City & County of San Francisco
London N. Breed, Mayor

Office of the City Administrator
Carmen Chu, City Administrator

San Francisco Generative AI Guidelines

Revised November 9, 2023

Introduction

Artificial Intelligence (AI) has great potential to provide public benefits, when used responsibly. Recently, Generative AI technology has gained mainstream attention and become available for use by staff of the City and County of San Francisco (City). Generative AI generates new data based on patterns learned from existing data and can produce content that mimics human creativity. Examples include text generation, image creation, and music composition. Generative AI differs from AI technology currently in use by the City, which supports informed decisions based on input data but does not create new content. Generative AI offers new opportunities and also poses unique challenges to ensure responsible and effective use.

Scope of Guidelines

The following guidelines apply to all city department personnel, including employees, contractors, consultants, volunteers, and vendors while working on behalf of the City. The guidelines will evolve based on legislative and regulatory developments and changes to Generative AI technology. The City Administrator's Office will update the guidelines as advancements, use cases and new information emerge.

Current Legislative and Regulatory Landscape

Governments, researchers, and tech policy experts are closely watching the evolution of Generative AI tools to understand potential risks and benefits for public service. In October 2023, President Biden issued an Executive Order aimed at improving the safety and security of AI in the public and private sectors. At the state level, Governor Newsom issued an Executive Order in September 2023 directing state agencies to study the development, use, and risks of AI and develop a process for deployment within California's government. Both federal and state legislatures have also debated numerous bills related to regulating the use of AI. In view of the early state of Generative AI, ongoing efforts at the federal and state level and the complexity of city operations and potential AI use cases, the City will continue to evaluate the field before issuing more proscriptive policies governing AI.

Initial Development of Guidelines

The City Administrator's Office (led by Digital & Data Services, the Department of Technology, and the Committee on Information Technology) developed these Generative AI-focused guidelines after reviewing recent Generative AI guidance issued by Boston, San Jose, the United Kingdom, the White House Office of Science and Technology Policy, and the Office of Governor Newsom.

The City and County of San Francisco embraces innovation with responsible and equitable use. Several city departments currently use various types of AI to support service delivery. For example, SF311 utilizes AI to categorize descriptions and photos submitted by the public to accelerate responses to service requests. The Office of the Assessor-Recorder uses AI to predict property prices and identify properties that require a full appraisal.

Evolution of Guidelines

As federal and state legislative and regulatory frameworks continue to evolve, San Francisco provides these preliminary guidelines for staff using Generative AI in city operations. City employees must understand and remain aware of both potential risks and benefits as the technology and the policies that govern it change. These guidelines are designed to provide sufficient guidance for employees to use the tools in a responsible manner, enhancing public service without stifling innovation.

This document represents a first step in an extended process to understand, test, and evaluate the use of AI broadly within San Francisco city government. Next steps include a comprehensive survey of current and proposed city department use of AI, meetings with experts in the field of AI and the creation of a user community, among other tasks, to maximize the benefits and minimize the risks of Artificial Intelligence in delivering service to San Francisco's residents and visitors.

Top 3 Guidelines for Exploring with Generative AI

- **Always** review and fact check AI-generated content before using it
- **Always** disclose usage of Generative AI in your output
- **Never** enter sensitive information into public Generative AI tools, like ChatGPT. The information you enter can be viewed by the companies that make the tools and, in some cases, members of the public

What are the benefits for City government?

Generative AI tools, used appropriately, have the potential to expand San Francisco's toolkit for public service. Text, code, and image-generation features can speed or improve common tasks when used carefully.

For example, used within guidelines, generative AI tools may assist with:

- Creating first drafts of documents, plans, memos, and briefs
- "Translating" text into levels of formality, reading levels, etc.
  - Rewriting an informal email into a draft of a memo
  - Summarizing technical or legal documentation in plain language and targeting summaries for different audiences
  - Turning frustrated thoughts into a polite interdepartmental request
- Repetitive coding and testing tasks for software developers, with appropriate engineering reviews
- Generating diagrams or other explanatory images
- Developing service interfaces such as chatbots with appropriate attention to language access and accuracy

These benefits all have their best effect when checked by a human who is:

1. Knowledgeable about the content and the service being provided
2. Aware of the common mistakes and limitations of Generative AI

What is Generative AI?

Generative AI refers to new software tools that can produce realistic text, images, audio, video, and other media based on a prompt provided by the user. Common generative AI applications include ChatGPT, Bard, and Dall-E. These tools use machine learning algorithms that have been trained on very large sets of text and image data culled from the internet. These models have extracted common language and image patterns from the training data and can respond to prompts quickly in a realistic way.

Generative AI applications are built using training datasets from various sources on the internet and often include gender, racial, political and other biases. As a result, Generative AI outputs can propagate biases. Additionally, even the most advanced current Generative AI tools may provide inconsistent answers to fact-based questions. Users should always check AI-generated content for accuracy, as well as for biases they may display.

What about Traditional AI?

Generative AI is distinct from Discriminative Machine Learning models which have been widely used since the early 2000s, including by the City. Discriminative Machine Learning models do not generate new content. They are limited to generating known and validated values. These models are primarily used to predict quantities (for example, predicting home prices) or to assign group membership (for example, classifying images into categories).

What are the risks of Generative AI?

Generative AI excels at creating content that appears authoritative and polished, making it easy to accept AI-generated content at face value. Without a knowledgeable person or expert system to review content for accuracy, Generative AI has the potential to mislead users and the public.

These risks are magnified if output is not labeled as created, drafted, or informed by AI. The risks can also apply when Generative AI technology is a component of other software, such as cloud business application or productivity tools, which may not be apparent to users.

Staff must use Generative AI tools with care to avoid possible negative outcomes:

- Making an inappropriate decision that affects residents based on AI-generated content
- Producing information, either to the public or internally, that is inaccurate
- Incorporating biases found in the AI's training data, resulting in inequities
- Cybersecurity problems or other errors due to the use of AI-generated code
- Exposing non-public data as part of training data sets. (Staff should assume all data entered in a Generative AI tool becomes part of the training set.)
- Inaccurately attributing AI-generated content to official SF sources

City Guidelines for Generative AI Use

To support the security of City systems and data and best serve the public, while upholding public trust, follow these Do's and Don'ts of Generative AI Usage.

✓ Do:

- Try it out! Experiment with Generative AI tools for drafting, leveling, and formatting text and explanatory images using public information
- Work with your department IT team and experiment thoroughly with various use cases before using generative AI in the delivery of programs or services
- Thoroughly review and fact check all AI-generated content (e.g. text, code, images, etc). You are responsible for what you create with generative AI assistance
- Disclose when and how generative AI was used in your output. For example:
  - "The header image was created using the AI tool MidJourney"
  - "This abstract was created using Bard, a generative AI tool"
  - "FYI, I used ChatGPT to revise this email"

✕ Don’t:

- Enter into public Generative AI tools (e.g. ChatGPT) any information that cannot be fully released to the public. This information can be viewed by the companies that make the tools and, in some cases, other members of the public. Once entered, this information becomes part of the public record. The handling and disclosure of sensitive information is already governed by several City policies, including but not limited to:
  - Charter Section 16.130, Privacy First Policy
  - Administrative Code Section 12M.2(a), Nondisclosure of Private Information
  - Campaign & Governmental Conduct Code section 3.228, Disclosure or Use of Confidential City Information.
  - The “Computers and Data Information Systems” section of the Department of Human Resource’s Employee Handbook” (January 2012, page 48)
  - Please refer to Citywide Data Classification Standard for more specifics on data classification and department responsible roles
- Publish Generative AI output (whether text, image, or code) without full knowledgeable review and disclosure
- Ask Generative AI tools to find facts or make decisions without expert human review
- Generate images, audio, or video that could be mistaken for real people, for example:
  - Making a fake photo or recording of a specific San Francisco official or member of the public (“deepfake”) – even with disclosure
  - Generating a fake image or recording which purports to be a San Franciscan or public official, even if not a specific one
  - Generating fake “respondents” or made-up profiles for surveys or other research
- Conceal use of Generative AI during interaction with colleagues or the public, such as tools that may be listening and transcribing the conversation or tools that provide simultaneous translation

Additional Guidance for Departmental IT Leaders:

Departmental IT leaders have a responsibility to support right-sized generative AI uses that deliver the greatest public benefit. IT leaders should consider the following additional guidance while working with staff to determine appropriate use cases for Generative AI.

- Expect these tools, and guidance regarding these tools, to evolve over time. This is the beginning. It is important to be aware of and track use to allow transparency with the public and ensure responsible use.

- Begin collecting use cases and be prepared to report your uses in a public forum to ensure transparency and accountability.
- Know whether software you manage – and its components – include Generative AI; inform your team how it is used and what the specific risks are.
- Ask questions about Generative AI in your procurement solicitations.
- Work with vendors to ensure that AI built into procured tools will be explainable and auditable. Vendors should be able to provide information and documentation on data sources, methods, and validation.
- Experiment with training internal models on internal data.
- When considering implementing chatbots for service to the public, thoroughly test and develop a language access plan.
- Consult with the Office of Cybersecurity early in the testing process when building or procuring applications using Generative AI technology.

Conclusion

Generative AI is rapidly developing and legislative and regulatory frameworks at the state and federal level continue to evolve. To best serve the public, uphold the public trust, and protect the security of city systems and data, city staff must be aware of the technology's potential risks and limitations, while exploring the potential benefits of Generative AI in public service delivery.

While these guidelines intend to educate city personnel about responsible usage of Generative AI, they are only the beginning. The City Administrator's Office will continue to work with the mayor, city departments, city technology vendors and external experts to support department use of AI technologies, manage risk, and protect resident and city data. Future actions include:

- Developing more detailed guidelines and training staff on specific uses of AI
- Developing ethical, transparent, and trusted AI use principles
- Defining AI governance and impact monitoring processes
- Adapting procurement processes for AI tools
- Collecting and documenting department use cases and supporting them in managing risk
- Continuing to protect City and resident data while working with technology vendors
- Seeking external expertise from other public-sector AI adopters and academics
- Sharing learnings with other government partners

The City Administrator's Office will revisit and revise these guidelines. For questions, please contact the Committee on Information Technology at COIT.staff@sfgov.org.

Glossary:

Algorithms: are a set of rules that a machine follows to generate an outcome or a decision.

Artificial Intelligence (AI): refers to a group of technologies that can perform complex cognitive tasks like recognizing and classifying images or powering autonomous vehicles. Many AI systems are built using machine learning models. For a task like image recognition, the model learns pixel patterns from a large dataset of existing images and uses these patterns to recognize and classify new images.

Auditability for AI: AI where the outputs are explainable, monitored and validated on a regular basis.

Bard: is a conversational Gen AI chatbot built by Google

Black box models: are those where you cannot effectively determine how or why a model produced a specific result.

Chatbots: are computer programs that simulate conversations. Chatbots have been around for a few decades. Basic chatbots (without Gen AI) use ML to understand human prompts and provide more-or-less scripted answers that can guide users through a process. Gen AI chatbots can provide more human-like, conversational answers.

chatGPT: is a conversational Gen AI chatbot built by OpenAI

Dall-e: is a Gen AI application that can generate images based on text prompts

Discriminative AI: In contrast to Gen AI, Discriminative AI models do not generate new content but can be used to predict quantities (for example, predicting home prices) or to assign group membership (for example, classifying images).

Generative AI (Gen AI): refers to a group of technologies that can generate new content based on a user provided prompt. Many are powered by LLMs.

Large language models (LLMs): are a type of machine learning model trained using large amounts of text data. These models learn nuanced patterns and structure of language. This allows the model to understand a user generated prompt and provide a text response that is coherent. The responses are based on predicting the most likely word in a sequence of words and as a result, the answers are not always contextually correct. The training datasets used to build these models can contain gender, racial, political and other biases. Since the models have learnt from biased data, their outputs can reflect these biases. Generative AI applications are built using these LLMs.

Machine Learning (ML): is a method for learning the rules of an algorithm based on existing data.

Machine learning model: is an algorithm that is built by learning patterns in existing data. For example, a machine learning model to predict house prices is constructed by learning from historical data on home prices. The model may learn that price increases with square footage, changes by neighborhood, and depends on the year of construction.

Model validation: methods to determine whether the outputs generated by a machine learning model are unbiased and accurate.

Training data: The dataset that is used by a machine learning model to learn the rules.