a watermark for large language models

A Watermark for Large Language Models: Protecting Intellectual Property in the Age of AI

Author: Dr. Anya Sharma, PhD in Computer Science, specializing in AI security and data provenance at Stanford University. Dr. Sharma has published extensively on the topic of AI safety and has consulted for several leading tech companies on the development and implementation of robust security measures for large language models (LLMs).

Publisher: The AI Ethics Journal, a leading peer-reviewed publication focusing on the ethical implications of artificial intelligence and its societal impact. The journal is widely respected within the academic and industry communities working on AI governance and responsible AI development.

Editor: Professor David Chen, PhD, Professor of Computer Science and Ethics at MIT. Professor Chen's expertise lies in the intersection of AI, law, and ethics.

Keywords: A watermark for large language models, LLM watermarking, AI watermarking, copyright protection, AI intellectual property, detection of AI-generated content, AI provenance, responsible AI, AI safety, deepfake detection.

Introduction: The Urgent Need for a Watermark for Large Language Models

The rapid advancement of large language models (LLMs) has ushered in a new era of technological possibilities, but it has also created significant challenges. One of the most pressing concerns is the potential for misuse, particularly the unauthorized reproduction and distribution of copyrighted material generated by LLMs. The lack of effective mechanisms to identify and track AI-generated content creates a pressing need for a watermark for large language models. This article explores the concept of a watermark for large language models, examining various approaches, technical challenges, and ethical implications.

Understanding the Problem: The Difficulty of Detecting AI-Generated Content

LLMs are remarkably adept at mimicking human writing styles. This capability, while beneficial for many applications, poses a serious threat to intellectual property rights. AI-generated content can easily be passed off as human-created work, leading to plagiarism, copyright infringement, and the spread of misinformation. The ability to reliably identify the origin of text, especially content generated by an LLM, is crucial. The development of a watermark for large language models is therefore not merely a technical challenge but a crucial step in ensuring responsible AI development.

Approaches to Implementing a Watermark for Large Language Models

Several approaches are being explored for implementing a watermark for large language models:

Statistical Watermarking: This technique involves subtly altering the statistical properties of the generated text, creating a hidden signature that can be detected using statistical analysis. These alterations might be in word choice, sentence structure, or the frequency of specific n-grams. The challenge lies in designing watermarks that are robust enough to withstand post-processing modifications (e.g., paraphrasing, translation) while remaining imperceptible to the human reader. The effectiveness of this approach for a watermark for large language models is still under investigation.

Semantic Watermarking: This approach focuses on embedding information within the meaning and structure of the text rather than its statistical properties. This might involve incorporating specific keywords or phrases in a carefully concealed manner, or using a specific narrative structure that reveals the AI origin upon close analysis. However, this approach is susceptible to sophisticated manipulation techniques, making it challenging to create a truly robust a watermark for large language models using semantic methods.

Cryptographic Watermarking: This approach employs cryptographic techniques to embed a digital signature directly within the generated text. This signature can then be extracted to verify the source and authenticity of the content. This method, while promising in its robustness, requires careful design to avoid compromising the readability or naturalness of the output. The use of cryptography for a watermark for large language models warrants further research into efficiency and resilience.

Model-Specific Signatures: Each LLM possesses unique characteristics in its output. Researchers are exploring methods to identify these unique signatures, effectively using the model itself as a watermark. This approach might involve analyzing the probabilistic distribution of word choices, or identifying patterns in the LLM's internal representations. However, the detection of these model-specific signatures is complicated by the continuous evolution and adaptation of LLMs.

Technical Challenges and Limitations

Implementing a truly effective a watermark for large language models presents several significant technical challenges:

Robustness: Watermarks must be resilient to various attacks, including paraphrasing, translation, and deliberate attempts to remove or obfuscate the watermark.
Imperceptibility: Watermarks must be undetectable to the human reader, preventing users from easily removing or modifying them.
Scalability: The watermarking system must be capable of handling large volumes of text data efficiently.
Privacy: The watermarking technique should not compromise the user's privacy.

Ethical Considerations and Societal Implications

The use of a watermark for large language models also raises significant ethical considerations:

Privacy concerns: The watermarking system must not inadvertently collect or reveal sensitive information about the users.
Freedom of expression: Watermarking should not unduly restrict users' freedom of expression or creativity.
False positives: The system must minimize false positives, avoiding the misidentification of human-created content as AI-generated.
Accessibility: The watermarking system should be accessible and usable by individuals and organizations of all sizes.

Conclusion: The Path Forward

The development of a robust and reliable a watermark for large language models is crucial for addressing the growing challenges related to intellectual property protection and the responsible use of AI. While significant technical and ethical hurdles remain, ongoing research and development offer hope for creating effective solutions. A multidisciplinary approach, involving computer scientists, ethicists, legal experts, and policymakers, is essential to ensure the development and deployment of watermarking technologies that are both effective and ethically sound. Collaboration and standardization will be key to achieving a widespread adoption of a watermark for large language models, ultimately safeguarding creativity and intellectual property in the age of AI.

FAQs

1. What is the primary purpose of a watermark for large language models? The primary purpose is to identify and track AI-generated content, protecting intellectual property and preventing unauthorized reproduction and distribution.

1. How does statistical watermarking differ from semantic watermarking? Statistical watermarking alters statistical properties of text, while semantic watermarking embeds information within the meaning and structure.

1. What are the main challenges in developing a robust watermark? Challenges include robustness to attacks, imperceptibility to the human reader, scalability, and privacy concerns.

1. Can watermarks be removed or bypassed? While highly robust watermarks are the goal, sophisticated techniques might eventually overcome some watermarking methods. Ongoing research aims to create increasingly resilient techniques.

1. What are the ethical implications of using watermarks? Concerns include potential restrictions on freedom of expression, privacy violations, and the risk of false positives.

1. Are there legal implications associated with watermarking AI-generated content? The legal landscape is still evolving. Watermarking could become a key factor in copyright infringement cases, but legal frameworks are still being developed.

1. How can a watermark be made imperceptible to the human reader? This involves subtle alterations to the text that do not significantly impact readability while embedding information detectable only through algorithmic analysis.

1. What role does cryptography play in AI watermarking? Cryptographic techniques offer a strong way to embed tamper-evident signatures, although careful design is needed to maintain text readability.

1. Is there a single "best" method for watermarking LLMs? There is no single best method yet. Different approaches have strengths and weaknesses, and an ideal solution might involve combining multiple techniques.

Related Articles:

1. "Detecting AI-Generated Text: A Survey of Current Techniques": This article reviews existing methods for identifying AI-generated text, highlighting their strengths, weaknesses, and applicability to different LLM architectures.

1. "The Ethical Implications of Watermarking AI-Generated Content": This paper delves into the ethical considerations surrounding AI watermarking, exploring potential biases, privacy concerns, and impacts on freedom of expression.

1. "A Novel Statistical Watermarking Scheme for Large Language Models": This research paper

details a new statistical watermarking algorithm specifically designed for LLMs, evaluating its robustness and imperceptibility.

1. "Cryptographic Watermarking for Secure Text Transmission: Application to AI-Generated Content": This article discusses the use of cryptography for securing AI-generated content, exploring various cryptographic techniques and their suitability for watermarking.

1. "Semantic Watermarking for Robust Copyright Protection of AI-Generated Text": This paper explores the challenges and opportunities of using semantic methods to create robust and tamper-proof watermarks for LLMs.

1. "Model-Specific Signatures for Tracing the Origin of AI-Generated Content": This research focuses on identifying and using unique characteristics of LLMs as a form of watermark.

1. "The Legal Landscape of AI-Generated Content and Copyright Protection": This article examines the current legal framework surrounding copyright and AI-generated content, discussing the implications of watermarking for legal disputes.

1. "Combating Deepfakes and Misinformation: The Role of AI Watermarking": This article explores the broader application of AI watermarking in combating the spread of deepfakes and other forms of AI-generated misinformation.

1. "A Comparative Analysis of Different Watermarking Techniques for LLMs": This article provides a detailed comparison of various LLM watermarking approaches, helping readers to understand their relative advantages and disadvantages.

a watermark for large language models: Large Language Models in Cybersecurity Andrei Kucharavy, 2024 This open access book provides cybersecurity practitioners with the knowledge needed to understand the risks of the increased availability of powerful large language models (LLMs) and how they can be mitigated. It attempts to outrun the malicious attackers by anticipating what they could do. It also alerts LLM developers to understand their work's risks for cybersecurity and provides them with tools to mitigate those risks. The book starts in Part I with a general

introduction to LLMs and their main application areas. Part II collects a description of the most salient threats LLMs represent in cybersecurity, be they as tools for cybercriminals or as novel attack surfaces if integrated into existing software. Part III focuses on attempting to forecast the exposure and the development of technologies and science underpinning LLMs, as well as macro levers available to regulators to further cybersecurity in the age of LLMs. Eventually, in Part IV, mitigation techniques that should allowsafe and secure development and deployment of LLMs are presented. The book concludes with two final chapters in Part V, one speculating what a secure design and integration of LLMs from first principles would look like and the other presenting a summary of the duality of LLMs in cyber-security. This book represents the second in a series published by the Technology Monitoring (TM) team of the Cyber-Defence Campus. The first book entitled Trends in Data Protection and Encryption Technologies appeared in 2023. This book series provides technology and trend anticipation for government, industry, and academic decision-makers as well as technical experts.

a watermark for large language models: The Developer's Playbook for Large Language Model Security Steve Wilson, 2024-09-03 Large language models (LLMs) are not just shaping the trajectory of AI, they're also unveiling a new era of security challenges. This practical book takes you straight to the heart of these threats. Author Steve Wilson, chief product officer at Exabeam, focuses exclusively on LLMs, eschewing generalized AI security to delve into the unique characteristics and vulnerabilities inherent in these models. Complete with collective wisdom gained from the creation of the OWASP Top 10 for LLMs list—a feat accomplished by more than 400 industry experts—this guide delivers real-world guidance and practical strategies to help developers and security teams grapple with the realities of LLM applications. Whether you're architecting a new application or adding AI features to an existing one, this book is your go-to resource for mastering the security landscape of the next frontier in AI. You'll learn: Why LLMs present unique security challenges How to navigate the many risk conditions associated with using LLM technology The threat landscape pertaining to LLMs and the critical trust boundaries that must be maintained How to identify the top risks and vulnerabilities associated with LLMs Methods for deploying defenses to protect against attacks on top vulnerabilities Ways to actively manage critical trust boundaries on your systems to ensure secure execution and risk minimization

a watermark for large language models: *Introduction to Generative AI* Numa Dhamani, Maggie Engler, 2024-02-27 Generative AI tools like ChatGPT are amazing—but how will their use impact our society? This book introduces the world-transforming technology and the strategies you need to use generative AI safely and effectively. Introduction to Generative AI gives you the hows-and-whys of generative AI in accessible language. In this easy-to-read introduction, you'll learn: How large language models (LLMs) work How to integrate generative AI into your personal and professional workflows Balancing innovation and responsibility The social, legal, and policy landscape around generative AI Societal impacts of generative AI Where AI is going Anyone who uses ChatGPT for even a few minutes can tell that it's truly different from other chatbots or question-and-answer tools. Introduction to Generative AI guides you from that first eye-opening interaction to how these powerful tools can transform your personal and professional life. In it, you'll get no-nonsense guidance on generative AI fundamentals to help you understand what these models are (and aren't) capable of, and how you can use them to your greatest advantage. Foreword by Sahar Massachi. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Generative AI tools like ChatGPT, Bing, and Bard have permanently transformed the way we work, learn, and communicate. This delightful book shows you exactly how Generative AI works in plain, jargon-free English, along with the insights you'll need to use it safely and effectively. About the book Introduction to Generative AI guides you through benefits, risks, and limitations of Generative AI technology. You'll discover how AI models learn and think, explore best practices for creating text and graphics, and consider the impact of AI on society, the economy, and the law. Along the way, you'll practice strategies for getting accurate responses and even understand how to handle misuse and security threats. What's inside How large

language models work Integrate Generative AI into your daily work Balance innovation and responsibility About the reader For anyone interested in Generative AI. No technical experience required. About the author Numa Dhamani is a natural language processing expert working at the intersection of technology and society. Maggie Engler is an engineer and researcher currently working on safety for large language models. The technical editor on this book was Maris Sekar. Table of Contents 1 Large language models: The power of AI Evolution of natural language processing 2 Training large language models 3 Data privacy and safety with LLMs 4 The evolution of created content 5 Misuse and adversarial attacks 6 Accelerating productivity: Machine-augmented work 7 Making social connections with chatbots 8 What's next for AI and LLMs 9 Broadening the horizon: Exploratory topics in AI

a watermark for large language models: Advances in Data-Driven Computing and Intelligent Systems Swagatam Das,

a watermark for large language models: Natural Language Processing and Chinese Computing Fei Liu, Nan Duan, Qingting Xu, Yu Hong, 2023-10-07 This three-volume set constitutes the refereed proceedings of the 12th National CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2023, held in Foshan, China, during October 12–15, 2023. The 143 regular papers included in these proceedings were carefully reviewed and selected from 478 submissions. They were organized in topical sections as follows: dialogue systems; fundamentals of NLP; information extraction and knowledge graph; machine learning for NLP; machine translation and multilinguality; multimodality and explainability; NLP applications and text mining; question answering; large language models; summarization and generation; student workshop; and evaluation workshop.

a watermark for large language models: Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track Gianmarco De Francisci Morales,

a watermark for large language models: Navigating AI in Academic Libraries: Implications for Academic Research Sacco, Kathleen, Norton, Alison, Arms, Kevin, 2024-09-25 Today's research scholars face the problem of how to effectively navigate the transformative impact of Artificial Intelligence (AI) while maintaining ethical integrity and scholarly rigor. AI technologies have permeated every aspect of scholarly inquiry, from information retrieval to research methodologies. As such, scholars grapple with the ethical implications, challenges, and opportunities presented by this technological revolution. Plagiarism, bias, and copyright issues in AI-assisted research threaten to undermine the integrity of academic scholarship. Navigating AI in Academic Libraries: Implications for Academic Research is presented as a groundbreaking solution to the complex challenges posed by AI integration in academia. This comprehensive volume serves as a guide for scholars seeking to navigate the intricacies of AI while upholding ethical standards and scholarly integrity. By addressing critical issues such as plagiarism detection, bias mitigation, and copyright concerns, the book equips scholars with the tools and strategies needed to harness the full potential of AI for academic inquiry.

a watermark for large language models: Towards AI-Aided Invention and Innovation Denis Cavallucci, Pavel Livotov, Stelian Brad, 2023-08-29 This book constitutes the proceedings of the 23rd International TRIZ Future Conference on Towards AI-Aided Invention and Innovation, TFC 2023, which was held in Offenburg, Germany, during September 12–14, 2023. The event was sponsored by IFIP WG 5.4. The 43 full papers presented in this book were carefully reviewed and selected from 80 submissions. The papers are divided into the following topical sections: AI and TRIZ; sustainable development; general vision of TRIZ; TRIZ impact in society; and TRIZ case studies.

a watermark for large language models: *Application of Large Language Models (LLMs) for Software Vulnerability Detection* Omar, Marwan, Zangana, Hewa Majeed, 2024-11-01 Large Language Models (LLMs) are redefining the landscape of cybersecurity, offering innovative methods for detecting software vulnerabilities. By applying advanced AI techniques to identify and predict weaknesses in software code, including zero-day exploits and complex malware, LLMs provide a

proactive approach to securing digital environments. This integration of AI and cybersecurity presents new possibilities for enhancing software security measures. Application of Large Language Models (LLMs) for Software Vulnerability Detection offers a comprehensive exploration of this groundbreaking field. These chapters are designed to bridge the gap between AI research and practical application in cybersecurity, in order to provide valuable insights for researchers, AI specialists, software developers, and industry professionals. Through real-world examples and actionable strategies, the publication will drive innovation in vulnerability detection and set new standards for leveraging AI in cybersecurity.

a watermark for large language models: Ki: Text Gerhard Schreiber, Lukas Ohly, 2024 Wenn Künstliche Intelligenz (KI) Texte generieren kann, was sagt das darüber, was ein Text ist? Worin unterscheiden sich von Menschen geschriebene und mittels KI generierte Texte? Welche Erwartungen, Befürchtungen und Hoffnungen hegen Wissenschaften, wenn in ihren Diskursen KI-generierte Texte rezipiert werden und Anerkennung finden, deren Urheberschaft und Originalität nicht mehr eindeutig definierbar sind? Wie verändert sich die Arbeit mit Quellen und welche Konsequenzen ergeben sich daraus für die Kriterien wissenschaftlicher Textarbeit und das Verständnis von Wissenschaft insgesamt? Welche Chancen, Grenzen und Risiken besitzen KI-Textgeneratoren aus Sicht von Technikethik und Technikfolgenabschätzung? Und welche Kriterien sind für einen ethisch bewussten Umgang mit KI-Textgeneratoren im Bildungskontext anzulegen? Diese und ähnliche Fragen verdeutlichen die Komplexität und Tragweite der Herausforderung, die der stetig wachsende und zunehmend selbstverständliche Einsatz von KI-Textgeneratoren für das zukünftige Arbeiten mit und an Texten mit sich bringt. Hierbei kristallisieren sich Problematiken heraus, die sich an ein breites Spektrum von Disziplinen wenden, deren spezifischen Diskurse in diesem Band beleuchtet werden.

a watermark for large language models: *Ethics in Online AI-Based Systems* Santi Caballé, Joan Casas-Roma, Jordi Conesa, 2024-04-10 Recent technological advancements have deeply transformed society and the way people interact with each other. Instantaneous communication platforms have allowed connections with other people, forming global communities, and creating unprecedented opportunities in many sectors, making access to online resources more ubiquitous by reducing limitations imposed by geographical distance and temporal constrains. These technological developments bear ethically relevant consequences with their deployment, and legislations often lag behind such advancements. Because the appearance and deployment of these technologies happen much faster than legislative procedures, the way these technologies affect social interactions have profound ethical effects before any legislative regulation can be built, in order to prevent and mitigate those effects. Ethics in Online AI-Based Systems: Risks and Opportunities in Current Technological Trends features a series of reflections from experts in different fields on potential ethically relevant outcomes that upcoming technological advances could bring about in our society. Creating a space to explore the ethical relevance that technologies currently still under development could have constitutes an opportunity to better understand how these technologies could or should not be used in the future in order to maximize their ethically beneficial outcomes, while avoiding potential detrimental effects. Stimulating reflection and considerations with respect to the design, deployment and use of technology will help guide current and future technological advancements from an ethically informed position in order to ensure that, tomorrow, such advancements could contribute towards solving current global and social challenges that we, as a society, have today. This will not only be useful for researchers and professional engineers, but also for educators, policy makers, and ethicists. - Investigates how intelligent technological advances might be used, how they will affect social interactions, and what ethical consequences they might have for society - Identifies and reflects on questions that need to be asked before the design, deployment, and application of upcoming technological advancements, aiming to both prevent and mitigate potential risks, as well as to identify potentially ethically-beneficial opportunities - Recognizes the huge potential for ethically-relevant outcomes that technological advancements have, and take proactive steps to anticipate that they be designed from an ethically-informed position - Provides reflections that

highlight the importance of the relationship between technology, their users and our society, thus encouraging informed design and educational and legislative approaches that take this relationship into account

a watermark for large language models: Advances in Cryptology - CRYPTO 2024 Leonid Reyzin, Douglas Stebila, 2024 Zusammenfassung: The 10-volume set, LNCS 14920-14929 constitutes the refereed proceedings of the 44th Annual International Cryptology Conference, CRYPTO 2024. The conference took place at Santa Barbara, CA, USA, during August 18-22, 2024. The 143 full papers presented in the proceedings were carefully reviewed and selected from a total of 526 submissions. The papers are organized in the following topical sections: Part I: Digital signatures; Part II: Cloud cryptography; consensus protocols; key exchange; public key encryption; Part III: Public-key cryptography with advanced functionalities; time-lock cryptography; Part IV: Symmetric cryptanalysis; symmetric cryptograph; Part V: Mathematical assumptions; secret sharing; theoretical foundations; Part VI: Cryptanalysis; new primitives; side-channels and leakage; Part VII: Quantum cryptography; threshold cryptography; Part VIII: Multiparty computation; Part IX: Multiparty computation; private information retrieval; zero-knowledge; Part X: Succinct arguments.

a watermark for large language models: Fen Eğitimi Araştırmalarına Güncel Bakış - VIII Ali GÜL, Semra BENZER, 2023-07-13

a watermark for large language models: Digital Watermarking for Machine Learning Model Lixin Fan, Chee Seng Chan, Qiang Yang, 2023-05-29 Machine learning (ML) models, especially large pretrained deep learning (DL) models, are of high economic value and must be properly protected with regard to intellectual property rights (IPR). Model watermarking methods are proposed to embed watermarks into the target model, so that, in the event it is stolen, the model's owner can extract the pre-defined watermarks to assert ownership. Model watermarking methods adopt frequently used techniques like backdoor training, multi-task learning, decision boundary analysis etc. to generate secret conditions that constitute model watermarks or fingerprints only known to model owners. These methods have little or no effect on model performance, which makes them applicable to a wide variety of contexts. In terms of robustness, embedded watermarks must be robustly detectable against varying adversarial attacks that attempt to remove the watermarks. The efficacy of model watermarking methods is showcased in diverse applications including image classification, image generation, image captions, natural language processing and reinforcement learning. This book covers the motivations, fundamentals, techniques and protocols for protecting ML models using watermarking. Furthermore, it showcases cutting-edge work in e.g. model watermarking, signature and passport embedding and their use cases in distributed federated learning settings.

a watermark for large language models: Teaching Economics Online Abdullah Al-Bahrani, Parama Chaudhury, Brandon J. Sheridan, 2024-08-06 In the light of the Covid-19 pandemic, this book is not only timely but essential reading, providing valuable insight into teaching economics both online and in a blended online/in person format. Diverse in scope, Teaching Economics Online combines past experience with innovative ideas on how to design teaching and improve the overall learning experience whilst remaining inclusive, effective and resilient.

a watermark for large language models: Kreative Intelligenz Mario Herger, 2023-11-30 Über ChatGPT hat man viel gelesen in der letzten Zeit: die künstliche Intelligenz, die ganze Bücher schreiben kann und der bereits jetzt unterstellt wird, Legionen von Autoren, Textern und Übersetzern arbeitslos zu machen. Und ChatGPT ist nicht allein, die KI-Familie wächst beständig. So malt DALL-E Bilder, Face Generator simuliert Gesichter und MusicLM komponiert Musik. Was erleben wir da? Das Ende der Zivilisation oder den Beginn von etwas völlig Neuem? Zukunftsforscher Dr. Mario Herger ordnet die neuesten Entwicklungen aus dem Silicon Valley ein und zeigt auf, welche teils bahnbrechenden Veränderungen unmittelbar vor der Tür stehen.

a watermark for large language models: *Digital Watermarking* Mohammad Ali Nematollahi, Chalee Vorakulpipat, Hamurabi Gamboa Rosales, 2016-08-08 This book presents the state-of-the-arts application of digital watermarking in audio, speech, image, video, 3D mesh graph, text, software,

natural language, ontology, network stream, relational database, XML, and hardware IPs. It also presents new and recent algorithms in digital watermarking for copyright protection and discusses future trends in the field. Today, the illegal manipulation of genuine digital objects and products represents a considerable problem in the digital world. Offering an effective solution, digital watermarking can be applied to protect intellectual property, as well as fingerprinting, enhance the security and proof-of-authentication through unsecured channels.

a watermark for large language models: Technik sozialisieren? / Technology Socialisation? Bruno Gransche, Jacqueline Bellon, Sebastian Nähr-Wagener, 2024-08-13 Anwendungen Künstlicher Intelligenz, Maschinellen Lernens sowie Robotik und weitere informatische Systeme spielen in immer mehr Bereichen der menschlichen Lebenswelt eine immer wichtigere Rolle. Technik wird dabei auch weiterhin und zunehmend Medium menschlicher Kommunikation und Interaktion sein, darüber hinaus wird jedoch auch immer mehr menschliche Interaktion nicht nur durch sondern mit Technik stattfinden. Eine Dimension neuer Mensch-Technik-Relationen ist dabei das Phänomen sozialer Angemessenheit. Obgleich sich dieses nicht auf ein einfaches Regelwerk reduzieren lässt, verhalten sich Menschen in zwischenmenschlicher Interaktion ganz selbstverständlich sozial angemessen und treffen in der Regel ohne Weiteres ‚den richtigen Ton'. Wie aber verhält es sich mit ‚intelligenten' technischen Systemen? Können – sollten – diese mit Fähigkeiten zu entsprechendem Sozialverhalten ausgestattet werden, um damit auch eine weitere bedeutsame Grenze zwischen Mensch und Technik zum Verschwinden zu bringen? Anders gefragt: Ist es möglich und erforderlich, Technik zu sozialisieren?

a watermark for large language models: Hardware Security Mark Tehranipoor,

a watermark for large language models: Who Wrote This? Naomi S. Baron, 2023-09-12 Would you read this book if a computer wrote it? Would you even know? And why would it matter? Today's eerily impressive artificial intelligence writing tools present us with a crucial challenge: As writers, do we unthinkingly adopt AI's time-saving advantages or do we stop to weigh what we gain and lose when heeding its siren call? To understand how AI is redefining what it means to write and think, linguist and educator Naomi S. Baron leads us on a journey connecting the dots between human literacy and today's technology. From nineteenth-century lessons in composition, to mathematician Alan Turing's work creating a machine for deciphering war-time messages, to contemporary engines like ChatGPT, Baron gives readers a spirited overview of the emergence of both literacy and AI, and a glimpse of their possible future. As the technology becomes increasingly sophisticated and fluent, it's tempting to take the easy way out and let AI do the work for us. Baron cautions that such efficiency isn't always in our interest. As AI plies us with suggestions or full-blown text, we risk losing not just our technical skills but the power of writing as a springboard for personal reflection and unique expression. Funny, informed, and conversational, Who Wrote This? urges us as individuals and as communities to make conscious choices about the extent to which we collaborate with AI. The technology is here to stay. Baron shows us how to work with AI and how to spot where it risks diminishing the valuable cognitive and social benefits of being literate.

a watermark for large language models: *Inside KI* Stephan Scheuer, Larissa Holzki, 2024-03-11 2023 fing die künstliche Intelligenz an, zu schreiben und zu sprechen wie ein Mensch. Damit steht die Welt vor einem fundamentalen Umbruch. Unsere Zivilisation, die Art und Weise, wie wir leben und arbeiten, wird so stark verändert werden wie durch die Erfindung des Internets. Noch lässt sich nur erahnen, welches Ausmaß an Innovationen in allen Lebensbereichen nun möglich ist. Fest steht aber, dass wir diese technische und gesellschaftliche Revolution nur dann in unserem Sinne steuern können, wenn wir sie und ihre Pioniere kennen und verstehen. Die Journalisten Larissa Holzki und Stephan Scheuer stellen die führenden Köpfe vor, die diese Entwicklung im Silicon Valley in den USA und in Europa vorantreiben, und zeigen, wie wir jetzt richtig reagieren.

a watermark for large language models: Digital Watermarking and Steganography Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, Ton Kalker, 2007-11-23 Digital audio, video, images, and documents are flying through cyberspace to their respective owners. Unfortunately, along the way, individuals may choose to intervene and take this content for

themselves. Digital watermarking and steganography technology greatly reduces the instances of this by limiting or eliminating the ability of third parties to decipher the content that he has taken. The many techiniques of digital watermarking (embedding a code) and steganography (hiding information) continue to evolve as applications that necessitate them do the same. The authors of this second edition provide an update on the framework for applying these techniques that they provided researchers and professionals in the first well-received edition. Steganography and steganalysis (the art of detecting hidden information) have been added to a robust treatment of digital watermarking, as many in each field research and deal with the other. New material includes watermarking with side information, QIM, and dirty-paper codes. The revision and inclusion of new material by these influential authors has created a must-own book for anyone in this profession. - This new edition now contains essential information on steganalysis and steganography - New concepts and new applications including QIM introduced - Digital watermark embedding is given a complete update with new processes and applications

a watermark for large language models: How AI Ate the World Chris Stokel-Walker, 2024-05-09 'An excellent starter for those who want to gain an insight into how AI works and why it's likely to shape our lives.' – The Daily Telegraph Artificial intelligence will shake up our lives as thoroughly as the arrival of the internet. This popular, up-to-date book charts AI’s rise from its Cold War origins to its explosive growth in the 2020s. Tech journalist Chris Stokel-Walker (TikTok Boom and YouTubers) goes into the laboratories of the Silicon Valley innovators making rapid advances in ‘large language models’ of machine learning. He meets the insiders at Google and OpenAI who built Gemini and ChatGPT and reveals the extraordinary plans they have for them. Along the way, he explores AI’s dark side by talking to workers who have lost their jobs to bots and engages with futurologists worried that a man-made super-intelligence could threaten humankind. He answers critical questions about the AI revolution, such as what humanity might be jeopardising and the professions that will win and lose – and whether the existential threat technologists Elon Musk and Sam Altman are warning about is realistic – or a smokescreen to divert attention away from their growing power. How AI Ate the World is a ‘start here’ guide for anyone who wants to know more about the world we have just entered. Reviews 'An excellent starter for those who want to gain an insight into how AI works and why it's likely to shape our lives.' The Daily Telegraph 'How AI Ate the World prodigiously captures the key issues and concerns around artificial intelligence.' Azeem Azhar, Exponential View 'From ancient China to Victorian England, How AI Ate The World is the story of the characters, moments, technologies, and relationships that populate the rich history of artificial intelligence... How AI Ate The World grapples with what the age of automation means for the people living through it.' Harry Law, University of Cambridge 'A witty, engaging book that takes us through AI's bumpy past to help us understand its present, and future, impacts. I highly recommend it to anyone who is impacted by AI tech – which is to say, everyone on the planet.' Sasha Luccioni, Hugging Face 'Easily the most comprehensive book on AI I have read so far, covering all the key issues' Peter Hunt, Business & Tech Correspondent, Evening Standard 'A comprehensive and compelling look at the technology that's transforming our world. It's an essential guide, full of surprises, to the technology you need to know.' Matt Navarra, social media expert 'Whether you are new to AI or have been following the AI hype for years, Chris Stokel-Walker offers an entertaining balance of history, context and insight that has something for everyone. The story of AI’s evolution is a complex one, but Stokel-Walker tackles it in a clear, direct way that will bring you up to speed while helping you grapple with what it all means — for individuals, the workplace, society and the planet.' Sharon Goldman, VentureBeat 'This book is a wild, brilliant ride through centuries of thinking about and decades of developing machines that can learn. As a crash course in how we got to this current point of thrilling chaos, it will take some beating. Whether or not you agree with Stokel-Walker’s solutions or not, How AI Ate The World is essential reading to understand where we are and how we got here' Ciaran Martin, former CEO, UK National Cyber Security Centre Buy the book to discover your future

a watermark for large language models: Parole de machines Alexei GRINBAUM,

2023-05-03T00:00:00+02:00 Jeunes ou vieux, simples utilisateurs de ChatGPT ou experts en informatique, nous devons faire face à l'avènement d'une ère nouvelle, celle des machines parlantes. Les nouveaux chatbots nous fascinent. Quelles connaissances possèdent-ils ? Quelles technologies opèrent dans leurs profondeurs ? Faut-il faire confiance à un système d'intelligence artificielle ? Autrefois, cette fascination était la marque de dialogues avec les entités non humaines qui peuplaient les mythes. Les agents conversationnels produisent sur nous un effet tout aussi illusoire, et tout aussi réel, que dieux, oracles, anges et démons. Entre les prouesses du numérique et les récits anciens, ce livre décrit notre situation technologique et métaphysique, politique et poétique, dans un monde où nous n'avons plus le monopole de l'expression linguistique.

a watermark for large language models: Lossless Information Hiding in Images Zhe-Ming Lu, Shi-Ze Guo, 2016-11-14 Lossless Information Hiding in Images introduces many state-of-the-art lossless hiding schemes, most of which come from the authors' publications in the past five years. After reading this book, readers will be able to immediately grasp the status, the typical algorithms, and the trend of the field of lossless information hiding. Lossless information hiding is a technique that enables images to be authenticated and then restored to their original forms by removing the watermark and replacing overridden images. This book focuses on the lossless information hiding in our most popular media, images, classifying them in three categories, i.e., spatial domain based, transform domain based, and compressed domain based. Furthermore, the compressed domain based methods are classified into VQ based, BTC based, and JPEG/JPEG2000 based. - Focuses specifically on lossless information hiding for images - Covers the most common visual medium, images, and the most common compression schemes, JPEG and JPEG 2000 - Includes recent state-of-the-art techniques in the field of lossless image watermarking - Presents many lossless hiding schemes, most of which come from the authors' publications in the past five years

a watermark for large language models: Chatbot Fouad Sabry, 2023-07-05 What Is Chatbot A chatbot is a piece of software that, through text or voice interactions, attempts to simulate human conversation. These conversations often take place online. Chatbots are a type of artificial intelligence (AI) system that can hold a conversation with a user in natural language and simulate the way a human would act as a conversational partner. Modern chatbots are capable of holding a conversation with a user in natural language. Deep learning and natural language processing are two areas that are frequently utilized in the development of such systems. How You Will Benefit (I) Insights, and validations about the following topics: Chapter 1: Chatbot Chapter 2: Internet bot Chapter 3: List of chatbots Chapter 4: Virtual assistant Chapter 5: OpenAI Chapter 6: Conversational commerce Chapter 7: LaMDA Chapter 8: ChatGPT Chapter 9: Hallucination (artificial intelligence) Chapter 10: Generative pre-trained transformer (II) Answering the public top questions about chatbot. (III) Real world examples for the usage of chatbot in many fields. (IV) 17 appendices to explain, briefly, 266 emerging technologies in each industry to have 360-degree full understanding of chatbot' technologies. Who This Book Is For Professionals, undergraduate and graduate students, enthusiasts, hobbyists, and those who want to go beyond basic knowledge or information for any kind of chatbot.

a watermark for large language models: *Transforming Conversational AI* Michael McTear,

a watermark for large language models: *Computer Vision – ECCV 2024* Aleš Leonardis,

a watermark for large language models: Handbook of Research on Multimedia Cyber Security Gupta, Brij B., Gupta, Deepak, 2020-04-03 Because it makes the distribution and transmission of digital information much easier and more cost effective, multimedia has emerged as a top resource in the modern era. In spite of the opportunities that multimedia creates for businesses and companies, information sharing remains vulnerable to cyber attacks and hacking due to the open channels in which this data is being transmitted. Protecting the authenticity and confidentiality of information is a top priority for all professional fields that currently use multimedia practices for distributing digital data. The Handbook of Research on Multimedia Cyber Security provides emerging research exploring the theoretical and practical aspects of current security

practices and techniques within multimedia information and assessing modern challenges. Featuring coverage on a broad range of topics such as cryptographic protocols, feature extraction, and chaotic systems, this book is ideally designed for scientists, researchers, developers, security analysts, network administrators, scholars, IT professionals, educators, and students seeking current research on developing strategies in multimedia security.

a watermark for large language models: Asiaccs '18 Asiaccs, 2018-10-29 We are pleased to present herein the proceedings of the 13th ACM Symposium on Information, Computer and Communications Security (ASIACCS 2018) held in Incheon, Korea, June 4-8, 2018. ASIACCS 2018 is organized by AsiaCCS 2018 organizing committee, supported by ACM SigSAC, Korea Institute of Information Security &Cryptography (KIISC). We received 310 submissions. This year's Program Committee comprising 103 security researchers from 24 countries, helped by 73 external reviewers, evaluated these submissions through thoughtful discussion and rigorous review procedure. The review process resulted in 52 full papers being accepted to the program, representing an acceptance rate of about 17%. In addition, 10 short papers and 15 posters are also included in the program. Once again we have a very strong technical program along with 5 specialized pre-conference workshops: CPSS'18, APKC'18, RESEC'18, BBC'18, and SCC'18. We are also fortunate to have distinguished invited speakers, Dr. Cliff Wang, Dr. Jaeyeon Jung, and Prof. Kevin Fu, who will provide various insights into Cyber Deception: an emergent research area, Securing a large scale IoT ecosystem, and Analog Sensor Cybersecurity and Transduction Attacks, respectively. These valuable and insightful talks can and will guide us to a better understanding on both fundamental and emerging topic areas in the field of information, computer and communications security.

a watermark for large language models: Disappearing Cryptography Peter Wayner, 2002 The bestselling first edition of Disappearing Cryptography was known as the best introduction to information hiding. This fully revised and expanded second edition describes a number of different techniques that people can use to hide information, such as encryption.

a watermark for large language models: Fuzzy Logic, Soft Computing and Computational Intelligence , 2005

a watermark for large language models: *Data Hiding* Michael T. Raggo, Chet Hosmer, 2012-12-31 As data hiding detection and forensic techniques have matured, people are creating more advanced stealth methods for spying, corporate espionage, terrorism, and cyber warfare all to avoid detection. Data Hiding provides an exploration into the present day and next generation of tools and techniques used in covert communications, advanced malware methods and data concealment tactics. The hiding techniques outlined include the latest technologies including mobile devices, multimedia, virtualization and others. These concepts provide corporate, goverment and military personnel with the knowledge to investigate and defend against insider threats, spy techniques, espionage, advanced malware and secret communications. By understanding the plethora of threats, you will gain an understanding of the methods to defend oneself from these threats through detection, investigation, mitigation and prevention. - Provides many real-world examples of data concealment on the latest technologies including iOS, Android, VMware, MacOS X, Linux and Windows 7 - Dives deep into the less known approaches to data hiding, covert communications, and advanced malware - Includes never before published information about next generation methods of data hiding - Outlines a well-defined methodology for countering threats - Looks ahead at future predictions for data hiding

a watermark for large language models: International Conference on Intelligent and Smart Computing in Data Analytics Siddhartha Bhattacharyya, Janmenjoy Nayak, Kolla Bhanu Prakash, Bighnaraj Naik, Ajith Abraham, 2021-03-12 This book is a collection of best selected research papers presented at International Conference on Intelligent and Smart Computing in Data Analytics (ISCDA 2020), held at K L University, Guntur, Andhra Pradesh, India. The primary focus is to address issues and developments in advanced computing, intelligent models and applications, smart technologies and applications. It includes topics such as artificial intelligence and machine learning, pattern recognition and analysis, computational intelligence, signal and image processing, bioinformatics,

ubiquitous computing, genetic fuzzy systems, hybrid evolutionary algorithms, nature-inspired smart hybrid systems, Internet of things, industrial IoT, health informatics, human–computer interaction and social network analysis. The book presents innovative work by leading academics, researchers and experts from industry.

a watermark for large language models: Reversible Digital Watermarking Ruchira Naskar, Rajat Subhra Chakraborty, 2022-06-01 Digital Watermarking is the art and science of embedding information in existing digital content for Digital Rights Management (DRM) and authentication. Reversible watermarking is a class of (fragile) digital watermarking that not only authenticates multimedia data content, but also helps to maintain perfect integrity of the original multimedia cover data. In non-reversible watermarking schemes, after embedding and extraction of the watermark, the cover data undergoes some distortions, although perceptually negligible in most cases. In contrast, in reversible watermarking, zero-distortion of the cover data is achieved, that is the cover data is guaranteed to be restored bit-by-bit. Such a feature is desirable when highly sensitive data is watermarked, e.g., in military, medical, and legal imaging applications. This work deals with development, analysis, and evaluation of state-of-the-art reversible watermarking techniques for digital images. In this work we establish the motivation for research on reversible watermarking using a couple of case studies with medical and military images. We present a detailed review of the state-of-the-art research in this field. We investigate the various subclasses of reversible watermarking algorithms, their operating principles, and computational complexities. Along with this, to give the readers an idea about the detailed working of a reversible watermarking scheme, we present a prediction-based reversible watermarking technique, recently published by us. We discuss the major issues and challenges behind implementation of reversible watermarking techniques, and recently proposed solutions for them. Finally, we provide an overview of some open problems and scope of work for future researchers in this area.

a watermark for large language models: Proceedings 2004 VLDB Conference VLDB, 2004-09-17 Proceedings of the 30th Annual International Conference on Very Large Data Bases held in Toronto, Canada on August 31 - September 3 2004. Organized by the VLDB Endowment, VLDB is the premier international conference on database technology.

a watermark for large language models: Dictionary of Information Science and Technology Khosrow-Pour, Mehdi, 2006-11-30 This book is the premier comprehensive reference source for the latest terms, acronyms and definitions related to all aspects of information science and technology. It provides the most current information to researchers on every level--Provided by publisher.

a watermark for large language models: Information Security Sokratis K. Katsikas, Michael Backes, Stefanos Gritzalis, Bart Preneel, 2006-10-04 This book constitutes the refereed proceedings of the 9th International Conference on Information Security, ISC 2006, held on Samos Island, Greece in August/September 2006. The 38 revised full papers presented were carefully reviewed and selected from 188 submissions. The papers are organized in topical sections.

a watermark for large language models: *Neural Machine Translation* Philipp Koehn, 2020-06-18 Learn how to build machine translation systems with deep learning from the ground up, from basic concepts to cutting-edge research.

a watermark for large language models: e-Business and Telecommunications Mohammad S. Obaidat, Joaquim Filipe, 2011-03-16 This book constitutes the refereed proceedings of the 6th International Joint Conference on e-Business and Telecommunications, ICETE 2009, held in Milan, Italy, in July 2009. The 34 revised full papers presented together with 4 invited papers in this volume were carefully reviewed and selected from 300 submissions. They have passed two rounds of selection and improvement. The papers are organized in topical sections on e-business; security and cryptography; signal processing and multimedia applications; wireless information networks and systems.

Additional Resources

A Watermark for Large Language Models - GitHub Pages
Discussion 3: Detecting portions with watermark ØDetecting which portions of a long text are watermarked leads to the “change point detection” problem in statistical analysis.

UNBIASED ∗ WATERMARK FOR LARGE LANGUAGE MODELS
Oct 18, 2023 · The recent advancements in large language models (LLMs) have sparked a growing apprehension regarding the potential misuse. One approach to mitigating this risk is to …

Scalable watermarking for identifying large language model …
In this work, we propose a generative watermarking scheme, SynthID-Text, which builds on previous generative watermarking components, but uses a novel sampling algorithm, …

Watermarking for Large Language Models - Li Lab at CMU
Gumbel Watermark (Aaronson, 2023) that uses a “traceable” pseudo-random softmax sampling when generating the next word. Green-Red Watermark (Kirchenbauer et al., 2023) that …

A Watermark for Large Language Models - GitHub Pages
Watermarked text can be generated using a standard language model without re-training. The watermark can be detected without any knowledge of the model parameters or access to the …

AN UNFORGEABLE PUBLICLY VERIFIABLE WATER MARK FOR …
Recently, text watermarking algorithms for large language models (LLMs) have been proposed to mitigate the potential harms of text generated by LLMs, including fake news and copyright issues.

Turning Your Strength into Watermark: Watermarking Large …
Large language model watermarking methods [33], [34] can be mainly divided into two types: embedding watermarks in the generated text and embedding watermarks in the LLMs.

A Watermark for Large Language Models - GitHub Pages
What is watermarking? • A watermark is a hidden pattern in text that is imperceptible to humans but can be used to identify synthetic text as being generated by a language model.

Watermarking for Large Language Models - Li Lab at CMU
He et al. Protecting Intellectual Property of Language Generation APIs with Lexical Watermark, AAAI 2022. He et al. CATER: Intellectual Property Protection on Text Generation APIs via …

1 Turning Your Strength into Watermark: Watermarking Large …
For watermark extraction, we query the suspicious LLM with questions related to the watermarked knowledge and extract the watermark from its corresponding response.

U WATERMARK FOR LARGE LANGUAGE MODELS - OpenReview
Therefore, in this simplified scenario, the undetectability of the watermark is achieved. However, there is a considerable gap between this toy example and the practical implementation of …

Improved Unbiased Watermark for Large Language Models
Unbiased watermarks offer a powerful solution by embedding statisti-cal signals into language model-generated text without distorting the quality. In this paper, we introduce MCMARK, a …

Watermarking Large Language Models and the Generated …

Watermarking serves as a promising approch to establish ownership, prevent unauthorized use, and trace the origins of LLM-generated content. This paper summarizes and shares the …

Watermarking for Large Language Models - Li Lab at CMU
Watermarking is a promising solution! I want to steal the model! Can we watermark the model? Is this text generated by my model? Model Watermarking Is this model from my model? (e.g., …

A Survey of Text Watermarking in the Era of Large Language …
Text watermarking involves embedding unique, imperceptible identifiers (watermarks) into textual content. These watermarks are designed to be robust yet inconspicuous, ensuring that the …

On the Reliability of Watermarks for Large Language Models
We study the robustness of watermarked text after it is re-written by hu-mans, paraphrased by a non-watermarked LLM, or mixed into a longer hand-written document. We find that watermarks …

Can Watermarks Survive Translation? On the Cross-lingual …
Large language models (LLMs) have exhibited impressive content generation capabilities. Mitigating the misuse of LLM is important. Tagging and identifying LLM-generated content …

Watermarking for Large Language Models - Li Lab at CMU
•Mark My Words: Analyzing and Evaluating Language Model Watermarks •WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models(ACL2024)

StealthInk: A Multi-bit and Stealthy Watermark for Large …
We propose a novel reweighting strategy for multi-bit water-marking that satisfies all of the properties listed in Table 1, which compares our proposed StealthInk scheme with SOTA multi …

A Watermark for Large Language Models - Proceedings of …
We propose a watermark-ing framework for proprietary language models. The watermark can be embedded with negligible impact on text quality, and can be detected using an eficient open …

A Watermark for Large Language Models - GitHub Pages
Discussion 3: Detecting portions with watermark ØDetecting which portions of a long text are watermarked leads to the “change point detection” problem in statistical analysis.

UNBIASED ∗ WATERMARK FOR LARGE LANGUAGE MODELS
Oct 18, 2023 · The recent advancements in large language models (LLMs) have sparked a growing apprehension regarding the potential misuse. One approach to mitigating this risk is to …

Scalable watermarking for identifying large language model …
In this work, we propose a generative watermarking scheme, SynthID-Text, which builds on previous generative watermarking components, but uses a novel sampling algorithm, …

Watermarking for Large Language Models - Li Lab at CMU
Gumbel Watermark (Aaronson, 2023) that uses a “traceable” pseudo-random softmax sampling when generating the next word. Green-Red Watermark (Kirchenbauer et al., 2023) that …

A Watermark for Large Language Models - GitHub Pages
Watermarked text can be generated using a standard language model without re-training. The watermark can be detected without any knowledge of the model parameters or access to the …

AN UNFORGEABLE PUBLICLY VERIFIABLE WATER MARK FOR …
Recently, text watermarking algorithms for large language models (LLMs) have been proposed to mitigate the potential harms of text generated by LLMs, including fake news and copyright issues.

Turning Your Strength into Watermark: Watermarking Large …
Large language model watermarking methods [33], [34] can be mainly divided into two types: embedding watermarks in the generated text and embedding watermarks in the LLMs.

A Watermark for Large Language Models - GitHub Pages
What is watermarking? • A watermark is a hidden pattern in text that is imperceptible to humans but can be used to identify synthetic text as being generated by a language model.

Watermarking for Large Language Models - Li Lab at CMU
He et al. Protecting Intellectual Property of Language Generation APIs with Lexical Watermark, AAAI 2022. He et al. CATER: Intellectual Property Protection on Text Generation APIs via …

1 Turning Your Strength into Watermark: Watermarking …
For watermark extraction, we query the suspicious LLM with questions related to the watermarked knowledge and extract the watermark from its corresponding response.

U WATERMARK FOR LARGE LANGUAGE MODELS
Therefore, in this simplified scenario, the undetectability of the watermark is achieved. However, there is a considerable gap between this toy example and the practical implementation of …

Improved Unbiased Watermark for Large Language Models …
Unbiased watermarks offer a powerful solution by embedding statisti-cal signals into language model-generated text without distorting the quality. In this paper, we introduce MCMARK, a …

Watermarking Large Language Models and the Generated …
Watermarking serves as a promising approch to establish ownership, prevent unauthorized use, and trace the origins of LLM-generated content. This paper summarizes and shares the …

Watermarking for Large Language Models - Li Lab at CMU
Watermarking is a promising solution! I want to steal the model! Can we watermark the model? Is this text generated by my model? Model Watermarking Is this model from my model? (e.g., …

A Survey of Text Watermarking in the Era of Large Language …
Text watermarking involves embedding unique, imperceptible identifiers (watermarks) into textual content. These watermarks are designed to be robust yet inconspicuous, ensuring that the …

On the Reliability of Watermarks for Large Language Models …
We study the robustness of watermarked text after it is re-written by hu-mans, paraphrased by a non-watermarked LLM, or mixed into a longer hand-written document. We find that …

Can Watermarks Survive Translation? On the Cross-lingual …
Large language models (LLMs) have exhibited impressive content generation capabilities. Mitigating the misuse of LLM is important. Tagging and identifying LLM-generated content …

Watermarking for Large Language Models - Li Lab at CMU
•Mark My Words: Analyzing and Evaluating Language Model Watermarks •WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models(ACL2024)

StealthInk: A Multi-bit and Stealthy Watermark for Large ...
We propose a novel reweighting strategy for multi-bit water-marking that satisfies all of the properties listed in Table 1, which compares our proposed StealthInk scheme with SOTA multi …

A Watermark For Large Language Models

A Watermark For Large Language Models

Related Articles

- a problem well defined is half solved
- a technological advance in the production of automobiles will
- a short term financial goal

Back to Home