A SURVEY ON CURRENT TRENDS AND RECENT ADVANCES IN TEXT ANONYMIZATION

Tobias Deußer[*,1,2]**, Lorenz Sparrenberg** [1]**, Armin Berger** [1,2]**, Max Hahnbück** [1,2]**, Christian Bauckhage** [1,2]**, and Rafet Sifa** [1,2]

[1]University of Bonn, Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin, Germany

ABSTRACT

The proliferation of textual data containing sensitive personal information across various domains requires robust anonymization techniques to protect privacy and comply with regulations, while preserving data usability for diverse and crucial downstream tasks. This survey provides a comprehensive overview of current trends and recent advances in text anonymization techniques. We begin by discussing foundational approaches, primarily centered on Named Entity Recognition, before examining the transformative impact of Large Language Models, detailing their dual role as sophisticated anonymizers and potent de-anonymization threats. The survey further explores domain-specific challenges and tailored solutions in critical sectors such as healthcare, law, finance, and education. We investigate advanced methodologies incorporating formal privacy models and risk-aware frameworks, and address the specialized subfield of authorship anonymization. Additionally, we review evaluation frameworks, comprehensive metrics, benchmarks, and practical toolkits for real-world deployment of anonymization solutions. This review consolidates current knowledge, identifies emerging trends and persistent challenges, including the evolving privacy-utility trade-off, the need to address quasi-identifiers, and the implications of LLM capabilities, and aims to guide future research directions for both academics and practitioners in this field.

Keywords Anonymization · Large Language Models · Named Entity Recognition · Natural Language Processing · Privacy · Trustworthy Machine Learning · Survey

1 Introduction

The digital age has led to an unprecedented generation and collection of textual data [1], from electronic health records and legal documents to social media posts and customer reviews. While this data holds immense value for research, analytics, and service improvement, it often contains sensitive personal information, posing significant privacy risks. Regulatory frameworks like GDPR [2, 3] mandate the protection of such data, making effective anonymization techniques indispensable.

Text anonymization aims to transform textual data in such a way that individuals cannot be re-identified, either directly or indirectly, while minimizing the loss of information utility for downstream tasks like contradiction detection [4], algorithmic trading [5], speech-based dementia detection [6], regulatory compliance verification [7], or legal contract analysis [8]. This involves complex challenges, including the accurate detection of diverse personally identifiable information (PII) types, handling contextual ambiguities, preserving semantic integrity, and balancing the often-competing goals of privacy and utility. The advent of sophisticated analytical tools, particularly Large Language Models (LLMs), has further complicated this landscape, offering both powerful new anonymization capabilities and potent de-anonymization threats.

---

[*]`tdeusser@uni-bonn.de`, ORCID-ID: 0000-0003-4685-0847

Our contributions are threefold. First, we provide a holistic survey that bridges foundational NER-based anonymization techniques with the emerging role of LLMs, highlighting their dual potential as both anonymization tools and de-anonymization threats. Second, we offer a broad synthesis of domain-specific challenges and solutions across healthcare, law, finance, and education, while also dedicating focused analysis to advanced privacy-preserving methodologies such as formal privacy models and authorship anonymization. Third, we emphasize the importance of robust evaluation, reviewing current benchmarks, metrics, and practical toolkits, and critically assessing the evolving trade-off between privacy and utility. This comprehensive perspective aims to guide both researchers and practitioners navigating the complex landscape of textual data anonymization.

The rest of this paper is structured as follows. We start by discussing foundational approaches, primarily centered around Named Entity Recognition (NER), which have long served as the cornerstone for PII identification. We then delve into the transformative impact of LLMs, examining their dual role as both sophisticated anonymizers and potential re-identification adversaries. Subsequent sections navigate the nuanced landscape of domain-specific anonymization challenges and tailored solutions, particularly in critical sectors like healthcare, finance, and law. We further explore advanced methodologies that incorporate formal privacy models (e.g., Differential Privacy) and risk-aware frameworks. The specialized subfield of authorship anonymization, focused on obscuring linguistic style, is also addressed. Finally, we emphasize the critical importance of robust evaluation frameworks, comprehensive metrics, and the development of practical toolkits and systems for real-world deployment. This review aims to consolidate current knowledge, highlight emerging trends and persistent challenges, and guide future research directions for both academics and practitioners in the field of anonymization.

2   Foundational and NER-Driven Anonymization

Named Entity Recognition (NER) has long been a cornerstone of automated text anonymization, serving as the primary mechanism for identifying explicit mentions of PII such as names, locations, organizations, and contact details [9]. Many early and ongoing efforts build upon NER, often augmenting it with rule-based systems, gazetteers, and traditional machine learning techniques.

For specific applications like call centers, [10] demonstrated the successful application of custom NER models (BiLSTM-CRF with contextual string embeddings) to noisy call-center transcripts for PII masking and privacy law compliance.

In educational contexts, [11] presented a method using set operations and filtering candidate private words (annotated manually or via ML) to redact PII in online discussion forums, achieving high recall. Similarly, [12] focused on anonymizing participant names in online discussions, including nicknames and spelling errors, using a pseudonymization process guided by class lists, NER, and heuristic rules, arguing that such guided methods can outperform deep neural networks for specific name variant challenges.

The development of practical tools often relies heavily on these foundational techniques. For instance, ANOPPI, a tool for semi-automatic anonymization of Finnish legal texts, combines rule-based, machine-learning-based, and gazetteer-based methods for NER, assigning identifiers for consistent pseudonym replacement [2]. Textwash, an open-source Python tool, also combines NER with pattern matching and fuzzy string matching to identify and mask PII in a language-agnostic manner [13].

These approaches highlight the importance of NER as a fundamental building block, often forming the first pass in more complex anonymization pipelines. However, they also underscore the limitations of relying solely on NER, particularly in handling implicit identifiers, contextual nuances, and ensuring comprehensive privacy protection against sophisticated re-identification attacks, paving the way for more advanced methodologies.

3   Anonymization for and via LLMs

The advent of LLMs has significantly impacted the field of text anonymization, offering new capabilities for both identifying and transforming sensitive information, as well as posing new challenges as potential de-anonymization tools.

Several works explore using LLMs directly as anonymizers. [14] and [15] investigated whether suitably prompted off-the-shelf LLMs (like GPT-3.5/4) can serve as effective zero-shot or few-shot text anonymizers. Their findings suggest that LLMs can remove or replace identifiable spans while preserving fluency, though consistency and completeness remain challenges. [16] proposed a novel approach to distill knowledge from LLMs into smaller, more resource-efficient encoder-only models for anonymization, using NER and regular expressions, thereby enabling deployment on less powerful devices.

In the medical domain, LLMs have shown particular promise. [17] introduced DeID-GPT, a framework leveraging GPT-4 for zero-shot de-identification of medical texts, reporting high accuracy in masking PHI while preserving text structure. [18] further evaluated GPT-3.5 and GPT-4 for clinical note de-identification, with GPT-4 demonstrating superior performance, highlighting its potential for safeguarding patient privacy.

LLMs are also being used to develop more robust and utility-preserving anonymization frameworks. [19] proposed RUPTA, a framework using LLMs as a privacy evaluator, utility evaluator, and optimizer to iteratively edit text for an optimal privacy-utility trade-off. [20] introduced IncogniText, an LLM-based method for conditional text anonymization that randomizes private attributes at the document level to prevent inference attacks while preserving meaning.

However, the power of LLMs also presents a new threat. [21] highlighted that LLMs like GPT-3.5 can act as effective "de-anonymizers," inferring identities from contextual information remaining after standard anonymization, calling for a re-evaluation of existing techniques. This dual role of LLMs—as both powerful anonymization tools and sophisticated adversaries—is a central theme in current research.

4 Domain-Specific Challenges and Solutions

Different domains present unique challenges for text anonymization due to varying data characteristics, types of sensitive information, and regulatory requirements. Researchers have developed tailored solutions for several key areas.

4.1 Healthcare and Clinical Notes

The healthcare domain is particularly sensitive, with strict regulations like HIPAA governing Protected Health Information (PHI). De-identifying clinical free text is crucial, and recent approaches have predominantly utilized machine learning. A systematic review by [22] covering 69 studies from 2010 to 2023 found that ML and hybrid (machine learning combined with rule-based) methods are the most common, with purely rule-based techniques becoming rare. This review also highlights ongoing challenges such as handling diverse data sources and balancing privacy with data utility.

Transformer-based models have shown significant promise. For instance, [23] presented Transformer-DeID, which employed models like BERT [24] and RoBERTa [25] for de-identifying clinical notes. Comparative evaluations, such as the one by [26], have explored the performance of various transformer architectures on benchmark datasets, often noting high overall accuracy but with performance variations for specific PHI categories. The generalizability of these models is a critical concern; [27] investigated the performance of pretrained de-identification transformers on narrative nursing notes, which can have different characteristics compared to other clinical texts like discharge summaries.

Several established tools and methodologies continue to be relevant. [28] introduced INCOGNITUS, a flexible platform for anonymizing clinical notes that offers multiple techniques, including one designed to guarantee 100% recall by substituting words with semantically similar ones, alongside a module for assessing information loss. [29] detailed a hybrid context-based model for the fully automated de-identification of over a billion clinical notes, which reportedly outperformed both NER-only models and commercial services. Ensemble learning approaches have also demonstrated effectiveness; [30] created a "best-in-class" automated de-identification tool for Electronic Health Records (EHRs) by combining rule-based methods, dictionary lookups, and ML models.

The advent of LLMs has opened new avenues. As previously noted, DeID-GPT [17] and the research by [18] specifically apply LLMs to the de-identification of clinical text. [31] conducted a comparative study investigating LLMs for clinical text anonymization, introducing novel evaluation metrics for generative anonymization and concluding that LLM-based methods are a reliable alternative to traditional approaches, achieving high accuracy while preserving utility. To address the poor cross-institutional generalization of de-identification models, [32] proposed using GPT-4 for privacy-safe data augmentation, generating synthetic clinical text (with PHI redacted before prompting) to improve F1 scores on unseen hospital data. Further exploring LLM utility, [33] introduced the LLM-Anonymizer, a tool using locally deployable LLMs to enhance privacy via on-premise processing. An alternative strategy involves "LLMs-in-the-loop" to develop expert small AI models for de-identification across multiple languages, which may offer advantages in accuracy and privacy over general-purpose LLMs [34].

However, de-identification alone may not suffice for robust privacy. [35] demonstrated that de-identified clinical notes can still be vulnerable to re-identification (e.g., via membership inference attacks) and explored state-of-the-art LLM-generated synthetic notes as an alternative, finding that while synthetic data can match real data utility, high-fidelity synthetic data can also carry similar privacy risks. Protecting privacy during the training of clinical language models is another critical area. [36] applied end-to-end pseudonymization to text data before training a Swedish Clinical BERT,

finding minimal to no performance loss on downstream NLP tasks compared to models trained on identified data, suggesting that training on fully de-identified data is feasible.

Finally, understanding the impact of these techniques is crucial. [37] systematically analyzed how various anonymization techniques affect downstream NLP task performance and the risk of re-identification in the medical domain, offering valuable guidance for practitioners.

4.2 Legal Documents

Legal documents, such as court decisions and parole hearing transcripts, often contain highly sensitive personal data, necessitating robust anonymization. The systematization of anonymizing court decisions has advanced, particularly with EU Member States' courts adopting algorithmic approaches to automate the process; these efforts also address inherent challenges like re-identification risks and ensuring system acceptance by court staff [38].

Several automated systems and methodologies have been developed for this domain. For example, [39] presented an automated process for anonymizing parole hearing transcripts in California, which reliably removes and pseudonymizes data while preserving structure. The ANOPPI tool [2], mentioned earlier, was specifically developed for Finnish legal texts. For German legal court decisions, an ML approach utilizing deep neural networks has been proposed for the automatic identification of sensitive text elements [40]. Efforts also extend to scanned documents, with systems designed for the automatic anonymization of images of Swiss Federal Supreme Court rulings [41] and more broadly for law enforcement documents, where machine learning is used to minimize manual effort and the extent of redacted areas [42]. Furthermore, [3] introduced PSILENCE, a pseudonymization tool for international arbitration documents that employs NER and coreference resolution for consistent entity replacement.

Highlighting the methodological needs, [43] reviewed the specific challenges of legal document anonymization, emphasizing that Named Entity Recognition (NER) alone is often insufficient and advocating for ML-based methods to complement it. The crucial issue of re-identification continues to be a research focus. The capabilities of Large Language Models (LLMs) to re-identify individuals in anonymized court decisions were assessed by [44], who found current risks to be low for well-anonymized Swiss court cases but acknowledged the potential future threat. Investigating deanonymization further, [45] introduced RedactBuster, a model that leverages sentence context to perform Named Entity Recognition on redacted documents. Tested on the Text Anonymization Benchmark (TAB), RedactBuster demonstrated high accuracy and also proposed countermeasures to enhance the privacy of redacted information.

4.3 Audio, Call Centers, and Spoken Conversations

Anonymizing spoken language introduces the complexity of Automatic Speech Recognition (ASR) errors and disfluencies. [46] presented a pipeline for anonymizing French audio data by using a forced aligner and an NER model to identify named entities in transcripts, then replacing corresponding audio segments with silence. For real-time applications, [47] introduced Trustera, a system that redacts PII in live spoken conversations in call centers by integrating ASR, NLU, and a live audio redactor. [10], as noted before, focused on applying NER to call center transcripts to aid privacy law compliance.

[48] explored the application of state-of-the-art NER algorithms to ASR-generated call center transcripts, creating models with low latency that can be readily integrated into existing pipelines. In the context of call center transcripts, [49] presented a process for anonymizing training data and a framework for assessing its effectiveness, comparing models fine-tuned on anonymized data with commercially available LLM APIs. [50] developed and evaluated an efficient speech de-identification system for Indonesian speech in low-resource transcripts, using speech recognition, information extraction, and masking.

4.4 Educational Data

Online learning environments generate vast amounts of student data, including discussion forum posts and essays, which require anonymization for research. [11] developed a method for discovering and redacting private information in online course forum texts. [12] specifically tackled the challenge of pseudonymizing names, nicknames, and spelling errors in student discussions. [51] explored a hybrid approach combining rule-based methods and a fine-tuned RoBERTa [25] model to de-identify student writing, aiming to balance privacy and data utility for educational researchers.

Beyond direct redaction of PII from text, the field is also exploring broader privacy-enhancing technologies for educational data, which often incorporates or is derived from sensitive textual student inputs. [52] highlighted the relatively low adoption of anonymization techniques in Learning Analytics despite their crucial role, while also demonstrating how such techniques can be effectively integrated. Recognizing that traditional anonymization methods

can be insufficient for the complexities of educational data, recent work has focused on more robust protection. For instance, [53] proposed a Differential Privacy framework specifically designed for LA, offering practical guidance for its implementation and validating its effectiveness in safeguarding data privacy against potential attacks, while also exploring the trade-offs between privacy and utility. Another established technique, K-anonymity, was applied by [54] to student data in Educational Data Mining patterns, showing its potential for protecting student privacy while noting the decrease in correlation coefficients that must be balanced with information loss. Furthermore, the generation of privacy-preserving synthetic educational data, as explored by [55], offers an alternative to anonymizing original datasets. This approach particularly addresses the risks of re-identification from naively pseudonymized data and includes an evaluation framework for comparing synthetic data generators and techniques to guarantee privacy.

4.5 Financial Reports

The financial sector, characterized by highly confidential data, presents a critical domain for textual anonymization. Efforts here aim to enable the use of financial documents for tasks like text classification and entity detection without compromising sensitive information. [56] specifically addressed the anonymization of German financial and legal documents by developing and evaluating methods based on neural network language models, including recurrent neural nets and transformer architectures. Their work also resulted in a web-based application for anonymizing such documents, demonstrating a practical approach to handling sensitive data. Building on the need for efficient solutions, [16] proposed a resource-efficient anonymization pipeline using knowledge distillation. This method transfers knowledge from large language models (LLMs) to smaller, more deployable encoder-only models, combined with named entity recognition and regular expressions, aiming to make anonymization scalable and accessible even without extensive computational resources or manually labeled data.

Beyond directly redacting information from documents, research also explores integrating privacy-preserving techniques into the analytical models themselves. For instance, [57] focused on privacy-enabled financial text classification. They proposed integrating Differential Privacy and Federated Learning with transformer-based models (BERT [24] and RoBERTa [25]) to train on sensitive financial data while preserving privacy, highlighting the crucial trade-offs between privacy and utility in this domain. These works underscore the dual challenge in the financial sector: robustly anonymizing textual data and ensuring that subsequent analytical processes maintain privacy.

5 Advanced Methodologies and Privacy-Preserving Techniques

Beyond foundational NER and direct LLM applications, a significant body of research focuses on developing more sophisticated anonymization methodologies. These often incorporate explicit privacy models, risk assessment, advanced machine learning, and techniques aimed at better preserving data utility while enhancing privacy.

One line of research aims to integrate explicit privacy risk measures into the anonymization process. [58] presented a three-step approach involving a privacy-enhanced entity recognizer, privacy-risk assessment measures (based on BERT [24], web search, or classifiers), and linear optimization to mask entities while minimizing semantic loss under a risk threshold. [59] proposed enhancing NER-based anonymization by using explainability techniques with a neural language model to iteratively detect and mask terms posing the greatest re-identification risk until a user-defined k-anonymity level is reached. [60] argued for moving beyond simple sequence labeling by incorporating such explicit disclosure risk measures. [61] presented a two-stage sanitization strategy using instruction-tuned LLMs to generate truth-preserving replacements for sensitive spans, then simulating inference attacks to select the most informative yet risk-resistant candidate.

Differential Privacy (DP) offers formal privacy guarantees and has been explored for text rewriting. [62] introduced DP-Rewrite, an open-source framework for reproducible DP text rewriting experiments. [63] later proposed DP-BART, a system for Local Differential Privacy (LDP) text rewriting that significantly improved utility over prior methods. [64] introduced DP-MLM, using a masked language model for DP text rewriting, claiming better context preservation and utility at lower privacy budgets. [65] presented DP-VAE, using a differentially private Variational Autoencoder to generate human-readable, privatized versions of online reviews.

Other novel approaches include bootstrapping anonymization models. [66] proposed using distant supervision from a knowledge graph to automatically annotate texts for k-anonymity, then fine-tuning a transformer model on this data. [67] explored word embedding-based anonymization, where risky terms are identified based on semantic similarity to a target profile and replaced with more general terms to preserve utility.

These advanced methodologies represent a shift towards more principled and robust anonymization, often striving for quantifiable privacy guarantees and a more nuanced balance between privacy protection and the utility of the anonymized data.

6 Authorship Anonymization

A distinct subfield within text anonymization is authorship anonymization, also known as authorship obfuscation. The goal here is not primarily to remove PII, but to modify the writing style of a text to prevent an author from being identified through linguistic patterns. This is crucial in contexts where an author wishes to remain anonymous, even if the content itself is not sensitive.

Several recent works have tackled this challenge using diverse techniques. [68] proposed MuCAAT, a multilingual contextualized authorship anonymization method for social media texts, designed to alter stylistic fingerprints while preserving the message. Reinforcement learning has emerged as a promising approach: [69] introduced TAROT, an unsupervised method using policy optimization to regenerate texts, fooling an author classifier while maintaining task utility.

Constrained decoding with smaller language models is another avenue. [70] presented JAMDEC, an inference-time algorithm that uses constrained decoding (e.g., with GPT-2 XL) to produce stylistic variations, outperforming prior methods with similar-sized models and even competing with much larger models in obfuscation effectiveness.

For languages other than English, research is also progressing. [71] evaluated GAN-based and sequence-to-sequence models for authorship obfuscation of Portuguese texts, finding the GAN approach offered a better trade-off between classifier fooling and content preservation.

Authorship anonymization addresses a different facet of privacy than PII redaction, focusing on the implicit "signature" an author leaves in their writing. The development of these techniques is vital for protecting authors in sensitive situations and for ensuring freedom of expression.

7 Evaluation Frameworks, Metrics, and Benchmarks

The ability to reliably evaluate the effectiveness of anonymization techniques is paramount. This involves assessing both the level of privacy protection achieved and the extent to which data utility is preserved. Recent research has focused on developing dedicated corpora, robust metrics, and realistic evaluation scenarios, including attacker models.

A significant contribution is The Text Anonymization Benchmark (TAB) [72], an open-source corpus of English-language court cases with comprehensive annotations of personal information, going beyond traditional de-identification by marking all spans needing masking. TAB also proposes evaluation metrics tailored for privacy protection and utility preservation.

However, traditional recall-based metrics have limitations. [73, 74] argued against relying solely on comparison to a single ground truth and proposed evaluating residual disclosure risk via an automated re-identification attack, formalized as a multi-class classification problem leveraging neural language models. This approach directly measures how anonymous a text truly is against an intelligent adversary. [75] also argued for better evaluation criteria, proposing TILD (Technical performance, Information Loss, De-identification by humans) to standardize assessment.

The impact of anonymization on downstream NLP tasks is another critical evaluation dimension. [76] investigated how different pseudonymization techniques affect text classification and summarization, exploring the trade-offs between data protection and utility.

Benchmarking different approaches is also crucial. [77] provided a comparative study of transformer-based models and LLMs against traditional architectures for text anonymization using the CoNLL-2003 dataset, highlighting relative strengths and weaknesses. [78] curated a challenging dataset with semi-synthetic sentences containing diverse PII types to expose performance gaps in widely used PII masking models, calling for better evaluation metrics and model transparency.

The re-identification capabilities of LLMs themselves are now part of the evaluation landscape. [44], as mentioned, assessed LLMs' ability to re-identify individuals in anonymized court decisions, providing insights into current risk levels. [59] also leverage re-identification risk within their enhancement methodology, implicitly evaluating against potential k-anonymity violations.

These efforts in developing robust evaluation frameworks are essential for advancing the field, ensuring that new anonymization methods provide genuine privacy protection while remaining practical for real-world applications.

8 Anonymization Toolkits and Systems

Beyond theoretical advancements and evaluation, the development of practical, usable toolkits and systems is paramount for deploying anonymization solutions in real-world scenarios. Several papers present such systems, often embodying the techniques discussed in previous sections.

For legal documents, ANOPPI [2] offers a semi-automatic anonymization tool for Finnish texts, available as both a web application with a UI for human review and a REST API. PSILENCE [3] is another tool focused on legal texts, specifically for pseudonymizing international arbitration documents in English, emphasizing consistent entity replacement. In the clinical domain, INCOGNITUS [28] provides a flexible platform for automated anonymization of clinical notes, incorporating multiple techniques and a performance-evaluation module. For spoken conversations, particularly in call centers, Trustera [47] is a system that redacts PII in real-time, preventing human agents from hearing sensitive details while capturing the PII for authorized uses.

More general-purpose tools and libraries have also been developed to support broader anonymization needs. Microsoft's Presidio [79] is a context-aware, pluggable, and customizable PII anonymization service designed for text and images, offering a robust framework for detecting and protecting sensitive information across various applications. PII-Codex [80] is a Python library designed for PII detection, categorization, and severity assessment, integrating with detection software to help users understand and manage the PII present in texts. Textwash [13] is another Python tool combining NER, pattern matching, and fuzzy string matching for language-agnostic text anonymization. For researchers working with differentially private text rewriting, DP-Rewrite [62] provides a modular and extensible open-source framework to facilitate reproducible experiments and ensure transparency.

9 Discussion and Future Directions

The landscape of text anonymization has evolved significantly, driven by advancements in NLP, particularly the rise of LLMs, and increasing regulatory and societal demands for data privacy. Several key trends and challenges emerge from the reviewed literature.

9.1 The Dual Role of LLMs

LLMs offer unprecedented capabilities for sophisticated anonymization, including nuanced understanding of context, fluent text generation for pseudonymization or generalization, zero-shot PII detection and transformation [14, 15, 17], and the leveraging of their knowledge for more efficient models [16]. Simultaneously, they represent powerful adversaries capable of re-identifying individuals from subtly anonymized texts [21, 44] and are one of the main reasons many strive to anonymize their data [16]. Future work must focus on developing LLM-based anonymization techniques that are robust against LLM-based attacks, potentially through adversarial training or by designing defenses that explicitly account for LLM inference capabilities and evaluation frameworks [19].

9.2 Balancing Privacy and Utility

This remains a central challenge. Overly aggressive anonymization can render data useless for downstream tasks, while insufficient anonymization leaves individuals vulnerable. In 2021, [60] emphasized moving beyond simple redaction to incorporate explicit disclosure risk measures . However, current techniques increasingly allow for a configurable trade-off, for example, by optimizing against a risk threshold [58] or a desired k-anonymity level [59]. Methods that preserve semantic structure and utility while ensuring privacy, such as through embedding-based generalization [67] or conditional randomization of private attributes [20], are crucial. Evaluating this balance requires comprehensive metrics and frameworks that capture both aspects effectively [72, 76].

9.3 Beyond Explicit PII

Many traditional methods focus on redacting explicit identifiers (names, addresses). However, quasi-identifiers, contextual information, and even linguistic style can also lead to re-identification. New methods are increasingly addressing this by considering broader re-identification risks from residual data [73], preventing attribute inference attacks [20], and modifying stylometric features to obscure authorship and prevent author re-identification [68, 70, 69]. This holistic view is essential for comprehensive privacy.

9.4 Formal Privacy Guarantees vs. Practical Anonymization

Techniques like Differential Privacy (DP), explored for text rewriting [63, 64, 62] and privacy-preserving data generation or analysis [65, 53, 57], offer formal, provable privacy guarantees. However, applying them effectively to complex, unstructured text while maintaining high utility can be challenging, often involving intricate model design and careful parameter tuning. Bridging the gap between theoretically sound privacy models and practical, high-utility text anonymization that is accessible to a wider range of practitioners remains an active area of research.

9.5 Domain Adaptation and Robustness

Anonymization models often require domain-specific tuning due to variations in PII types, language use, and regulatory contexts. Developing techniques that generalize well across diverse text types and languages with minimal labeled data, potentially through innovative data augmentation strategies [32] or knowledge distillation from larger models [16], is important for wider applicability across varied domains like healthcare [29, 22], legal [38], and finance [56]. The robustness of anonymization against noisy inputs (e.g., ASR errors in call center transcripts [10, 48], or informal language in user-generated text like educational forums [11]) also needs continuous improvement.

9.6 Evaluation and Benchmarking

The development of standardized benchmarks like TAB [72] and attacker-centric evaluation methodologies, such as simulating re-identification attacks [73, 74, 45] or using challenging datasets to expose model weaknesses [78], is vital. These approaches help move evaluation beyond simple recall against a single ground truth [75] towards a more realistic assessment of privacy risks. Future efforts should expand these resources to more languages and domains, and incorporate more sophisticated and adaptive attack models, including those leveraging LLMs. Transparency in model capabilities, limitations, and evaluation processes is also key for building trust and fostering responsible innovation [78].

9.7 Resource Efficiency and Accessibility

While powerful models achieve strong results, their computational cost can be prohibitive for many applications. Research into knowledge distillation [16] and the development of efficient, specialized smaller models [34] is important for deploying anonymization on edge devices or in resource-constrained environments. A growing ecosystem of open-source tools and platforms [13, 62, 2, 3, 79, 80] plays a crucial role in democratizing access to anonymization technologies, fostering reproducible research, and enabling practical adoption.

9.8 Open Research Questions

Despite these advancements, several critical challenges and open research questions persist, demanding further investigation:

- *Multilingual and Low-Resource Anonymization:* While many tools and techniques focus on high-resource languages like English, robust anonymization for a wider array of languages, especially low-resource ones, remains a significant hurdle. The lack of standardized multilingual benchmarks further complicates comparative evaluation and progress.

- *Tension between Formal Privacy and Textual Quality:* Formal methods like Differential Privacy offer strong guarantees but can sometimes lead to text that is unnatural, lacks fluency, or has significantly diminished utility. Striking a better balance between provable privacy and the preservation of semantic integrity and readability is crucial, particularly for generative anonymization techniques.

- *Dynamic and Adaptive Anonymization:* Current anonymization is often static. Developing systems that can dynamically adapt the level and type of anonymization based on context, data sensitivity, user permissions, or evolving threat models is an open area.

- *Explainability and Trustworthiness of Anonymization Models:* Especially with complex models like LLMs, understanding *why* certain information was (or was not) anonymized is important for trust and debugging. Methods for explaining anonymization decisions are needed.

- *Anonymization of Emerging Data Modalities:* While this survey focuses on text, the principles extend. However, specific challenges arise with multimodal data (text + image/audio), code, and other structured or semi-structured textual forms.

Addressing these multifaceted challenges will require continued interdisciplinary collaboration, innovative algorithmic development, and a commitment to robust, context-aware evaluation to ensure that the benefits of natural language processing and text data can be realized responsibly while safeguarding privacy and adhering to existing and future regulations.

10   Conclusion

In this survey, we detailed the landscape of textual data anonymization. We have traced the evolution from foundational NER-driven techniques to the transformative impact of Large Language Models. Our review spanned domain-specific challenges and solutions in areas such as healthcare, law, and finance, alongside explorations of advanced privacy-preserving methodologies like Differential Privacy and the specialized domain of authorship anonymization. Furthermore, we underscored the indispensable role of robust evaluation frameworks, metrics, benchmarks, and the development of practical toolkits for real-world applicability.

Several key themes have emerged: the profound influence of LLMs as both powerful anonymizers and sophisticated re-identification threats; the enduring challenge of achieving a nuanced balance between robust privacy protection and data utility; the expanding scope of anonymization to address implicit identifiers and stylometric features; and the critical need for rigorous, attacker-aware evaluation. The development of resource-efficient and accessible tools also remains paramount for broader adoption.

As the amount of textual data is only increasing [1], the imperative to protect individual privacy while still enabling beneficial data use will only intensify. Addressing this will likely steer future efforts towards several key technical advancements: developing anonymization techniques specifically hardened against sophisticated LLM-driven re-identification attacks; refining the privacy-utility trade-off with more dynamic, context-aware, and even personalized mechanisms; and creating more sophisticated methods to obscure not just explicit PII but also subtle quasi-identifiers and stylometric fingerprints that can compromise anonymity. Furthermore, enhancing the practical applicability of formal privacy models like Differential Privacy to unstructured text, improving the robustness of anonymization against noisy and diverse data sources, and fostering the development of more resource-efficient models for wider deployment are all critical avenues. Progress in these domains, alongside the continuous evolution of comprehensive evaluation benchmarks and robust, adaptive attacker models, will be paramount for the advancement of text anonymization into an even more reliable tool, capable of meeting the privacy challenges of the future.

Acknowledgments

References

[1] Kevin Bartley. Big data statistics – how much data is there in the world?, 2025. URL `https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/`. Accessed: 2025-05-15.

[2] Arttu Oksanen, Eero Hyvönen, Minna Tamper, Jouni Tuominen, Henna Ylimaa, Katja Löytynoja, Matti Kokkonen, and Aki Hietanen. An anonymization tool for open data publication of legal documents. In *International Workshop on Artificial Intelligence Technologies for Legal Documents*, 2022.

[3] Luis Adrián Cabrera-Diego and Akshita Gheewala. PSILENCE: A pseudonymization tool for international law. In *Proc. Workshop on CALD-Pseudo*, 2024.

[4] Tobias Deußer, David Leonhard, Lars Hillebrand, Armin Berger, Mohamed Khaled, Sarah Heiden, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, et al. Uncovering inconsistencies and contradictions in financial reports using large language models. In *Proc. BigData*, 2023.

[5] Xinli Yu, Zheng Chen, and Yanbin Lu. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In *Proc. EMNLP*, 2023.

[6] Tobias Deußer, Abdul Mohsin Siddiqi, Lorenz Sparrenberg, Tobias Adams, Christian Bauckhage, and Rafet Sifa. Fusing speech and language models for dementia detection. In *Proc. BigData*, 2024.

[7] Armin Berger, Lars Hillebrand, David Leonhard, Tobias Deußer, Thiago Bell Felix De Oliveira, Tim Dilmaghani, Mohamed Khaled, Bernd Kliem, Rudiger Loitz, Christian Bauckhage, et al. Towards automated regulatory compliance verification in financial auditing with large language models. In *Proc. BigData*, 2023.

[8] William Watson, Nicole Cho, Nishan Srishankar, Zhen Zeng, Lucas Cecchi, Daniel Scott, Suchetha Siddagangappa, Rachneet Kaur, Tucker Balch, and Manuela Veloso. LAW: Legal agentic workflows for custody and fund services contracts. In *Proc. COLING*, 2025.

[9] Tobias Deußer, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. Informed named entity recognition decoding for generative language models. In *Proc. BigData*, 2024.

[10] Micaela Kaplan. May I ask who's calling? Named Entity Recognition on call center transcripts for privacy law compliance. In *Proc. W-NUT*, 2020.

[11] Nigel Bosch, R. Wes Crues, Najmuddin Shaik, and Luc Paquette. "Hello, [REDACTED]": Protecting student privacy in analyses of online discussion forums. In *Proc. EDM 2020*, 2020.

[12] Elaine Farrow, Johanna D. Moore, and Dragan Gašević. Names, nicknames, and spelling errors: Protecting participant identity in learning analytics of online discussions. In *Proc. LAK 2023*, 2023.

[13] Bennett Kleinberg, Toby Davies, and Maximilian Mozes. Textwash – automated open-source text anonymisation, 2022. URL https://arxiv.org/abs/2208.13081.

[14] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Large language models are anonymizers. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

[15] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Language models are advanced anonymizers. In *Proc. ICLR*, 2025.

[16] Tobias Deußer, Max Hahnbück, Tobias Uelwer, Cong Zhao, Christian Bauckhage, and Rafet Sifa. Resource-efficient anonymization of textual data via knowledge distillation from large language models. In *Proc. COLING*, 2025.

[17] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. DeID-GPT: Zero-shot medical text de-identification by GPT-4. arXiv Preprint, March 2023. URL https://arxiv.org/abs/2303.11032.

[18] Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, and et al. Evaluating GPT models for clinical note de-identification. *Scientific Reports*, 2025.

[19] Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models, 2024. URL https://arxiv.org/abs/2407.11770.

[20] Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Incognitext: Privacy-enhancing conditional text anonymization via LLM-based private attribute randomization. In *Neurips Safe Generative AI Workshop 2024*, 2024.

[21] Constantinos Patsakis and Nikolaos Lykousas. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 2023.

[22] Aleksandar Kovačević, Bojana Bašaragin, Nikola Milošević, and Goran Nenadić. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial intelligence in medicine*, 2024.

[23] Callandra Moore, Lucas Bulgarelli, Tom Pollard, and Alistair Johnson. Transformer-deid: Deidentification of free-text clinical notes with transformers. PhysioNet, 2023.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, 2019.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.

[26] Christopher Meaney, Wali Hakimpour, Sumeet Kalia, and Rahim Moineddin. A comparative evaluation of transformer models for de-identification of clinical text data, 2022. URL https://arxiv.org/abs/2204.07056.

[27] Fangyi Chen, Syed Mohtashim Abbas Bokhari, Kenrick Cato, Gamze Gürsoy, and Sarah Rossetti. Examining the generalizability of pretrained de-identification transformer models on narrative nursing notes. *Applied Clinical Informatics*, 15(02):357–367, 2024.

[28] Bruno Ribeiro, Vitor Rolla, and Ricardo Santos. INCOGNITUS: A toolbox for automated clinical notes anonymization. In *Proc. EACL*, 2023.

[29] Veysel Kocaman, D Talby, and H Ul Hak. Rwd143 beyond accuracy: Automated de-identification of large real-world clinical text datasets. *Value in Health*, 2023.

[30] Karthik Murugadoss, Ajit Rajasekharan, Bradley Malin, Vineet Agarwal, Sairam Bade, Jeff R. Anderson, Jason L. Ross, William A. Faubion Jr., John D. Halamka, Venky Soundararajan, and Sankar Ardhanari. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, 2021.

[31] David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. Unlocking the potential of large language models for clinical text anonymization: A comparative study. In *Proc. Workshop on Privacy in Natural Language Processing*, 2024.

[32] Woojin Kim, Sungeun Hahm, and Jaejin Lee. Generalizing clinical de-identification models by privacy-safe data augmentation using gpt-4. In *Proc. EMNLP*, 2024.

[33] Isabella C. Wiest, Marie-Elisabeth Leßmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P. Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. Anonymizing medical documents with local, privacy preserving large language models: The llm-anonymizer. *medRxiv*, 2024.

[34] Murat Gunay, Bunyamin Keles, and Raife Hizlan. Llms-in-the-loop part 2: Expert small ai models for anonymization and de-identification of phi across multiple languages, 2024. URL https://arxiv.org/abs/2412.10918.

[35] Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific Reports*, 2024.

[36] Thomas Vakili, Aron Henriksson, and Hercules Dalianis. End-to-end pseudonymization of fine-tuned clinical bert models: Privacy preservation with maintained data utility. *BMC Medical Informatics and Decision Making*, 2024.

[37] Iyadh B. C. Larbi, Aljoscha Burchardt, and Roland Roller. Clinical text anonymization, its influence on downstream NLP tasks and the risk of re-identification. In *Proc. EACL*, 2023.

[38] Kalliopi Terzidou. Automated anonymization of court decisions: Facilitating the publication of court decisions through algorithmic systems. In *Proc. ICAIL*, 2023.

[39] Abed El Rahman Itani, Wassiliki Siskou, and Annette Hautli-Janisz. Automated anonymization of parole hearing transcripts. In *Proc. Natural Legal Language Processing Workshop*, 2024.

[40] Ingo Glaser, Tom Schamberger, and Florian Matthes. Anonymization of german legal court rulings. In *Proc. ICAIL*, 2021.

[41] Joel Niklaus, Robin Mamié, Matthias Stürmer, Daniel Brunner, and Marcel Gygli. Automatic anonymization of swiss federal supreme court rulings. In *Proc. Workshop Natural Legal Language Processing*, 2023.

[42] Manuel Eberhardinger, Patrick Takenaka, Daniel Grießhaber, and Johannes Maucher. Anonymization of documents for law enforcement with machine learning, 2025. URL https://arxiv.org/abs/2501.07334.

[43] Gergely M. Csányi, Dániel Nagy, Renátó Vági, J. Pál Vadász, and Tamás Orosz. Challenges and open problems of legal document anonymization. *Symmetry*, 2021.

[44] Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. Anonymity at risk? assessing re-identification capabilities of large language models in court decisions. In *Findings of the ACL: NAACL 2024*, 2024.

[45] Mirco Beltrame, Mauro Conti, Pierpaolo Guglielmin, Francesco Marchiori, and Gabriele Orazi. Redactbuster: Entity type recognition from redacted documents. In *Proc. ESORICS*, 2024.

[46] Guillaume Baril, Patrick Cardinal, and Alessandro Lameiras Koerich. Named entity recognition for audio de-identification. In *Proc. IJCNN 2022*, 2022.

[47] Evandro Gouvêa, Ali Dadgar, Shahab Jalalvand, Rathi Chengalvarayan, Badrinath Jayakumar, Ryan Price, Nicholas Ruiz, Jennifer McGovern, Srinivas Bangalore, and Ben Stern. Trustera: A live conversation redaction system. In *Proc. ICASSP*, 2023.

[48] Sam Davidson, Jordan Hosier, Yu Zhou, and Vijay Gurbani. Improved named entity recognition for noisy call center transcripts. In *Proc W-NUT*, 2021.

[49] Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In Elena Volodina, David Alfter, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, and Xuan-Son Vu, editors, *Proc CALD-pseudo 2024*, 2024.

[50] Rifqi Naufal Abdjul, Dessi Puji Lestari, Ayu Purwarianti, Candy Olivia Mawalim, Sakriani Sakti, and Masashi Unoki. Indonesian speech content de-identification in low resource transcripts. In *Proc. Workshop in South East Asian Language Processing*, 2025.

[51] Langdon Holmes, Scott A. Crossley, Wesley Morris, Harshvardhan Sikka, and Anne Trumbore. De-identifying student writing with rules and transformers. In *Proc. AIED*, 2023.

[52] Janneth Chicaiza, Ma Carmen Cabrera-Loayza, Rene Elizalde, and Nelson Piedra. Application of data anonymization in learning analytics. In *Pro. APPIS*, 2020.

[53] Qinyi Liu, Ronas Shakya, Mohammad Khalil, and Jelena Jovanovic. Advancing privacy in learning analytics using differential privacy. In *Proc. LAK*, 2025.

[54] Agung Triayudi, Iskandar Fitri, Sitti Rachmawati Yahya, and Sumiati Sumiati. Educational data mining patterns k-anonymity for the analytics of student privacy data. In *Proc. ICCoSITE*, 2023.

[55] Jill-Jênn Vie, Tomas Rigaux, and Sein Minn. Privacy-preserving synthetic educational data generation. In *Proc. EATL*, 2022.

[56] David Biesner, Rajkumar Ramamurthy, Robin Stenzel, Max Lübbering, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. Anonymization of german financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, 2022.

[57] Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, and Zumrut Muftuoglu. Privacy enabled financial text classification using differential privacy and federated learning. In *Proc. Workshop ECONLP*, 2021.

[58] Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. Neural text sanitization with explicit measures of privacy risk. In *Proc. AACL-IJCNLP*, 2022.

[59] Benet Manzanares-Salor and David Sánchez. Enhancing text anonymization via re-identification risk-based explainability. *Knowledge-Based Systems*, 2025.

[60] Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In *Proc. ACL-IJCNLP*, 2021.

[61] Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Truthful text sanitization guided by inference attacks, 2024. URL https://arxiv.org/abs/2412.12928.

[62] Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. DP-Rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proc. COLING 2022*, 2022.

[63] Timour Igamberdiev and Ivan Habernal. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the ACL: ACL 2023*, 2023.

[64] Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the ACL: ACL 2024*, 2024.

[65] Benjamin Weggenmann and Valentin Hartmann. DP-VAE: Human-readable text anonymization for online reviews with differentially private VAE. In *Proc. WWW*, 2022.

[66] Anthi Papadopoulou, Pierre Lison, Lilja Øvrelid, and Ildikó Pilán. Bootstrapping text anonymization models with distant supervision. In *Proc. LREC*, 2022.

[67] Fadi Hassan, David Sánchez, and Josep Domingo-Ferrer. Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[68] Vladimir Panov, Mikhail Kovalchuk, Anastasiia Filatova, and Sergey Teryoshkin. MuCAAT: Multilingual contextualized authorship anonymization of texts from social networks. *Procedia Computer Science*, 2022.

[69] Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. TAROT: Task-oriented authorship obfuscation using policy optimization methods. In *Proc. PrivateNLP Workshop*, 2025.

[70] Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. JAMDEC: Unsupervised authorship obfuscation using constrained decoding over small language models. In *Proc. NAACL-HLT*, June 2024.

[71] Antônio M. R. Franco, Ítalo S. Cunha, Leonardo B. Souza, and et al. Evaluation of deep neural network architectures for authorship obfuscation of portuguese texts. *Natural Language Processing Journal*, 2024.

[72] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 2022.

[73] Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. *Data Mining and Knowledge Discovery*, 2024.

[74] Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Automatic evaluation of disclosure risks of text anonymization methods. In *Proc. PSD*. Springer, 2022.

[75] Maximilian Mozes and Bennett Kleinberg. No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization, 2021. URL https://arxiv.org/abs/2103.09263.

[76] Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proce. Workshop TrustNLP*, 2023.

[77] Dimitris Asimopoulos, Ilias Siniosoglou, Vasileios Argyriou, Thomai Karamitsou, Eleftherios Fountoukidis, Sotirios K. Goudos, Ioannis D. Moscholios, Konstantinos E. Psannis, and Panagiotis Sarigiannidis. Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches. In *Proc. MOCAST*, 2024.

[78] Devansh Singh and Sundaraparipurnan Narayanan. Unmasking the reality of pii masking models: Performance gaps and the call for accountability, 2025. URL `https://arxiv.org/abs/2504.12308`.

[79] Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images, 2018. URL `https://microsoft.github.io/presidio`.

[80] Eidan J Rosado. Pii-codex: a python library for pii detection, categorization, and severity assessment. *The Journal of Open Source Software*, 2023.

[81] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models, 2025. URL `https://arxiv.org/abs/2312.11805`.

[82] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.