Landscape Study of **Generative Artificial Intelligence in the Criminal Justice System**

June
2025

NIJ Contact:
Steven Schuetz
Senior Physical Scientist
steven.schuetz@ojp.usdoj.gov

CJTTEC Contacts:
Jeri D. Ropero-Miller, PhD, F-ABFT
Senior Fellow, Principal Scientist,
Project Director, CJTTEC
jerimiller@rti.org

Jacob Smith, PhD
Innovation Advisor
jacobsmith@rti.org

NIJ | NATIONAL INSTITUTE OF JUSTICE
ADVANCING JUSTICE THROUGH SCIENCE

Criminal Justice Technology Testing and Evaluation Center
A Program of the National Institute of Justice

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY
2 WHAT IS GENERATIVE AI?
- Inputs and Outputs
- Utility of Generative AI
- Risks of Generative AI

5 USE CASES OF GENERATIVE AI IN THE CRIMINAL JUSTICE SYSTEM: APPLICATIONS, BENEFITS, AND LIMITATIONS
- Example Applications of Generative AI in Criminal Justice Settings

10 CRIMINAL JUSTICE GENERATIVE AI IMPLEMENTATION CONSIDERATIONS
- Technical Considerations
- Operational Considerations
- Governance Considerations

17 FUTURE OUTLOOK
18 APPENDIX I: GLOSSARY
21 APPENDIX II: GENERATIVE AI BASICS
- Technical Basics: Architecture and Access
- Technical Basics: Foundation Models and Training
- Advanced AI Concept Considerations for Criminal Justice Practitioners

24 REFERENCES

Cover image created via Midjourney on May 9th, 2025 using the following prompt:

"Professional, clean, minimalistic image featuring a realistic background such as a justice-related environment (e.g., courtroom, government building, or institutional architecture) with three modern, clean white icons in the foreground in clear focus: a gavel for courts, a police badge for law enforcement, and a simplified representation of a correctional facility. Position the icons in a balanced horizontal layout. Visually link them using vibrant digital interconnections glowing neural networks, flowing data lines, and abstract AI patterns to represent the role of generative AI in connecting the justice system. The background should be subtle and professional, while the icons remain clear and symbolic. Use a dominant color palette in blues: #12284C, #2B5597, #3CB4E5, with accents in #A7A8A9, #54565A, and black (#000000). Composition should be sleek, modern, and tech-forward, ideal for a report cover."

This report was authored primarily by RTI International's Innovation Advisors—Jacob Smith, with support from Rebecca Shute, Meghan Camello, and Kristina Cooley—and RTI's Justice Practice Area—Michael Planty and Jeri Ropero-Miller. CJTTEC would like to thank Chief Phil Lukens (ret.), NIJ LEADS Scholar, and Dr. Ian T. Adams, Department of Criminology & Criminal Justice at the University of South Carolina for serving as external reviewers for this document.

This project was supported by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, through award number 15PNIJ-23-GK-00931-NIJB. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice. The products detailed in this landscape study are intended to be a good-faith overview, but not an exhaustive list of commercially available products and products approaching market readiness. The inclusion of a product or company in this report does not represent National Institute of Justice's or Criminal Justice Technology Testing and Evaluation Center's recommendation, endorsement, or validation of product claims.

Suggested Citation:

Smith, J., Camello, M., & Planty, M. (2025). *Landscape study of generative artificial intelligence in the criminal justice system*. Research Triangle Park, NC: RTI International. https://cjttec.org/

Criminal Justice Technology Testing and Evaluation Center (CJTTEC)

CJTTEC is a program of the National Institute of Justice (NIJ) that uses research-based methodologies to enhance the capabilities of law enforcement, courts, and corrections agencies. CJTTEC leverages expertise from varied criminal justice community stakeholders to understand and test technologies and practices in a variety of NIJ's research areas.

RTI International

RTI International is an independent, scientific research institute dedicated to improving the human condition. Our vision is to address the world's most critical problems with technical and science-based solutions in pursuit of a better future. Clients rely on us to answer questions that demand an objective and multidisciplinary approach—one that integrates expertise across social, statistical, data, and laboratory sciences, engineering, and other technical disciplines to solve the world's most challenging problems. For more information, visit www.rti.org.

RTI leads CJTTEC. CJTTEC leverages RTI's expertise in criminal justice, investigative and forensic sciences, innovation, technology application, economics, data analytics, statistics, program evaluation, public health, and information science.

EXECUTIVE SUMMARY

Generative artificial intelligence (AI) refers to AI[1] used to create content, such as text, images, music, audio, and videos.[2] Generative AI offers many potential benefits, enabling users to automate, augment, and accelerate a wide range of workflows, from simple administrative tasks like transcription and translation to more-complex functions such as investigation and decision support. In the criminal justice system, generative AI offers promising solutions to address human resource and budget challenges, allowing practitioners to focus on more-impactful work. Generative AI–integrated tools may enhance data analysis, improve detection and objective assessment of evidence, and streamline administrative processes. However, its integration, particularly in the criminal justice domain, raises some concerns, including potential biases, privacy issues, and the need for rigorous oversight to ensure effective implementation. It is unclear whether these tools can deliver on their promised efficiencies in practice, as evidenced by early research evaluating time savings of implementing AI-assisted report writing software.[3] These concerns highlight the necessity for addressing bias and accuracy, maintaining strict data privacy and security protocols, and promoting transparency and accountability in AI-driven decisions and processes.

This report is intended to help criminal justice decision-makers do the following:

- Understand what generative AI is and how it relates to the criminal justice system
- Identify how generative AI may be applied to tasks and jobs within the criminal justice system and the potential benefits, realities, and limitations
- Consider the technical, operational, and governance factors that may influence adoption and implementation
- Understand what makes up the generative AI technology stack and how models can be trained

Key Takeaways

- Generative AI represents an acceleration and advancement in technological innovation that already impacts the criminal justice system and will continue to do so—it is no longer a question of if or when, but how and to what extent.
- Generative AI–powered software tools may offer many potential benefits, such as improving efficiency and augmenting capabilities across an extremely broad set of applications for criminal justice system stakeholders. Although these products hold promise, little empirical evidence currently supports or refutes promised benefits from these products.
- Generative AI models can be deployed in various forms, including cloud-based models that centralize data processing and federated models that enable decentralized training across multiple locations, preserving data privacy and enhancing security for sensitive criminal justice applications.
- Decision-makers should be aware of the substantial technical, operational, and governance risks associated with generative AI–powered software tools prior to implementation.
- Responsible use of generative AI requires addressing bias and accuracy concerns, maintaining strict data privacy and security protocols, adhering to ethics and legal standards, and promoting transparency and accountability in AI-driven decisions and processes.
- Generative AI technology is evolving faster than the legal or policy environment for AI—the criminal justice community must be proactive and must implement robust internal training and policy frameworks rather than relying solely on external legal or regulatory guidance.

WHAT IS GENERATIVE AI?

Inputs and Outputs

Generative AI refers to advanced software programs specifically designed to create new content, such as text, code, images, music, audio, and videos, from given inputs like text, images, or audio. Unlike discriminative AI, which primarily classifies or predicts outcomes, generative AI focuses on producing novel outputs based on the prompts it receives. The models on which many generative AI tools are built operate probabilistically, meaning they generate outputs by predicting likelihoods for all possibilities of next elements (e.g., words, pixels) based on patterns learned from their training data. For example, facial recognition is a well-known application of AI in the criminal justice system. Discriminative AI tools may classify and predict whether detected faces match known individuals. However, generative AI-based facial redaction or anonymization tools may generate anonymized synthetic faces for redaction in video footage, which enables privacy while maintaining realistic visuals.[4] Appendices I and II describe generative AI common terms and provide a basic overview of underlying technology concepts, respectively.

Figure 1 provides a high-level summary of the tech stack that makes up a generative AI tool. Users typically interact with these tools through various interfaces, including web applications like chatbots or native software applications with application program interfaces (APIs) that provide access to AI models. In these interfaces, users provide requests for specific tasks (such as answering questions or generating content) in natural (i.e., plain) language, referred to as a prompt. The application sends the prompt to an AI model, which processes it and generates the output. At the core of many of these generative AI tools are foundation models, a general type of large-scale AI model. Prominent examples of foundation models include large language models (LLMs)—such as OpenAI's GPT, Anthropic's Claude, or Google's Gemini—which were initially developed to specialize in text generation. These models are pretrained on broad diverse data sets using self-supervised learning techniques, giving them diverse capabilities without specific training.[5] Foundation model parameters can be further refined and fine-tuned through human feedback to adapt them for specific tasks, such as those within the criminal justice domain. Typically, freely accessible tools are optimized as generalists and not specific to any domain. These models are typically hosted on cloud platforms, and the physical infrastructure of computer hardware to power model training and processing can be found in large, remote data centers.

> Users of generative AI interfaces can word their prompts in simple human language, significantly reducing technical barriers to use. For example, the user could use the prompt, "Summarize the following body-camera footage transcript for use in a court brief," for an output like the following:
>
> "The footage shows Officer [name] approaching a suspect, [suspect's name], who is visibly agitated. After initial questioning, the officer attempts to de escalate the situation. The suspect resists, and a brief struggle ensues before the suspect is subdued and taken into custody."

Utility of Generative AI

Broadly, generative AI offers benefits across different domains and industries:

- Increased Efficiency: Generative AI can automate complex tasks and produce high-quality outputs quickly, reducing time and resource requirements.[6]

- Improved Productivity: By handing repetitive or labor-intensive tasks, generative AI increases productivity, freeing up resources for more-complex tasks.[7]
- Enhanced Accessibility: AI-generated summaries and translations make information more digestible and widely available.
- Improved Decision-Making: By quickly developing data-driven insights, generative AI tools can provide suggestions that inform user decision-making.

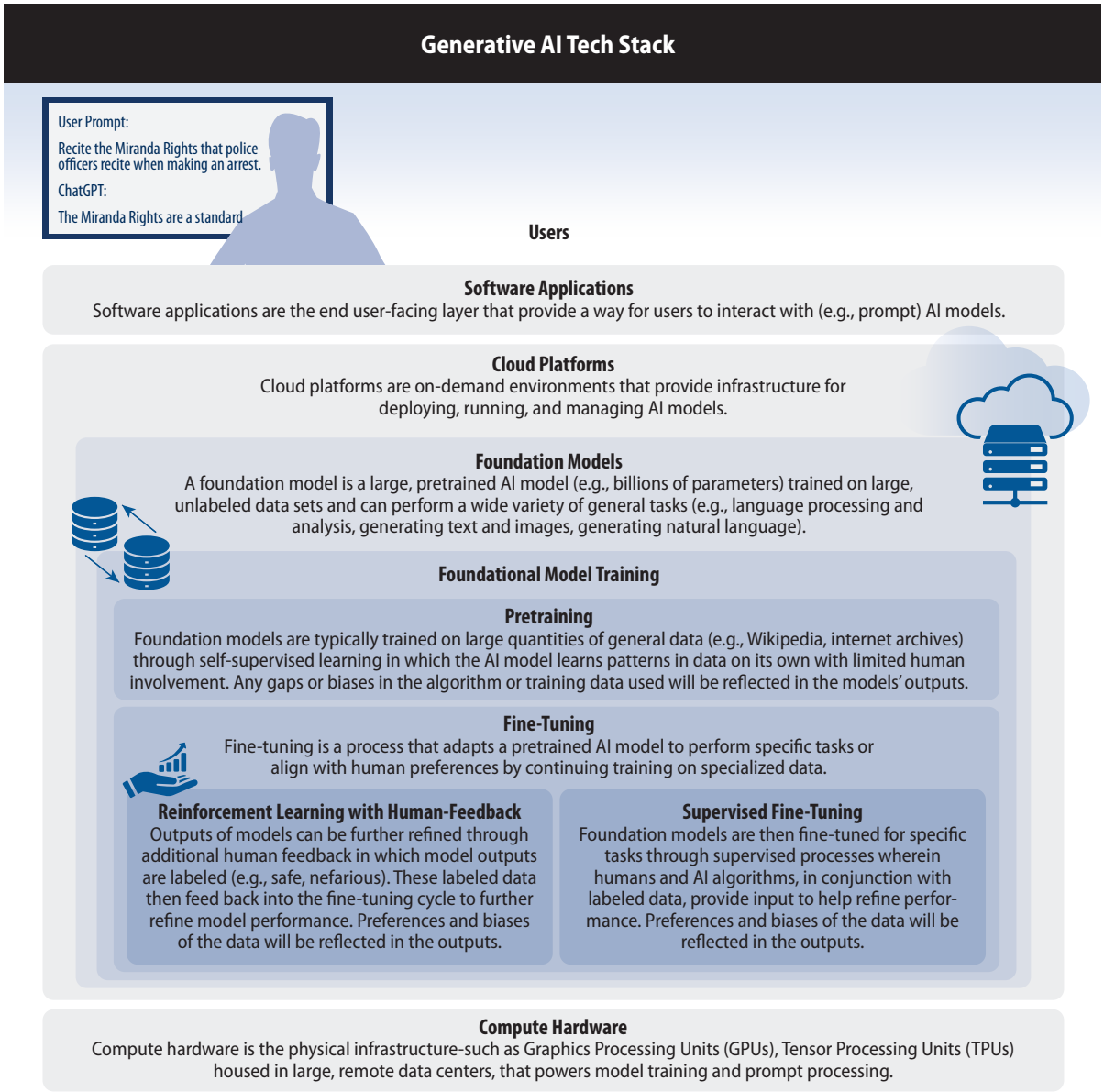

Figure 1: High-level technology stack overview, showing how users interact with the generative AI interface and the technology powering this interface.

Risks of Generative AI

Although generative AI tools offer potential benefits, the associated risks are equally significant for criminal justice practitioners.

Bias, Accuracy, and Hallucinations

A notable concern is inherent bias in generative AI models, which often stems from training on and operational use[5] of biased data sets such as Wikipedia, which has known biases.[8] This concern demands vigilant monitoring and mitigation by all stakeholders (e.g., software developers, practitioners, auditors) involved with the development and implementation of generative AI tools.

Additionally, generative AI models can generate inaccurate or entirely fabricated outputs, known as hallucinations. If undetected, these errors can compromise critical documents, from routine emails to sensitive court briefs or motions. These models may also have blind spots, due to missing data or knowledge cutoffs, where they fail to consider relevant information or nuances. Addressing both hallucinations and blind spots requires rigorous validation of outputs.

Data Security and Privacy

Data security and privacy are additional concerns, particularly as many generative AI tools operate on cloud infrastructure, where sensitive legal information may be processed or stored. Many widely used free AI tools (e.g., ChatGPT, Claude, Gemini) collect user data and may not be suitable for handling confidential or sensitive information. However, some cloud platforms—such as Microsoft Azure's FedRAMP-compliant offerings—support more-secure, private deployments that may be suitable for sensitive use cases.[9] Criminal justice agencies must ensure that AI tools comply with relevant data privacy regulations and implement safeguards to protect confidential information. For example, using a tool without proper data protections could lead to the inadvertent exposure of personally identifiable information or unauthorized access to sealed case files. See the following implementation considerations for more details on risks.

Intentional Exploitation of Models

Another emerging threat is the skeleton key risk, whereby AI model safeguards can be bypassed through malicious or misleading prompts, allowing users to exploit and manipulate a model's intended behavior and outputs.[10] By exploiting vulnerabilities in generative AI, malicious actors can manipulate the model's behavior to produce harmful, unlawful, or unethical outputs. For instance, convincing an AI tool to generate dangerous or illegal content, such as instructions for building a bomb, under the guise of legitimate educational or research purposes demonstrates how easily such models can be weaponized.

USE CASES OF GENERATIVE AI IN THE CRIMINAL JUSTICE SYSTEM: APPLICATIONS, BENEFITS, AND LIMITATIONS

In criminal justice settings, generative AI tools can be used to automate, augment, and accelerate a wide range of tasks, from administrative activities to decision support (as shown in **Table 1**). Although the technology is already being employed to perform different tasks, many more applications will likely emerge over time as the technology matures and improves, new benefits are discovered, and limitations are addressed.

Table 1: Example use cases of generative AI and associated benefits and limitations

Category	Example Tasks	Potential Benefits	Limitations
Administrative & Record Management	• Automating document drafting, case file management, and scheduling • Auto-generating summaries of incident and arrest reports • Drafting routine correspondence and reports • Creating automated task reminders and notifications* • Streamlining record organization and categorization* • Generating compliance reports for legal and regulatory standards*	• Reduce manual workload, improving efficiency • Enhance document accuracy and consistency • Ensure better compliance with legal standards	• Can sometimes generate content that appears accurate but may contain subtle errors or hallucinations • Requires ongoing human oversight to ensure compliance with complex legal standards and correct interpretation of details
Data Analysis & Decision Support	• Generating insights and recommendations from large data sets (e.g., legal research, investigative support, jury selection,* patrol route generation,* inmate classification,* real-time crime center support*) • Enhancing data analysis of body-worn camera audio and legal databases • Facilitating the conversion and integration of various file types • Enhancing digital evidence handling and processing for accuracy and reliability • Detecting potential forgeries by analyzing handwriting, signatures, and documents* • Creating predictive models for criminal justice strategies* • Analyzing facial features to generate high-resolution images of suspects from low-quality evidence* • Analyzing behavioral data to assist in creating accurate behavioral profiles of suspects*	• Accelerate data-driven decision-making • Improve accuracy in evidence assessment • Increase efficiency in handling large data sets • Aid in proactive crime prevention by narrowing suspect lists and predicting potential criminal behavior • Enhance identification accuracy with high-resolution facial recognition • Provide high accuracy in forgery detection, strengthening legal evidence	• Limited by data processing capabilities and context window size • May struggle with open-ended or complex legal queries • Risks of inaccuracies due to model drift • Prone to contrafactual bias, requiring human oversight

Table 1 cont.

Category	Example Tasks	Potential Benefits	Limitations
Content Generation & Modeling	• Drafting summaries of interactions based on transcripts and reports • Augmenting composite sketch creation and suspect identification • Auto-generating court bundles with documents, evidence, and legal arguments* • Creating 3D models and visualizations of crime scenes by analyzing data such as witness statements, photographs, and physical evidence*	• Speed up case preparation and analysis • Increase accuracy and consistency in visual and textual content • Improve the reliability of evidence interpretation	• Technological limits in rendering complex models and sketches; some applications are speculative and unproven • Risk of bias in automated outputs • Requires calibration and oversight to prevent inaccuracies • Initial research[3] and anecdotal reports[11] note that efficiency and other potential benefits of tools (such as AI-assisted report writing) have not yet been realized
Training & Simulation	• Automating content creation for training and anti recidivism programs • Prisoner education • Developing dynamic training scenarios for law enforcement* • Generating realistic virtual environments for training*	• Enhance training realism and adaptability • Tailor learning experiences to individual needs • Facilitate continuous feedback and improvement	• Early-stage technology with technical limitations • Computational demands may exceed institutional capacity • May not fully address unique learning styles
Communication & Coordination	• Crafting real-time updates and briefings for responders* • Generating consistent communication across multiple channels*	• Ensure timely and accurate information dissemination • Improve coordination during large-scale incidents	• Risks of generating inaccurate or misleading information • Requires continuous monitoring by human operators
Public Interaction & Assistance	• Using generative AI chatbots to deliver personalized responses to public inquiries • Generating educational content and rapid response FAQ systems for public use • Routing nonemergency services*	• Improve public access to information • Enhance responsiveness to public inquiries • Reduce demand on human resources	• Prone to errors in interpreting nuanced queries • May produce incorrect or unclear responses • Susceptible to contrafactual bias; requires human oversight to ensure reliability
Knowledge Management for Investigative Purposes	• Enabling multimodal search across disparate law enforcement data sets to identify signals or patterns, analyze large amounts of data, and note connections	• Save time sifting through large amounts of data • Suggest actionable patterns and insights that would be time-intensive or impossible to conduct without visualization and analysis	• Complexity of multimodal data may cause difficulties in analyzing diverse data types (e.g., text, images, video) • Lack of standardization across data sets may hinder effective analysis of multimodal data

* Indicates tasks that are not currently under development but could be potential future tasks.

Example Applications of Generative AI in Criminal Justice Settings

Generative AI–powered tools are being developed, tested, piloted, or employed for tasks that include transcription and translation, legal document summarization, legal document drafting, legal research and citations, legal document review, scheduling, data and evidence query, multimodal body-worn camera data analysis, investigation support, and even personalized diversion support.

Table 2: Example use cases of generative AI in the criminal justice system and associated benefits and limitations

Category	Example use cases of generative AI in the criminal justice system		
Example Use Case	Generative AI-Powered Draft Report Generation	Generative AI-Powered Legal Research and Document Review and Preparation	
Target Users	Law Enforcement Officers	Courts, Legal Professionals	
Availability	Available Now	Available Now	
Prominent Classes of AI Used	Generative AI	Generative AI	
Description	Generative AI–powered tools assist law enforcement officers in drafting detailed and accurate police reports by converting input data, such as voice notes or bullet points, into well-structured narratives. These tools can also provide templates or suggest content based on the type of incident, ensuring consistency and thoroughness in reports.	Generative AI–powered software tools that leverage the enhanced text generation and data query capabilities of generative AI to draft documents, assist with research, and help stakeholders prepare for depositions.	
Pain Points Addressed	Drafting police reports can be time-consuming and inconsistent due to varying writing skills among officers. The risk of omitting crucial details or introducing errors when manually drafting police reports, especially under time constraints or heavy workload, is a significant concern. Downstream users of police reports (e.g., media, prosecutors, victims) benefit from expedited access to this information. Additionally, the administrative burden of writing reports reduces the time available for active policing.	Generative AI addresses many pain points: large caseloads, substantial time required for manual document drafting and case research, the high propensity for errors in citations and factual accuracy, the intensive resource demands for document review and compliance checks, and the complexity of managing and linking extensive legal documents.	
Specific Tasks Augmented	Generative AI augments drafting incident reports, converting structured input into comprehensive narratives, auto-generating sections of reports, ensuring compliance with legal standards, and translating content for multilingual reports.	Generative AI augments reviewing discovery documents, conducting contract policy compliance checks, performing legal research, drafting memos, document review, preparing for depositions, creating hyperlinked legal documents, generating timelines, spotting fake cases, cite-checking, generating exhibits for filing, and analyzing opposing counsel's briefs for factual weaknesses.	
Potential Benefits	Generative AI has the potential to enhance efficiency by reducing report drafting time, improve accuracy and completeness of reports, ensure consistency in report formatting and content, and allow officers more time for active duty.	Generative AI tools could enable faster and more efficient legal analysis, streamline legal document preparation, and ensure accuracy in citations and evidence presentation. These tools may be implemented to automatically redact race-related information from case narratives to enable blinded prosecutorial decision-making, reducing instances of bias.[13]	
Key Considerations	Although generative AI tools can streamline report drafting, they must ensure data privacy and security, especially when handling sensitive or confidential information. Officers should verify the accuracy of AI-generated reports, and the tools must be integrated with existing law enforcement systems to maximize effectiveness. Monitoring for potential biases and hallucinations in AI-generated content is also crucial. Little empirical evidence demonstrates the impact of these tools; for example, results of early research assessing the time savings impact of implementing AI-based report writing assistants to law enforcement workflows indicated that AI assistance did not significantly impact the duration of writing police reports.[3] Likewise, the deputy chief of the Anchorage Police Department noted in an Alaska Public Media interview that piloting AI-assisted report writing did not contribute to significant time savings for the officers.[12] More research and evaluation is needed to understand how and where these tools can provide value.	Cloud-based legal software tools, including generative AI tools, may pose greater privacy and security risks than traditional on-premises legal software. Cloud infrastructure often involves remote data processing and storage, increasing the chances of unauthorized access or data breaches. Ensuring that these tools offer robust encryption, strict access controls, and comply with data privacy regulations like CJIS is essential. Additionally, careful integration with existing legal platforms is necessary to prevent vulnerabilities.	
Example Tools & Vendors (tool | vendor)	Axon | Draft One	Thomas Reuters | CoCounsel	DecoverAI | DecoverAI
	365 Labs | CoPilot.AI	Clearbrief | Clearbrief	DeepJudge | DeepJudge
	TRULEO | Field Notes	LexisNexis | Lexis+ AI®	Harvey | Harvey
		Wordsmith AI | Wordsmith	

Table 2 cont.

Category	Example use cases of generative AI in the criminal justice system	
Example Use Case	Body-Worn Camera Transcription and Analysis	Knowledge Management for Investigative Purposes
Target Users	Law Enforcement Agencies that Employ Body-Worn Cameras	Law Enforcement
Availability	Available Now	Available Now
Prominent Classes of AI Used	Generative AI and Discriminative AI	Generative AI and Discriminative AI
Description	Software tools that integrate generative AI and advanced traditional AI technologies and techniques, such as computer vision, natural language processing, and speech processing, can automate transcription and analysis of body-worn camera data.	Platforms use AI to assist in multimodal analysis across law enforcement data (e.g., CAD/RMS, police reports, digital evidence, agency documents, open-source intelligence sources), providing actionable insights and intelligence. These platforms leverage generative AI to provide semantic search across data, summaries and reports, and data visualization and analysis.
Pain Points Addressed	Vast quantities of footage are captured but rarely used by agencies due to the massive effort required to review, transcribe, and analyze the content manually and the need for specialized personnel and technology to handle the data efficiently, making it cumbersome and resource-intensive for law enforcement to fully use the footage for investigative and training purposes.	The platforms address the following pain points: disparate data sources requiring manual review, time-consuming searching across databases and systems, and slow search, retrieval, and analysis processes.
Specific Tasks Augmented	AI can augment audio transcription, tone and speech analysis, recognition and classification of events (e.g., use-of-force, pursuits, frisking, noncompliance incidents), recognition of professional and unprofessional language use, and analysis of community–police interactions, and redaction.	AI augments document and report generation and summarization, semantic searching, pattern identification, trend prediction, and data visualization.
Potential Benefits	AI offers efficient and cost-effective solutions that enhance public trust and safety. By identifying patterns in police–community interactions, these systems can help optimize community engagement, patrol operations, supervision, and training. Additionally, they facilitate automated supervision and coaching, thereby enhancing police professionalism.[14] These tools can help recognize behavioral patterns, which can lead to improved compliance and reduced escalations in police–community interactions, further contributing to a more effective and community-focused policing approach.	AI can help streamline workflows, reduce manual workload, enhance efficiency and speed of data retrieval, and support investigations.
Key Considerations	AI tools that transcribe and analyze body-worn camera footage raise concerns about data security, audio misinterpretation, and potential bias. Background noise, dialects, or overlapping speech can lead to inaccurate transcripts, and foundation models may reflect biases in their training data. Mitigation strategies include deploying Criminal Justice Information Services (CJIS)–compliant tools, encrypting audio and transcript data, and incorporating human-in-the-loop review processes. Vendors or agencies can also fine-tune models on law enforcement–specific audio and implement audit trails and logging to enhance accountability and accuracy.	Key risks include hallucinated outputs, data leakage, and unauthorized access to sensitive materials. To mitigate these risks, agencies should adopt CJIS-compliant tools with strong access controls, encryption protocols, and auditing capabilities. Ensuring transparency and human review of AI-generated insights remains essential to ensure accuracy and reliability in legal procedures.
Example Tools & Vendors (tool | vendor)	Polis Solutions | TRUSTSTAT	C3.AI | C3.AI
	TRULEO | Body Camera Analytics	Penlink | CoAnalyst
		ForceMetrics | ForceMetrics

Table 2 cont.

Category	Example use cases of generative AI in the criminal justice system
Example Use Case	Transcription, Translation, and Facilitation of Next Steps in Courts
Target Users	Courts and Correctional Systems
Availability	Available but Custom Development Required
Prominent Classes of AI Used	Generative AI and Discriminative AI
Description	Generative AI–powered software tools provide assistance to the full spectrum of court system stakeholders and augment capabilities that range from paperwork completion to judgment preparation for live courts, hybrid courts, and fully virtual courts.
Pain Points Addressed	AI addresses many concerns, including error-prone manual form completion; inconsistencies in case detail writing; delays from coordination difficulties in hearing preparation and scheduling; time-consuming case and judgment preparations; transcription inaccuracies and slow transcript analysis, leading to misinterpretations; inefficient minute summarization; and time-consuming data retrieval from outdated legacy systems.
Specific Tasks Augmented	AI augments form assistance, case detail writing, hearing preparation, automated scheduling, case preparation, judgment preparation, transcription, transcript analysis, transcript summarization, minute preparation, and data searching across legacy systems.
Potential Benefits	AI can enhance customer-service centric paradigms in courts, enable privileged attorney-client conversations in digital settings, support multilingual communication, and provide real-time insights for court management.
Key Considerations	AI tools used in court-related processes face challenges in accurately interpreting legal terminology, regional language nuances, and informal speech while maintaining privacy. Infrastructure limitations can further impact reliability. Mitigation strategies include fine-tuning models on legal and courtroom-specific language, securing data with encryption, and using human reviewers to verify translations or AI-suggested next steps before they are used in official proceedings.
Example Tools & Vendors (tool | vendor)	Microsoft + CX Unicorn | Custom Instances
	Steno | Transcript Genius
	Filevine | Depo Copilot

CRIMINAL JUSTICE GENERATIVE AI IMPLEMENTATION CONSIDERATIONS

The adoption of generative AI tools in the criminal justice system requires careful evaluation of technical feasibility, operational viability, and governance considerations. As these tools advance, their potential to improve efficiency, productivity, accessibility, and decision-making must be weighed against challenges such as system integration, data privacy, and accuracy in high-stakes legal contexts. Additionally, the governance risks related to bias, transparency, and oversight demand particular scrutiny. By addressing these factors, stakeholders can implement generative AI responsibly, ensuring that these tools contribute to, rather than undermine, justice.[15]

Technical Considerations

Criminal justice practitioners should evaluate the technical considerations before implementing a generative AI tool. Technical considerations focus on six key aspects of quality: poor specificity, hallucination, real-time data or training cutoff dates, context window limitations, model drift, and homogenization.

Quality Monitoring: Key challenges for generative AI tools in criminal justice involve ensuring the relevance and accuracy of outputs, including addressing several specific issues:

- **Specificity:** The foundation models that underpin generic generative AI tools are trained on broad publicly available data. Although this gives these models a broad knowledge base, it also means they may lack specific context or knowledge relevant to the criminal justice system. To improve output relevance, several approaches can be considered. Prompt engineering is a method involving the careful design of user inputs to guide the model's responses. For example, generic generative AI tools may not properly consider specific nuances of legal jargon, mannerisms of law enforcement communication, or localized trends when prompted. Prompt engineering can be used to tailor the input prompts to better fit the specific context in many cases, such as having the model adopt a role, giving it definitions of domain-specific jargon, or evaluating model performance and improving iteratively. Retrieval-augmented generation (RAG) is another method that allows a model to retrieve relevant external documents—such as local case law or policy guidance—at runtime, improving specificity without requiring changes to the model itself. Fine-tuning involves augmenting the model leveraging domain-specific data (e.g., regional laws, communication styles, case files) to improve contextual accuracy, although this approach requires technical expertise and access to sensitive data sets. Human feedback can be incorporated during model development or fine-tuning to further refine outputs for accuracy and appropriateness, particularly in high-stakes legal contexts.
- **Hallucination:** The outputs from generative AI–powered tools are prone to hallucinations, wherein the model generates content that is factually incorrect, yet appears plausible. These hallucinations occur when the model makes connections in its latent space—abstract representations within the AI—that do not exist in reality. For example, the model might incorrectly generate a statement that combines unrelated legal concepts, such as referencing a nonexistent court ruling or merging details from different cases into one, creating a fictitious precedent. These errors are not a result of probabilistic sampling; rather, they stem from the way

the model processes and generates novel content based on the vast and varied data it has been trained on. Generative AI–powered tools have been documented to create fake legal cases,[16] fabricate citations,[17] and make up sentences during transcription-related tasks.[18]

- **Real-Time Data or Training Cutoff Dates:** Foundation models that underlie generative AI tools are trained on historical data often only up to a specific cutoff date. Therefore, the models may not reflect the most current information, such as legal precedents or case law. Criminal justice practitioners should be aware of cutoff dates for generative AI tools used and should consider how they themselves or tool vendors plan to ensure models are continually updated over time. Some of the most recent models from several companies are highlighted in **Table 3**.

Model Provider	Model Name	Cutoff Date
OpenAI	GPT-4o	June 2024[19]
Google	Gemini 2.0 Flash	August 2024[20]
Anthropic	Claude 3.7 Sonnet	November 2024[21]

Table 3: Cutoff dates for select generative AI models

  If an investigator were using a generative AI tool to compile a timeline of recent events in a specific area, the tool alone without access to other data sources would lack knowledge of any events occurring after its underlying model's training cutoff date. This limitation is compounded by the fact that the training data may not include comprehensive coverage of local events, particularly in areas with limited news reporting. Therefore, the investigator would need to manually provide any relevant recent information or retrieve it from current sources to ensure the accuracy and completeness of the timeline generated by the AI tool.

- **Context Window Limitations:** Foundation models that underlie generative AI tools can only consider a set amount of text at once, usually due to technical and architectural constraints. This limitation impacts the model's ability to interpret and maintain context over long stretches of text. For example, when processing lengthy legal documents or complex narratives, the model might lose track of earlier information, leading to less-coherent or -accurate outputs as it processes further into the text. Inputs to foundation models within generative AI tools are tokenized, or split into smaller parts, to be processed by the models. Several of the most-popular foundation models have context windows that range from 128,000 to 200,000 tokens or approximately 85,000 to 133,000 words. Exceptions to this general window are already emerging as technology advances, with models like Google's Gemini 2.0 Flash having an input context window of 1,048,576 tokens or roughly 700,000 words.[22] Although these windows are expanding, window size does not always directly correlate with accuracy in outputs for longer prompts—model architecture can lead to the loss of nuance in context well below these token limits.
- **Model Drift:** The outputs produced by generative AI tools are subject to model drift, meaning the quality of AI-generated completions can degrade over time due to changes in the underlying data or shifts in societal behaviors, legal standards, or environmental factors that were not part of the model's original training. As a result, without regular updates or retraining with new data, the model's outputs may become less accurate and relevant, leading to a misalignment between the tool's responses and the current needs or conditions it is meant to address. In the context of criminal justice, using modern generative AI to review old cases could inadvertently uncover major errors that might lead to legal challenges, lawsuits, or even acquittals if these outdated models were applied without proper oversight. Furthermore, historical biases, such as those related to racial disparities in policing, embedded in the training data could perpetuate these issues in future outputs, exacerbating existing problems rather than addressing them.

- **Homogenization:** Homogenization in foundation models, commonly used in generative AI, refers to the reuse of similar architectures—the underlying structures and design frameworks that determine how AI models process information—across multiple applications. Although this approach streamlines development and improves efficiency, it can degrade the quality of outputs. When a narrow range of models and training data is applied broadly, outputs may lack the specificity, contextual sensitivity, or robustness required in sensitive areas. This risk is particularly concerning for criminal justice, where outputs that fail to account for context or nuance can reinforce biases, create errors, and compromise fairness. Additionally, the concentration of resources needed to create these models means that only a few entities can develop them, limiting the diversity in training data and perspectives.

Operational Considerations

The operational integration of generative AI tools within the criminal justice system encompasses several key considerations: **policies and procedures, training, funding,** and **data management.**

In light of the risks of using tools with generative AI capabilities, the criminal justice community may prohibit use of these tools. For example, in Washington, the King County Prosecuting Attorney's Office will not accept law enforcement narratives that have been written with the help of AI-assistant report writing tools.

1. **Policies and Procedures**
   - **Privacy and Data Security:** Privacy and data security are foundational considerations when integrating generative AI into internal processes, particularly within the criminal justice system where sensitive data and high-stakes decisions may be involved. Many generative AI–powered tools operate via cloud-based applications that transmit, store, and sometimes retain data for future model training. This approach raises significant concerns about data ownership, unauthorized access, and data breaches. To mitigate these risks, robust cybersecurity measures and strict data handling policies must be in place. In parallel, privacy risks may also arise during model training when sensitive data are used. Advanced AI techniques, such as federated learning, provide an additional layer of security by enabling AI training across decentralized locations. This approach enhances privacy by keeping sensitive data local, making it particularly valuable in criminal justice applications. Furthermore, synthetic data generation can create artificial data sets for training models without compromising real-world sensitive information, thereby bolstering privacy. Although federated learning and synthetic data generation offer important security benefits, they remain supplementary rather than essential components of AI implementation. Stakeholders must also ensure that appropriate authorities retain ownership of sensitive data. Tools like Google Gemini Apps[23] and Open AI's ChatGPT app,[24] for instance, may store user data—including conversations and uploaded files—which can be reviewed or used for model improvement, as highlighted in their privacy policies. Organizations must assess how new tools fit within existing operational structures—who manages the tools, who approves data access, how audits are conducted, and whether vendors' practices (e.g., storing user interactions for model improvement) align with internal policies and procedures.
   - **Oversight and Accountability:** Establishing policies and procedures that ensure accountability for generative AI outputs is critical, especially with the relevance and accuracy issues discussed in the technical considerations, and the biases discussed in the governance considerations

below. Challenges such as bias, AI hallucinations, inaccuracies, model drift, and the difficulty of maintaining context in extensive texts present significant risks. Having mechanisms in place and humans in the loop is essential to monitor and manage these outputs, ensuring they remain reliable and appropriate for important and sensitive criminal justice applications.

2. **Training:** Effective use of generative AI tools in criminal justice requires more than basic technical proficiency—it demands specialist training in both how to interact with the tools and how to critically evaluate their products. Although prompt engineering—the process of asking exact, contextually relevant questions—remains key, users also need to be trained to act as the human-in-the-loop to ensure outputs are correct, equitable, and appropriate for use in high-stakes settings. This training entails developing the ability to recognize when a model is likely to have hallucinated, how to recognize signs of model drift over time, and how to audit AI-generated content for alignment with legal standards and institutional protocols. This ability requires ongoing education, practical training, and clear organizational guidance to achieve responsible and effective AI integration into criminal justice applications.
3. **Funding (Including Maintenance and Updates):** The dynamic nature of the criminal justice system mandates regular updates and maintenance of generative AI tools to ensure continued relevance, fairness, and effectiveness. Vendors may shoulder this burden for subscription-based tools, but for agency-tailored solutions, organizations must integrate the cost and logistics of these activities into broader strategic, operational, and fiscal plans.
4. **Data Management**
   - **Data Access and Ownership:** Before generative AI tools are used in criminal justice settings, guidelines must be established regarding access, use, and ownership of data and results by stakeholders and AI providers. These guidelines apply to both input data, which may include sensitive or nondeidentified data, and results generated by the AI, such as synthesized summaries, legal inferences, or case-related insights. Without definite agreements, data or results could be exploited, stored without the consent of the subject, or redirected to other purposes without adequate surveillance. Having definition on such terms is vital to maintaining privacy, data integrity, and public trust.

Governance Considerations

The implementation of generative AI in the criminal justice system entails several governance considerations, primarily focused on fairness, legal and regulatory, and community.

1. **Fairness**
   - **Human Bias:** Biases that developers, data annotators, or users might have can be explicit or implicit and of personal, cultural, or societal origins. Such biases can affect the design, training, or tuning of generative AI systems, and through this process, might generate outputs that mirror or perpetuate these biases. Bias might also occur at runtime, as users might craft prompts that capture implicit assumptions or biased perspectives. Because LLMs generally align with the user's intent, they tend to amplify biased prompts by creating responses that support or agree with those views, a phenomenon referred to as sycophantic behavior. Sycophantic behavior implies critical consciousness and diligence when generative AI is used in sensitive domains like criminal

justice. For example, if one poses a leading question to a model concerning the credibility of a particular witness type (e.g., "Why are teenage witnesses typically not credible?"), the model may affirm the assumption rather than disprove it.

- **Algorithmic Bias:** This bias occurs when AI algorithms yield systematically discriminatory results, often due to their design or the data on which they were trained. For example, LLMs trained predominantly on data from a particular demographic group may show a bias toward or against that group. As part of algorithmic bias, LLMs are susceptible to sycophancy bias, where the model echoes or agrees with the user input—regardless of whether it is true or neutral—and thus reinforcing biased assumptions or misinformation.
- **System Bias:** This bias encompasses those embedded in the broader systems within which AI operates, including data collection, deployment environments, and feedback loops. For example, if a generative AI tool is trained using historical police reports or case records that have been subject to charging disparities or disparities in arrest rates for specific racial or socioeconomic populations, the model can learn and perpetuate the same patterns—such as associating certain neighborhoods or demographics with higher rates of criminality. These findings are not due to prejudices within the algorithm itself, but rather the systemic inequalities embedded within the data and processes of its implementation. Systemic biases can cause a generative AI system to produce consistently biased outputs in specific contexts or for particular groups.
- **Anthropomorphic Bias, Automation Bias, and Model Capability:** The hyper-realistic outputs of generative AI–powered tools can induce anthropomorphic bias, wherein users attribute human-like understanding to these systems, which fundamentally operate on statistical prediction rather than reasoning. This mischaracterization can lead to overreliance on AI, known as automation bias, where users uncritically accept AI-generated content and overly trust the system's outputs, even in high-stakes criminal justice contexts. Such biases can result in flawed decision-making and can have potentially serious consequences. To mitigate these risks, stakeholders must be trained on the tools' actual capabilities and limitations, and robust verification protocols for AI-generated information must be implemented to ensure that AI is used as a support tool rather than an unquestioned authority.

2. **Community**

- **Transparency and Explainability:** Transparency and explainability are essential for building trust, accountability, and community-centered decision-making in generative AI systems, particularly in criminal justice systems. Transparency involves providing clear, accessible information on how a model has been developed and how it functions—including details about training data sources, design choices, embedded assumptions, and known limitations—so that stakeholders can anticipate and understand the system's potential impacts. Explainability refers to the ability to trace how a particular output was generated from a given input, providing insights on the model's logic or reasoning. These qualities are especially important in the context of the criminal justice system, where communities can be directly affected by decisions made with the aid of generative AI. However, transparency and explainability are often difficult to achieve with deep-learning models, which act like black boxes, and with proprietary systems, in which internal mechanisms are not disclosed.[5] In criminal justice, where accountability is critical, practitioners must weigh the benefits of generative AI against the need for clear, explainable, and accountable outputs to maintain ethical and legal standards.

3. **Legal and Regulatory Compliance:** Users of generative AI in the criminal justice realm must navigate a complex landscape of laws and regulations unique to this sector.[25] These legal frameworks and laws are constantly evolving and address topics such as ensuring the admissibility of generative AI-generated evidence in court and adhering to local, state, and federal regulations that may restrict the use of new technologies in certain contexts, such as policing.

The technical, operational, and governance implementation considerations presented in this section are addressed in varying degrees within the current laws and policies being introduced or implemented in the United States and around the world to define guidelines for the development and use of AI systems in both the public and private sectors. In recognition of the regulatory complexity of AI, federal, state,[26] and local government stakeholders and international institutions are developing frameworks, standards, and oversight structures that promote responsible innovation.[27] For instance, the Toolkit for Responsible AI Innovation in Law Enforcement,[28] developed by Interpol and UNICRI with support from the European Union, provides guidance to help agencies implement AI while safeguarding human rights, ethics, and policing principles. Although approaches vary, there is a shared understanding that generative AI tools—particularly when used in sensitive areas like criminal justice—must be guided by principles that ensure transparency, fairness, accountability, and protection of individual rights. These efforts reflect growing commitment to establishing the terms under which AI can be used safely and ethically across sectors.

With the changing policy climate, organizations must remain vigilant and adaptive. Technical innovation is not enough; successful implementation rests on a solid foundation of organizational readiness, legal clarity, and human-centered design. Agencies that adopt generative AI technologies in criminal justice environments will need to take great care in considering how the technologies map to existing mandates and values, implement oversight frameworks, and prioritize the needs and rights of the populations they serve. An engaged, informed, and ethical approach is critical to unlocking the benefits of AI while minimizing unintended harms.

Considerations	Questions to Ask
Purpose and Goals	☐ Where in criminal justice workflows might AI address inefficiencies and limitations? ☐ How might generative AI support or enhance the work currently being done by practitioners? ☐ What are the key performance indicators (KPIs) that will help measure the success of implementing this tool?
Technical Considerations	**Quality Monitoring** ☐ What processes will be used to regularly assess the accuracy and reliability of generative AI outputs? **Compatibility and Integration** ☐ Does our agency have the necessary hardware and software infrastructure to support the deployment of generative AI tools? ☐ How will generative AI integrate with existing systems and workflows without causing disruption?

Considerations	Questions to Ask
Operational Considerations	**Policy and Procedure** ☐ How will data ownership and access be managed between our agency and external vendors providing generative AI solutions? ☐ What documentation and standard operating procedures (SOPs) are necessary for effectively integrating generative AI into our existing workflows? ☐ What protocols will be established to ensure ongoing accountability and oversight of AI-generated decisions? ☐ How will unauthorized access to the outputs and insights generated by these tools be prevented? **Workforce and Culture** ☐ Who within our agency will be responsible for overseeing the legality, use, and outputs of generative AI tools? **Training** ☐ How will staff be trained to use generative AI tools effectively and responsibly (from identification of use cases to effective prompt engineering and critical evaluation of AI outputs)? ☐ How can practitioners be trained to recognize and manage their own biases and the limitations of the AI tools they use? ☐ What continuous training programs will be established to ensure our workforce stays updated on best practices for generative AI usage? **Data Management** ☐ Does our agency have clear data governance policies to manage the flow of data between our agency and generative AI tool vendors? ☐ What measures ensure that any sensitive data used by generative AI tools are securely handled and stored? ☐ What specific data sets will be used by the generative AI tool, and how will access to these data be controlled and monitored? **Funding** ☐ What are the initial and ongoing costs associated with implementing and maintaining generative AI tools, and how will they be funded? ☐ Are grants or other funding sources available to support the adoption of generative AI in our agency?
Governance Considerations	**Fairness** ☐ How can we identify and mitigate individual, algorithmic, and systemic biases within generative AI tools? ☐ What measures can we implement to continuously improve the fairness of AI-generated outcomes? ☐ How can we communicate fairness considerations effectively to the communities served? ☐ What ethical guidelines should we follow to ensure responsible AI usage, particularly concerning bias and overreliance on AI decisions? **Legal and Regulatory** ☐ What legal frameworks govern the use of generative AI in our specific jurisdiction? ☐ How can compliance with constitutional rights, such as privacy and due process, be ensured or further protected by the use of generative AI tools? **Community** ☐ How can we engage the community to address concerns about the transparency and fairness of generative AI tools? ☐ What systems can we implement to ensure transparency and explainability in the decisions or outputs generated by AI? ☐ How can we measure and communicate the accuracy and error rates of AI tools to build trust with the public?

FUTURE OUTLOOK

As generative AI technology evolves, generative AI tools are set to revolutionize various roles across and throughout all branches of the criminal justice system. Many general and criminal justice–specific generative AI tools are already operational, whereas others are being developed to automate, augment, and accelerate a range of tasks—from administrative automation to decision support—and paving the way for new workflows. The exploration, evaluation, and deployment of these tools should be approached with a mix of skepticism and caution, considering the full spectrum of technical, operational, and governance factors highlighted in this report and associated resources. To significantly and consistently improve outcomes within the criminal justice system using generative AI, a collaborative approach involving practitioners, technologists, legal experts, and policymakers is essential to ensure that these tools align with the broader goals of the criminal justice system, balancing benefits against potential risks.

APPENDIX I: GLOSSARY

Application program interface (API)
An API is a set of rules or protocols that enables software applications to communicate with each other to exchange data, features, and functionality.[29]

Autoregressive models
In generative AI, an autoregressive model generates each output token (e.g., word, character) one at a time, using previously generated tokens as input to predict the next one. These models are trained to predict the next element in a sequence based on prior context, enabling coherent generation of text, code, or other sequential data. For example, language models like GPT use an autoregressive approach to autocomplete sentences by predicting each next word based on the words that came before.[30]

Bias
In the context of AI, bias refers to systematic differences in an AI system's behavior or predictions that arise from a range of sources, including human, systemic, and statistical or computational factors. These forms of bias may be introduced during data collection, model design, deployment, or usage and can lead to outcomes that are inaccurate, unfair, or discriminatory. Bias is not always the result of intent; it can emerge inadvertently through institutional practices, skewed training data, algorithmic assumptions, or cognitive heuristics. Effective management of AI bias requires a sociotechnical approach that accounts for the full life cycle of AI systems and the broader social context in which they operate.[31]

Blind spots
Blind spots are data gaps resulting from missing or incomplete training data within specific domains or due to knowledge cutoffs.

Chatbot
A chatbot is a computer program that simulates human conversation with an end user.[32]

Closed-source foundation model
This model is a type of AI model where access to its architecture, weights, and training data is restricted, typically by the organization that developed it. Unlike open-source models, closed-source models limit public access to prevent modification or misuse.[33]

Cloud-based application
Also known as software-as-a-service (SaaS), cloud-based application is application software hosted in the cloud. Users access the application through a web browser, a dedicated desktop client, or an API that integrates with a desktop or mobile operating system.[34]

Cloud platform
Cloud platforms are platforms by which cloud service providers provide on-demand access of computing resources—physical servers or virtual servers, data storage, networking capabilities, application development tools, software, AI-powered analytic tools, and more—over the internet.[34]

Context window
Context window refers to the maximum amount of text an LLM can process at one time, including the user's prompt. It defines how much context the model can use to generate relevant and coherent responses.[22]

Contrafactual bias
A form of bias specific to LLMs that are trained to respond to questions or commands, contrafactual bias is when the model accepts a false assumption in a prompt as true and responds accordingly on the basis of that false assumption. This behavior can result in outputs that sound reasonable but are misleading or inaccurate.[35]

Discriminative AI
Algorithms and models that are well suited to classification tasks due to their focus on modeling boundaries between specific classes of data aiming to predict the conditional probability of a given data point falling into a certain class (e.g., car vs. not car).[36]

End-to-end software application
This type of application is where a single provider supplies all the necessary software and hardware components to meet a customer's needs, without requiring any involvement from other vendors.

Federated learning
Federated learning is a decentralized method for training AI models collaboratively across multiple devices or servers, without data leaving the local environment. Each participant trains a model locally on their private data, which aggregates the improvements.[37]

Fine-tuning
This process of adapting a pre-trained model for specific tasks or use cases allows the model to retain general knowledge from its initial training while refining its performance for specialized tasks.[38]

Foundation model
Machine learning models trained on a broad spectrum of generalized and unlabeled data and capable of being adapted to perform a wide variety of general tasks such as language processing and analysis, generating text and images, and generating natural language.[39]

Generative AI
This class of deep-learning models that can generate new content, such as text, images, or other data types, by learning patterns from existing data sets. These models create outputs that resemble the data they were trained on without replicating them exactly.[2]

Hallucination
A phenomenon wherein a LLM perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate.[40]

Homogenization
In the context of foundation models, homogenization refers to the consolidation of methodologies and architectures across various AI applications. This process provides efficiency by using a single foundation model for diverse tasks but creates risks, as any flaws or biases in the foundational model can be inherited by all downstream applications built from it.[5]

Human-labeled data
These raw data (e.g., text, images, videos) have been annotated with tags or labels by humans to specify its context. These labels guide machine learning models during training, enabling them to learn patterns and make accurate predictions.[41]

Knowledge cutoff
In the context of foundation models, knowledge cutoff is the date at which model training data or knowledge base does not cover any online content released after that point.[42]

Large language model (LLM)
This foundation model is trained on vast amounts of text data for the purpose of predicting and creating language outputs based on patterns observed during training.[43]

Model drift
Model drift is a phenomenon where a machine learning model's performance degrades over time due to changes in the environment, data, or conditions it was trained on. This occurs because the model no longer aligns with the real-world scenario it is meant to operate in.[44]

Multimodal
Multimodal refers to machine learning models capable of processing and integrating information from multiple modalities or types of data (e.g., text, images, audio, video, computer code).[45]

Natural language processing (NLP)
This subfield of computer science and AI uses computers to process, interpret, and analyze human language.[46]

Nonautoregressive
In the context of generative AI, a nonautoregressive model generates output tokens in parallel rather than one at a time, as opposed to autoregressive models that generate each token based on previously generated ones. This parallel generation enables faster inference but may reduce output quality for complex tasks. For example, in machine translation, a nonautoregressive model might generate an entire sentence at once using the overall context, rather than building it word by word.[47]

Open-source foundation model
Open-source foundation model is a type of foundation model whose architecture and weights are publicly accessible. This approach enables external developers to study, modify, and build on the model, fostering collaboration and innovation.[33]

Probabilistic
In the context of LLMs, probabilistic refers to how responses are generated by sampling from a probability distribution over possible next tokens, rather than deterministically choosing the single most likely token. These probabilities are computed using a softmax function, which converts internal model scores into a distribution over possible outputs. Many LLMs incorporate a temperature parameter to adjust the randomness of sampling—higher temperatures produce more-varied outputs, whereas lower temperatures yield more-focused, deterministic responses. This probabilistic nature is central to how generative AI creates coherent and contextually appropriate language without true understanding.[48]

Prompt
A prompt is a user's input into an AI system, such as a chatbot application powered by an LLM, to obtain specific results.[49]

Prompt engineering
Prompt engineering is the process of designing and refining prompts to guide generative AI models in producing desired outputs.[50]

Reinforcement learning from human feedback
This machine learning technique uses human feedback to train machine learning models. Reinforcement learning techniques train models to provide outputs that maximize rewards, making their outcomes more accurate.[51]

Self-supervised learning
In this machine learning technique, a model is trained using augmented input data in lieu of labeled output data, reducing the need for human-annotated data sets.[52]

Supervised learning
In this machine learning technique, a model is trained using labeled output data.[53]

Synthetic data
Synthetic data refers to computer-generated information used to train or test AI models when real-world data are scarce, difficult to obtain, or sensitive.[54]

Token
In the context of LLMs, token refers to machine-readable representation of words, parts of words, or even punctuation.[55]

Web application
Also known as web app, a web application is a software program hosted on a web server and accessed through a web browser on a user's device.[56]

Web-enabled (native) software application
Also known as hybrid apps, this type of software is installed directly on a user's device and uses web technologies to access remote data or services.[56]

APPENDIX II: GENERATIVE AI BASICS

Technical Basics: Architecture and Access

Users typically interact with these tools through various interfaces, including web applications like chatbots or native software applications with application program interfaces (APIs) that provide access to AI models (as seen in **Figure 2**). In these interfaces, users provide requests for specific tasks (such as answering questions or generating content) in natural (i.e., plain) language, referred to as a prompt. The application sends the prompt to an AI model, which processes it and generates the desired output. Depending on the model type, this process could involve predicting the output word by word, as with autoregressive models like GPT-4, or generating the entire output at once, as seen in nonautoregressive models, which are often used for tasks like translation. These models generally require significant computational resources to run and are often hosted on cloud platforms that provide access to the processing resources required for development and deployment.

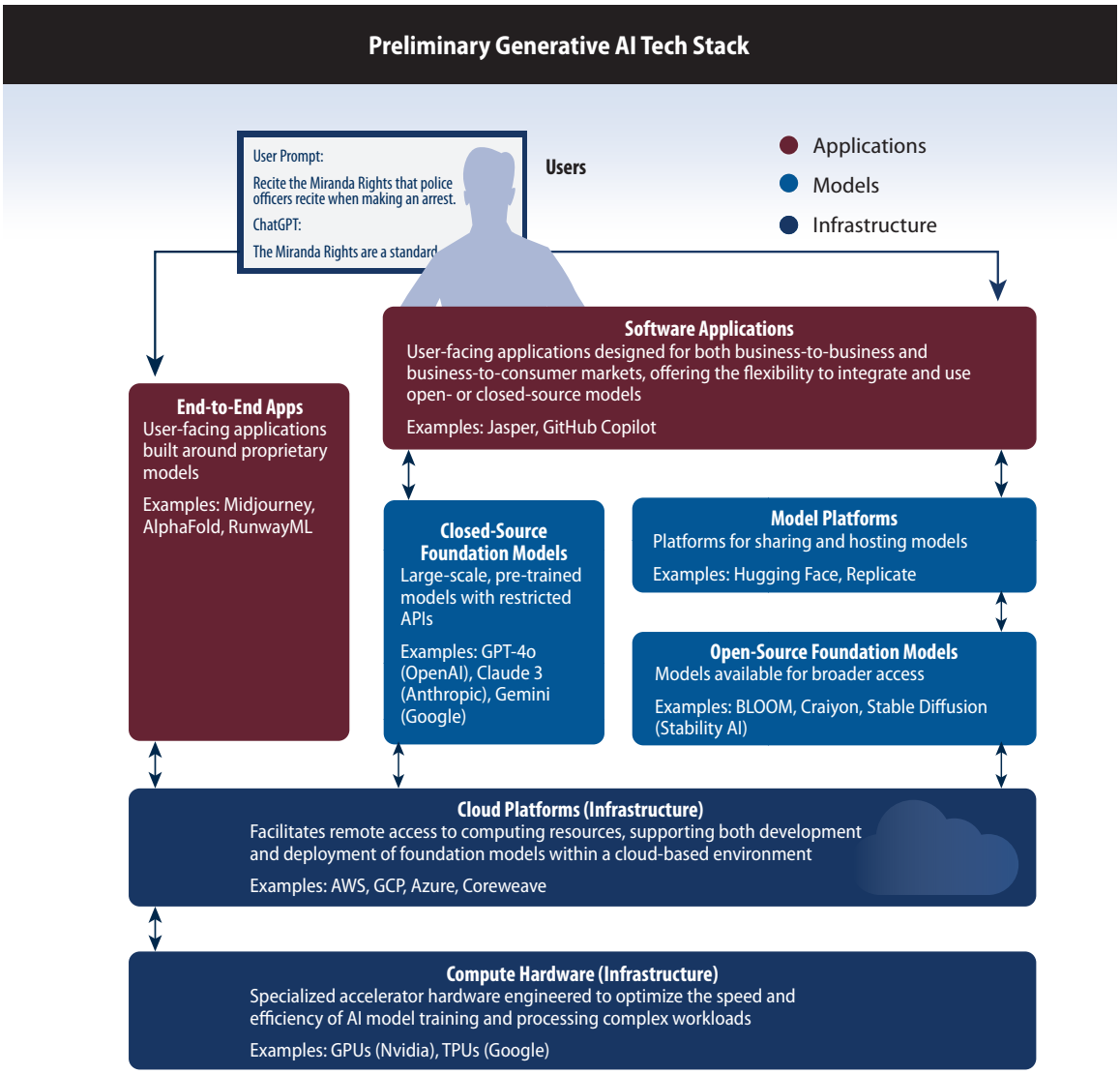

Figure 2: Overview of user experience using a generative AI–powered chatbot (ChatGPT) with an example prompt and response from the tool.

Prominent publicly accessible generative AI tools include OpenAI's ChatGPT,[57] Anthropic's Claude,[58] and Google's Gemini,[59] which transform text prompts into human-like responses. Tools like Midjourney[60] and OpenAI's DALL-E[61] create images from text descriptions. Many recent generative AI models are multimodal, meaning they can interpret and generate content across different types of data, such as text, images, and audio, enhancing their versatility and application in various fields. For example, ChatGPT-4 and Claude can handle both image and text inputs, enhancing analysis and transformation capabilities. Enterprise software providers like Microsoft, Google, and Salesforce are already beginning to integrate generative AI into their suites, exemplified by Windows' Copilot and Salesforce's Einstein GPT, boosting daily workflows.

Technical Basics: Foundation Models and Training

Foundation models, which form the core of generative AI tools, are trained on broad data sets and can be adapted for various tasks.[5] Open-source foundation models, which are publicly accessible and can be modified by anyone, allow users to tailor AI tools to specific needs, fostering innovation and collaboration within the field. Closed-source foundation models, on the other hand, are proprietary and controlled by specific organizations, often providing more-robust support and security and ensuring compliance with legal standards. Many of these foundation models, including LLMs, learn patterns from vast data sources (e.g., the entirety of Wikipedia up to a certain date) through self-supervised learning, which involves the AI system itself (i.e., with very limited human involvement) predicting or reconstructing parts of training data to capture complex relationships.[5] This method enables rapid scaling of training and provides foundation models with a broad and generalizable knowledge base. **Figure 3** highlights a generic overview of foundation model training.

Techniques such as supervised fine-tuning and reinforcement learning from human feedback (RLHF) further refine these models for specific applications. Supervised fine-tuning involves training the model on labeled data to improve its performance on particular tasks, ensuring it meets application-specific needs (e.g., the needs of criminal justice practitioners). RLHF enhances this process by incorporating feedback from human evaluators, aligning the model's outputs with human values and preferences. Rejection sampling is used during fine-tuning to filter out low-quality or irrelevant outputs, ensuring that only high-quality responses are used for further training. This sampling method helps in maintaining high standards of accuracy and relevance in the model's performance. This iterative process of generating outputs, receiving feedback, and updating the model helps create a robust and reliable AI tool tailored for a particular application. Unlike earlier AI models that were tailored for hyper-specific tasks and trained using only human-labeled data, foundation models are adaptable to diverse applications, making them highly valuable in addressing complex and evolving needs.[62]

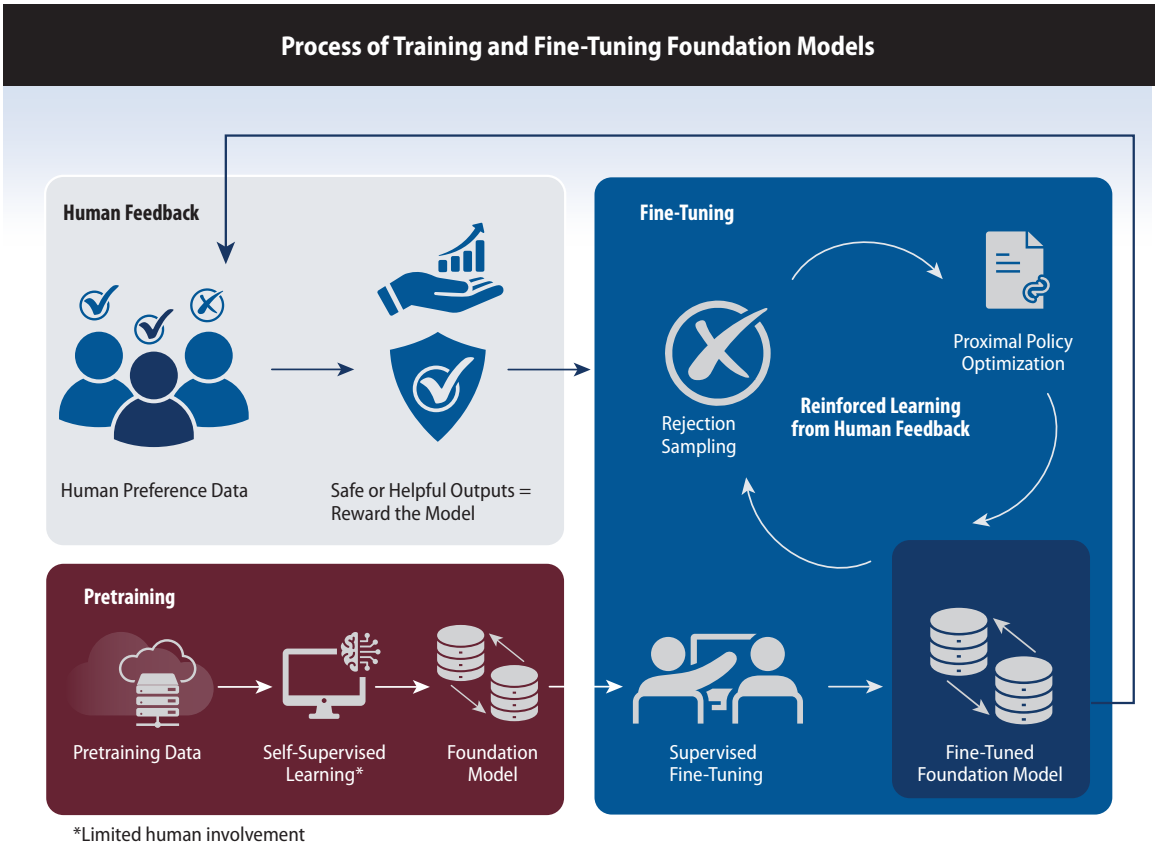

Figure 3: General overview of how foundation models are trained using three different types of learning and capability of foundation models to generate a variety of modal outputs.

Advanced AI Concept Considerations for Criminal Justice Practitioners

For criminal justice practitioners, the integration of advanced AI concepts like synthetic data generation and federated learning can significantly enhance the effectiveness and security of generative AI applications. Synthetic data generation focuses on creating artificial data that closely mimic real-world data and that can be used to train foundation models. This approach allows vendors and criminal justice system stakeholders to train generative AI systems on sensitive scenarios—such as crime pattern analysis or suspect identification—without using sensitive real-world data, thereby preserving privacy and overcoming data scarcity challenges. Federated learning enables the creation and continuous improvement of AI systems by leveraging decentralized data sources while maintaining data privacy and security. Implementing these advanced AI concepts typically requires collaboration with specialized vendors who can provide the necessary expertise and infrastructure. Understanding and implementing these advanced AI concepts is crucial for criminal justice professionals looking to leverage generative AI's full potential while addressing the unique challenges of data sensitivity, privacy, and interdepartmental collaboration.

REFERENCES

1. Redden, J., Dix, M. O., and Criminal Justice Testing and Evaluation Consortium. (2020). *Artificial intelligence in the criminal justice system*. RTI International. https://cjttec.org/artificial-intelligence-in-the-criminal-justice-system
2. Martineau, K. (2023, April 20). *What is generative AI?* IBM Research Blog. https://research.ibm.com/blog/what-is-generative-AI
3. Adams, I. T., Barter, M., McLean, K., Boehme, H. M., & Geary, I. A. (2024). No man's hand: Artificial intelligence does not improve police report writing speed. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-024-09644-7
4. Brighter AI. (n.d.). https://brighter.ai/
5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Catellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . , & Liang, P. (2022). *On the opportunities and risks of foundation models*. Stanford University. https://crfm.stanford.edu/report.html
6. Al Naqbi, H., Bahroun, Z., & Ahmed, V. (2024). Enhancing work productivity through generative artificial intelligence: A comprehensive literature review. *Sustainability, 16*(3), 1166. https://doi.org/10.3390/su16031166
7. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y. Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . , & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71. https://doi.org/10.1016/j.ijinfomgt.2023.102642
8. Wikipedia. (n.d.). *Racial bias on Wikipedia*. https://en.wikipedia.org/w/index.php?title=Racial_bias_on_Wikipedia&oldid=1227157155
   Wikipedia. (n.d.). *Gender bias on Wikipedia*. https://en.wikipedia.org/w/index.php?title=Gender_bias_on_Wikipedia&oldid=1239493218
   Wikipedia. (n.d.). *Ideological bias on Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ideological_bias_on_Wikipedia&oldid=1239456587
9. Microsoft. (2025, January 27). *Azure, Dynamics 365, Microsoft 365, and Power Platform services compliance scope*. https://learn.microsoft.com/en-us/azure/azure-government/compliance/azure-services-in-fedramp-auditscope
10. Perplexity. (2024, June 28). *The skeleton key AI jailbreak*. https://www.perplexity.ai/page/the-skeleton-key-ai-jailbreak-OuIr1gvxRQ0002Bu6ZBI1Q
11. Early, W. (2024, December 4). Anchorage police not moving forward with using AI to write reports, for now. *Alaska Public Media*. https://alaskapublic.org/news/2024-12-04/anchorage-police-not-moving-forward-with-using-ai-to-write-reports-for-now
12. Guariglia, M. (2025, March 12). *Anchorage Police Department: AI-generated police reports don't save time*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2025/03/anchorage-police-department-ai-generated-police-reports-dont-save-time
13. Chohlas-Wood, A., Nudell, J., Yao, K., Lin, Z. (Jerry), Nyarko, J., & Goel, S. (2021). Blind Justice: Algorithmically Masking Race in Charging Decisions. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 35–45. https://doi.org/10.1145/3461702.3462524
14. Adams, I. T., McLean, K., & Alpert, G. (2024). Improving police behavior through artificial intelligence: Pre-registered experimental results in two large US agencies. *CrimRxiv*. https://doi.org/10.21428/cb6ab371.160e914f
15. Guthrie Ferguson, A., (2024). *AI-assisted police reports and the challenge of generative suspicion*. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=4897632
16. Wes Davis, W. (2023, May 27). A lawyer used ChatGPT and now has to answer for its 'bogus' citations. *The Verge*. https://www.theverge.com/2023/5/27/23739913/chatgpt-ai-lawsuit-avianca-airlines-chatbot-research
17. Weise, K. & Metz, C. (2023, May 1). When A.I. chatbots hallucinate. *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html
18. Koenecke, A. Choi, A. S. G., Mei, K. X., Schellmann, H., Sloane, M. (2024). Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658996
19. OpenAI. (2025). *Model release notes.* https://help.openai.com/en/articles/9624314-model-release-notes
20. Google AI for Developers. (n.d.). *Gemini models*. https://ai.google.dev/gemini-api/docs/models
21. Anthropic. (2025). *How up-to-date is Claude's training data?* https://support.anthropic.com/en/articles/8114494-how-up-to-date-is-claude-s-training-data
22. Bergmann, D. (2024, November 7). *What is a context window?* IBM Research Blog. https://www.ibm.com/think/topics/context-window
23. Gemini Apps Help. (updated 2025, April 15). *Gemini Apps privacy hub*. Accessed May 23, 2024. https://support.google.com/gemini/answer/13594961?hl=en
24. OpenAI. (2024). *Privacy policy*. https://openai.com/policies/privacy-policy/
25. Harris, J. (2024, September 27). King County's new policy: No AI in police report writing over error concerns. *KOMO*. https://komonews.com/news/local/king-county-prosecutor-tells-police-not-to-use-ai-artificial-intelligence-for-official-reports-for-now-errors-concerns-law-enforcement-perjury-criminal-justice
26. Sentinella, R., & Zweifel-Keegan, C. (2025). *US state AI governance legislation tracker*. IAPP. https://iapp.org/resources/article/us-state-ai-governance-legislation-tracker/
27. IAPP. (2024). *Global AI law and policy tracker*. https://iapp.org/resources/article/global-ai-legislation-tracker/
28. Advancing Innovation, Governance and Responsible AI in Law Enforcement (AI-POL). (n.d.). *Welcome to your responsible AI innovation journey*. https://www.ai-lawenforcement.org/
29. Goodwin, M. (2024, April 9). *What is an API (application programming interface)?* IBM Research Blog. https://www.ibm.com/topics/api
30. Noble, J. (2024, June 12). *What is an autoregressive model?* IBM Research Blog. https://www.ibm.com/think/topics/autoregressive-model
31. Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence*. NIST. https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence
32. IBM Research Blog. (2021, October 15). *What is a chatbot?* https://www.ibm.com/topics/chatbots
33. Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Heigeartaigh, S. O., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., . . . , Gupta, A. (2023). *Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models
34. Susnjara, S., & Smalley, I. (2025, February 10). *What is cloud computing?* IBM Research Blog. https://www.ibm.com/topics/cloud-computing

35. Ho, D. E. (2024). *Hallucinating law: Legal mistakes with large language models are pervasive*. Stanford University. https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive
36. IBM Research Blog. (n.d.). *What is an AI model?* https://www.ibm.com/topics/ai-model
37. Martineau, K. (2022, August 24). *What is federated learning?* IBM Research Blog. https://research.ibm.com/blog/what-is-federated-learning
38. Bergmann, D. (2024, March 15). *What is fine-tuning?* IBM Research Blog. https://www.ibm.com/topics/fine-tuning
39. Amazon Web Services. (n.d.) *What are foundation models?* https://aws.amazon.com/what-is/foundation-models/
40. IBM Research Blog. (2023, September 1). *What are AI hallucinations?* https://www.ibm.com/topics/ai-hallucinations
41. IBM Research Blog. (n.d.). *What is data labeling?* https://www.ibm.com/topics/data-labeling
42. Belcic, I., & Stryker, C. (2024, September 18). *What is GPT (generative pretrained transformer)?* IBM Research Blog. https://www.ibm.com/think/topics/gpt
43. IBM Research Blog. (2023, November 2). *What are large language models (LLMs)?* https://www.ibm.com/topics/large-language-models
44. Holdsworth, J., Belcic, I., & Stryker, C. (2024, July 16). *What is model drift?* IBM Research Blog. https://www.ibm.com/topics/model-drift
45. Stryker, C. (2024, July 15). *What is multimodal AI?* IBM Research Blog. https://www.ibm.com/think/topics/multimodal-ai
46. Stryker, C., & Holdsworth, J. (2024, August 11). *What is NLP (natural language processing)?* IBM Research Blog. https://www.ibm.com/topics/natural-language-processing
47. Xiao, Y., Wu, L., Guo, J., Li, J., Zhang, M., Qin, T., & Liu, T.-Y. (2023). A survey on non-autoregressive generation for neural machine translation and beyond. *Journal of Latex Class Files*. https://arxiv.org/pdf/2204.09269
48. Moyer, P. (2023, October 25). *The prompt: Probability, data, and the gen AI mindset*. Google Cloud Blog. https://cloud.google.com/transform/prompt-probability-data-and-the-gen-ai-mindset
49. MIT Sloan Teaching & Learning Technologies. (2025). *Effective prompts for AI: The essentials*. https://mitsloanedtech.mit.edu/ai/basics/effective-prompts/
50. IBM Research Blog. (n.d.). *What is prompt engineering?* https://www.ibm.com/topics/prompt-engineering
51. Amazon Web Services. (2025). *What is RLHF?* https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/
52. Bergmann, D. (2023, December 5). *What is self-supervised learning?* IBM Research Blog. https://www.ibm.com/topics/self-supervised-learning
53. Belcic, I., & Stryker, C. (2024, December 28). *What is supervised learning?* IBM Research Blog. https://www.ibm.com/topics/supervised-learning
54. Martineau, K., & Feris, R. (2023, February 8). *What is synthetic data?* IBM Research Blog. https://research.ibm.com/blog/what-is-synthetic-data
55. Martineau, K. (2024, July 24). *Why larger LLM context windows are all the rage*. IBM Research Blog. https://research.ibm.com/blog/larger-context-window
56. Amazon Web Services. (2025). *What's the difference between web apps, native apps, and hybrid apps?* https://aws.amazon.com/compare/the-difference-between-web-apps-native-apps-and-hybrid-apps/
57. OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/index/chatgpt/
58. Anthropic. (n.d.). *Claude*. https://www.anthropic.com/claude
59. Google DeepMind. (n.d.). *Gemini*. https://deepmind.google/technologies/gemini/
60. Midjourney. (n.d.) *About*. https://www.midjourney.com/
61. OpenAI. (n.d.). *DALL·E 3*. https://openai.com/index/dall-e-3/
62. For more on the technical aspects, review the following resources:
    Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Catellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . ., & Liang, P. (2022). *On the opportunities and risks of foundation models*. Stanford University. https://crfm.stanford.edu/report.html
    Susnjara, S., & Smalley, I. (2025, February 10). *What is cloud computing?* IBM Research Blog. https://www.ibm.com/topics/cloud-computing
    Ashish Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I.(2017; updated 2023). *Attention is all you need*. arXiv. https://arxiv.org/abs/1706.03762
    Google Cloud Tech. (2023, May 8). *Introduction to generative AI* [Video]. YouTube. https://www.youtube.com/watch?v=G2fqAlgmoPo
    Google Cloud Tech. (2023, May 8). *Introduction to large language models* [Video]. YouTube. https://www.youtube.com/watch?v=zizonToFXDs