out behavior alone.

Literature

- Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. <https://doi.org/10.1145/2976749.2978318>
- Ashuach, T., Tutek, M., & Belinkov, Y. (2025). REVS: Unlearning sensitive information in language models via rank editing in the vocabulary space. *Findings of the Association for Computational Linguistics: ACL 2025*, 14774–14797. <https://doi.org/10.18653/v1/2025.findings-acl.763>
- Bourtole, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine unlearning. *2021 IEEE Symposium on Security and Privacy*, 141–159. <https://doi.org/10.1109/SP40001.2021.00019>
- Cao, Y., & Yang, J. (2015). Towards making systems forget with machine unlearning. *2015 IEEE Symposium on Security and Privacy*, 463–480. <https://doi.org/10.1109/SP.2015.35>
- Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. *28th USENIX Security Symposium (USENIX Security 19)*, 267–284. USENIX Association. <https://www.usenix.org/conference/usenixsecurity19/presentation/carlini>
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650. USENIX Association. <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2023). Quantifying memorization across neural language models. *International Conference on Learning Representations*. https://openreview.net/forum?id=TatRHT_1cK
- Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In S. Halevi & T. Rabin (Eds.), *Theory of Cryptography: TCC 2006 (Lecture Notes in Computer Science, Vol. 3876, pp. 265–284)*. Springer. https://doi.org/10.1007/11681878_14
- Ginart, A. A., Guan, M. Y., Valiant, G., & Zou, J. Y. (2019). Making AI forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32. <https://proceedings.neurips.cc/paper/2019/hash/cb79f8fa58b91d3af6c9c991f63962d3-Abstract.html>

- Golatkar, A., Achille, A., & Soatto, S. (2020). Eternal sunshine of the spotless net: Selective forgetting in deep networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9304–9312. https://openaccess.thecvf.com/content_CVPR_2020/html/Golatkar_Eternal_Sunshine_of_the_Spotless_Net_Selective_Forgetting_in_Deep_CVPR_2020_paper.html
- Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 79–90. <https://doi.org/10.1145/3605764.3623985>
- Guo, C., Goldstein, T., Hannun, A., & van der Maaten, L. (2020). Certified data removal from machine learning models. Proceedings of the 37th International Conference on Machine Learning, 119, 3832–3842. PMLR. <https://proceedings.mlr.press/v119/guo20c.html>
- Kandpal, N., Wallace, E., & Raffel, C. (2022). Deduplicating training data mitigates privacy risks in language models. Proceedings of the 39th International Conference on Machine Learning, 162, 10697–10707. PMLR. <https://proceedings.mlr.press/v162/kandpal22a.html>
- Kerrigan, G., Slack, D., & Tuyls, J. (2020). Differentially private language models benefit from public pre-training. Proceedings of the Second Workshop on Privacy in NLP, 39–45. <https://doi.org/10.18653/v1/2020.privatenlp-1.5>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Liu, J., Zhang, J., & Wang, S. (2026). Exposing privacy risks in graph retrieval-augmented generation. Findings of the Association for Computational Linguistics: ACL 2026, 18073–18093. <https://doi.org/10.18653/v1/2026.findings-acl.899>
- Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models. 2023 IEEE Symposium on Security and Privacy, 346–363. <https://doi.org/10.1109/SP46215.2023.10179300>
- Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). TOFU: A task of fictitious unlearning for LLMs. First Conference on Language Modeling (COLM 2024). <https://openreview.net/forum?id=B41hNBwLo>
- Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., & Berg-Kirkpatrick, T. (2023). Membership inference attacks against language models via neighbourhood comparison. Findings of the Association for Computational Linguistics: ACL 2023, 11330–11343. <https://doi.org/10.18653/v1/2023.findings-acl.719>
- McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2018). Learning differentially private recurrent language models. International Conference on Learning Representations. <https://openreview.net/forum?id=BJ0hFIZ0b>
- Mireshghallah, F., Inan, H. A., Hasegawa, M., Rühle, V., Berg-Kirkpatrick, T., & Sim, R. (2021). Privacy regularization: Joint privacy-utility optimization in language models. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3799–3807. <https://doi.org/10.18653/v1/2021.naacl-main.298>
- Mori, J., Kakizaki, K., Miyagawa, T., & Sakuma, J. (2026). Differentially private synthetic text generation for retrieval-augmented generation (RAG). Findings of the Association for Computational Linguistics: ACL 2026, 1216–1235. <https://doi.org/10.18653/v1/2026.findings-acl.62>
- Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Tramèr, F., & Lee, K. (2025). Scalable extraction of training data from aligned, production language models. International Conference on Learning Representations. <https://>

proceedings.iclr.cc/paper_files/paper/2025/hash/cce0e917b050208170151f77b497fc71-Abstract-Conference.html

- Qi, Z., Zhang, H., Xing, E. P., Kakade, S., & Lakkaraju, H. (2025). Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/79cafa874121a3435d8a54f454b646b4-Abstract-Conference.html
- Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy*, 3–18. <https://doi.org/10.1109/SP.2017.41>
- Shi, W., Cui, A., Li, E., Jia, R., & Yu, Z. (2022). Selective differential privacy for language modeling. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2848–2859. <https://doi.org/10.18653/v1/2022.naacl-main.205>
- Shi, W., Shea, R., Chen, S., Zhang, C., Jia, R., & Yu, Z. (2022). Just fine-tune twice: Selective differential privacy for large language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6327–6340. <https://doi.org/10.18653/v1/2022.emnlp-main.425>
- Shi, W., Lee, J., Huang, Y., Malladi, S., Zhao, J., Holtzman, A., Liu, D., Zettlemoyer, L., Smith, N. A., & Zhang, C. (2025). MUSE: Machine unlearning six-way evaluation for language models. *International Conference on Learning Representations*. <https://openreview.net/forum?id=TArmA033BU>
- Song, C., & Shmatikov, V. (2019). Auditing data provenance in text-generation models. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 196–206. <https://doi.org/10.1145/3292500.3330885>
- Tirumala, K., Markosyan, A. H., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35, 38274–38290. https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa0509f4dab6807e2cb465715bf2d249-Abstract-Conference.html
- Wan, Y., Ramakrishna, A., Chang, K.-W., Cevher, V., & Gupta, R. (2025). Not every token needs forgetting: Selective unlearning balancing forgetting and utility in large language models. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 1827–1835. <https://doi.org/10.18653/v1/2025.findings-emnlp.96>
- Wang, Y., Qu, W., Zhai, S., Jiang, Y., Liu, Z., Liu, Y., Dong, Y., & Zhang, J. (2026). Silent leaks: Implicit knowledge extraction attack on RAG systems. *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2026/hash/2974844555dc383ea16c5f35833c7a57-Abstract-Conference.html
- Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., & Zhang, H. (2022). Differentially private fine-tuning of language models. *International Conference on Learning Representations*. <https://openreview.net/forum?id=Q42f0dfjECO>
- Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y., & Tang, J. (2024). The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). *Findings of the Association for Computational Linguistics: ACL 2024*, 4505–4524. <https://doi.org/10.18653/v1/2024.findings-acl.267>
- Zhang, M., Dong, J., Lu, B., Li, W., Zhang, X., Zhang, T., & Qiu, H. (2026). LeakDojo: Decoding the leakage threats of RAG systems. *Findings of the Association for Computational Linguistics: ACL 2026*, 5790–5811. <https://doi.org/10.18653/v1/2026.findings-acl.287>
- Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. *34th USENIX Security Symposium (USENIX Security 25)*, 3827–3844. USENIX Association. <https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag>

