Generative AI Guidelines

Last updated: September 23, 2023

Executive Summary:

Generative Artificial Intelligence (AI) is a new branch of AI technology that can generate content—such as stories, poetry, images, voice, and music— at the request of a user. Many organizations have banned Generative AI, while others allow unrestricted usage. The City recognizes the opportunity for a controlled and responsible approach that acknowledges the benefits to efficiency while minimizing the risks around AI bias, privacy, and cybersecurity.

This is the first step in a collaborative process to develop the City's overall AI policy. Registered users will be invited to join the Information Technology Department in a working group to share their experience and co-develop the City's AI policies.

At a baseline, users must follow these rules while using Generative AI for City work, this includes direct services like ChatGPT and extensions like Compose.ai:

1. Information you enter into Generative AI systems could be subject to a Public Records Act (PRA) request, may be viewable and usable by the company, and may be leaked unencrypted in a data breach. Do not submit any information to a Generative AI platform that should not be available to the general public (such as confidential or personally identifiable information).
2. Review, revise, and fact check via multiple sources any output from a Generative AI. Users are responsible for any material created with AI support. Many systems, like ChatGPT, only use information up to a certain date (e.g., 2021 for ChatGPT).
3. Cite and record your usage of Generative AI. See how and when to cite in the "Citing Generative AI" section. Record when you use Generative AI through this form.
4. Create an account just for City use to ensure public records are kept separate from personal records. See "Getting started with Generative AI for City use." If a user agrees to

the terms and conditions of a system that the City does not have a formal agreement with, he/she is responsible for complying with those terms and conditions.

5. Departments may provide additional rules around Generative AI. Consult your manager or department contact if there are additional department-specific rules.
6. Refer to this document quarterly, as guidance will change with the technology, laws, and industry best practices. Check the "Change Log" to identify changes. Bookmark this link for easy access to the latest doc. You can subscribe to updates to the guidelines here.
7. Users are encouraged to participate in the City's established workgroups to help advance AI usage best practice in the City and enhance the Guidelines. See "Joining AI Working Group" section.

Table of Contents

Change Log

Date	Content
July 20, 2023	First Release
September 23, 2023	New Sections: 1. **"Ownership of Generated Content"**. Emphasized users' responsibility to verify, edit, and manage the generated content 2. **"Verify the Copyright of All Generated Content"**. Advised users to verify content does not infringe copyright and, if unsure, to edit before using. 3. **"Understand the Terms and Conditions".** Emphasized need for users to review terms and conditions. Edited Sections: 1. **"Create an Account Specific for City Usage".** Clarified that inputs and outputs of a Generative AI system may be subject to a PRA request. 2. **"Accuracy".** Noted that Generative AI companies do not guarantee the content they generate is accurate. Miscellaneous Edits: 1. Clarified applicability of Public Records Acts requests; 2. Added acknowledgement of Generative AI extensions; 3. Reinforced the date limit of ChatGPT; 4. Tweaked names of principles to match the evolving terminology; 5. Updated link to usage reporting form and to the latest Guidelines; 6. Updated acceptable uses of Generative AI for programming;

Definitions

User: staff, contractors, or others using Generative AI for City work purposes
City: the city government of San José located in California, United States of America
Generative AI: a machine that automatically creates content such as text, audio, or image
Artificial Intelligence (AI): machines doing tasks that typically require human intelligence
Machine Learning: a type of AI in which computers use data to "learn" tasks through algorithms
Algorithm: a set of steps, such as mathematical operations (e.g., addition) or logical rules

Purpose of Guidelines

"Generative AI", such as ChatGPT, grew from a niche topic to a variety of publicly available tools with hundreds of millions of adopters in less than one year. Among other things, Generative AI presents an incredible opportunity for people to increase their efficiency and efficacy in work. Generative AI has also been used for several irresponsible applications including faking news headlines,[1] leaking personal information,[2] and enabling phishing cyber-attacks.[3]

The City is actively working to create policies and procedures around AI in general. This document serves as part of an evolving governance structure around responsible AI usage.

Application of the Guidelines

This document applies to all use of Generative AI by a City staff member, contractor, volunteer, or other person while performing a role for the City (collectively "users"). This document does not apply to users of Generative AI for personal purposes or business purposes unassociated with the City.

Generative AI does not refer to algorithms that a person directly defines. For example, a spreadsheet a human created to calculates taxes owed based on income is not "Generative AI". A general rule is that if you cannot write the system's entire algorithm, either because you do not understand the math or because it would take years to write down, then it is probably AI.

Departments may provide additional rules on the usage of Generative AI. Users should consult their manager if there are additional rules specific to their department.

[1] "ChatGPT is making up fake Guardian articles. Here's how we're responding", The Guardian
[2] "OpenAI says a bug leaked sensitive ChatGPT user data", Engadget
[3] "Council post: How ai is changing social engineering forever", Forbes

Principles for Using Generative AI

Usage of Generative AI shall follow the City's AI principles:

1. Privacy: Submit information to Generative AI tools that is ready for public disclosure. This includes any text, photos, videos, or voice recordings you share with the AI. Be mindful that the AI output may include unexpected personal information from another user and ensure removing any potential private information before publishing.
2. Accuracy: The City maintains trust with its residents and partners by providing accurate information. Review and fact check all outputs you receive from a Generative AI. Users should consult trustworthy sources to confirm that the facts and details in the AI-generated content are accurate. Trustworthy sources include official City documents and peer-reviewed journals. Consult your supervisor for other trustworthy sources (e.g., newspapers, blogs, or datasets). Be aware that many systems, like ChatGPT, may only use information up to a certain date (e.g., 2021 for ChatGPT) and cannot guarantee the content they generate is accurate.
3. Transparency: The user shall be clear when he/she uses Generative AI. This can often include citing that you used AI in creating a product. See how and when to cite Generative AI in the "Citing Generative AI" section under "Guidance while Using Generative AI".
4. Equity: AI system responses are based on patterns and relationships learned from large datasets derived from existing human knowledge, which may contain errors and is historically biased across race, sex, gender identity, ability, and many other factors. Users of Generative AI need to be mindful that Generative AI may make assumptions based on past stereotypes and need to be corrected. Establish guidelines to address equity as it relates to services in your department.
5. Accountability: The person using AI is accountable for the content it generates. Use Generative AI with a healthy dose of skepticism. The level of caution used should correspond to the risk level of the use case (see "Assessing Risk in Generative AI Use Cases"). It is always important to verify information provided by Generative AI.
6. Beneficial: User should be open to responsibly incorporating Generative AI into their work where it can make services better, more just, and more efficient. For example, a tool like ChatGPT can help users go from an outline to a draft Council memorandum quickly, enabling them to focus more time on the analyses and findings that inform recommendations to Council.

Getting Started with Generative AI for City Use

Usage of Generative AI may be Subject to the Public Records Act

Any retained conversations relating to City work may be subject to public records requests and must comply with the City's retention policies. In addition, users will need to comply with the California Public Records Act and other applicable public records laws for all City usage of Generative AI. This means any prompts, outputs, or other information used in relation to a Generative AI tool may be released publicly. Do not use any prompts that may include information not meant for public release.

Create an Account Specifically for City-related Work

If you choose to use Generative AI for City-related work, you shall have an account for all Generative AI usage in your role at the City using a City email address. The purpose of this is to ensure proper retention of public records and avoid comingling of public and personal records. This account should not be used for any personal purpose. Users can use their City email address for City usage, or they can create a shared account using a different work email address. For example, the Digital Privacy Office might create a shared ChatGPT account using the digitalprivacy@sanjoseca.gov email address. Regardless of whether a shared or work email address is used to create an account, users should use a unique password for the service. **Like any other account which uses a City email address, the password should not be the same password used to log in to any City devices.** For example, if a data breach occurs on ChatGPT (which happened in March 2023)[4] and your password is stolen, a hacker should not be able to log into your laptop with that information.

If users use personal devices or accounts to conduct City work, the records generated may still be subject to search and disclosure. The records generated may include both the content users input and the content users receive from the Generative AI system.

Understand the Terms and Conditions

The City does not currently have agreements in place for common Generative AI systems, such as ChatGPT or Bing AI. If you choose to use Generative AI for City work and agree to the terms and conditions of a system without a City agreement in place, you are responsible for complying

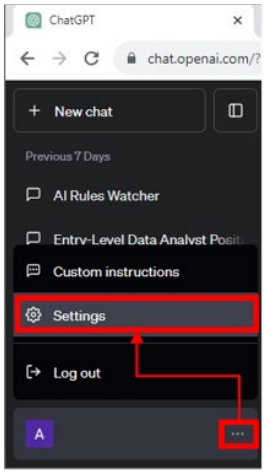
[4] ChatGPT confirms data breach, raising security concerns. (2023, May 16). Security Intelligence.

with those terms and conditions. In the event that the City forms an agreement with a Generative AI service, this section will list those services.

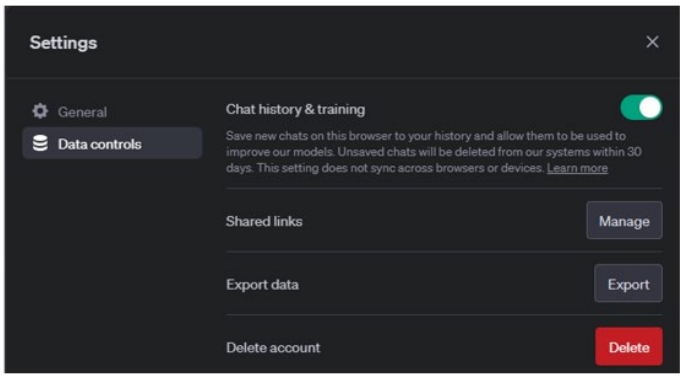
Opt out of data collection if possible

Some services offer an option to opt out of data collection. This means the generative AI system will not keep the data you provide, and it will not be used in the system's models. Opt out of data collection and model training whenever possible. For example, you can opt out of ChatGPT by going to "settings" → "data controls" → "chat history and training".

Opting out of ChatGPT data collection (as of August 2023)



1. On the side bar, click on the three dots and select "Settings"



2. Go to "Data controls" and uncheck "Chat history & training"

Verify the Copyright of All Generated Content

Users shall verify the content they use from any Generative AI systems does not infringe any copyright laws. For example, City employees could check the copyright of text-based content with plagiarism software and the copyright of image-based content with reverse Google searches, although neither of these approaches guarantees protection against copyright infringements. If users are uncertain if content violates copyright, they should either edit the content to be original or not use it.

Ownership of Generated Content

In most cases, the user owns the content they input into a Generative AI service and the information they receive as an output. The user can use the content at their discretion, in accordance with City policy and any terms and conditions he/she has agreed to. However, many Generative AI companies still retain the right to use both the input and output content for their own commercial purposes. For example, this could include a Generative AI company using City

data to train their models or distributing City output data for marketing campaigns. This emphasizes the importance that only information the City is ready to make public should be entered into a Generative AI system.

Joining AI Working Groups

The City is dedicated to providing practical guidance around AI that protects people from harm while providing the best services to residents. To accomplish this, the City has three engagement groups dedicated to informing AI use in the City:

1. City AI working group: City staff discuss AI policy, use cases, and guidelines. Users can learn more about AI in the City, discuss potential ideas in their departments, and flag any potential concerns.
2. Digital Privacy Advisory Taskforce: External Taskforce of experts around Digital Privacy and AI. The Taskforce advises and recommends on the City's digital privacy practices, including responsible AI.[5]
3. GovAI Coalition: The City of San José is collaborating with government agencies across the country to ensure that the AI systems we use serve all of our communities. The group collaborates on items including responsible AI governance, vendor accountability, and sharing use case experiences. If you are an agency interested in joining, you can do so at sanjoseca.gov/govai.

In addition to these three groups, the City holds opportunities for the public to provide feedback, including in-person sessions in San José, virtual sessions, and online at sanjoseca.gov/digitalprivacy. Members of the public are also able to contact the Privacy and AI team directly at digitalprivacy@sanjoseca.gov.

Guidance while using Generative AI

Citing Generative AI

When to Cite:

Users must cite the Generative AI when a substantial portion of the content used in the final version comes from the Generative AI. A "substantial portion" will be further defined in future working group discussions. **Any statements used as fact must cite a credible source rather than**

[5] Digital Privacy Advisory Taskforce webpage: https://www.sanjoseca.gov/your-government/departments-offices/information-technology/digital-privacy/digital-privacy-advisory-task-force

the AI. Credible sources include official City documents and peer-reviewed journals. Consult your supervisor for other trustworthy sources (e.g., newspapers, blogs, or datasets).

All images and videos must cite any AI used in their creation, even if the images are substantially edited after generation.

How to Cite:
Generative AI can be cited as a footnote, endnote, header, or footer. Citations for text-generated content must include the following:

- Name of Generative AI system used (e.g., ChatGPT-4, Google Bard, Stable Diffusion)
- Confirmation that the information was fact-checked.

For example: "This document was drafted with support from ChatGPT. The content was edited and fact-checked by City staff. Sources for facts and figures are provided as they appear."

Citations for images and video must be embedded into every frame of the image or video. For support on how to do this, see the "Creating Images or Video" use case in the appendix or reach out to digitalprivacy@sanjoseca.gov.

Recording usage of Generative AI

The City needs to understand how users are using Generative AI tools in their work. When you choose to use Generative AI to support your work, report that usage through this form: https://forms.office.com/g/3Znipym4k5. The form will take 1 minute. You do not need to wait for a response after filling out the form to use Generative AI, unless required by your department or manager. This is only meant to track usage in aggregate.

Additional guidance and advice around using Generative AI can be found in the Appendix.

Assessing Risk in Generative AI Use Cases

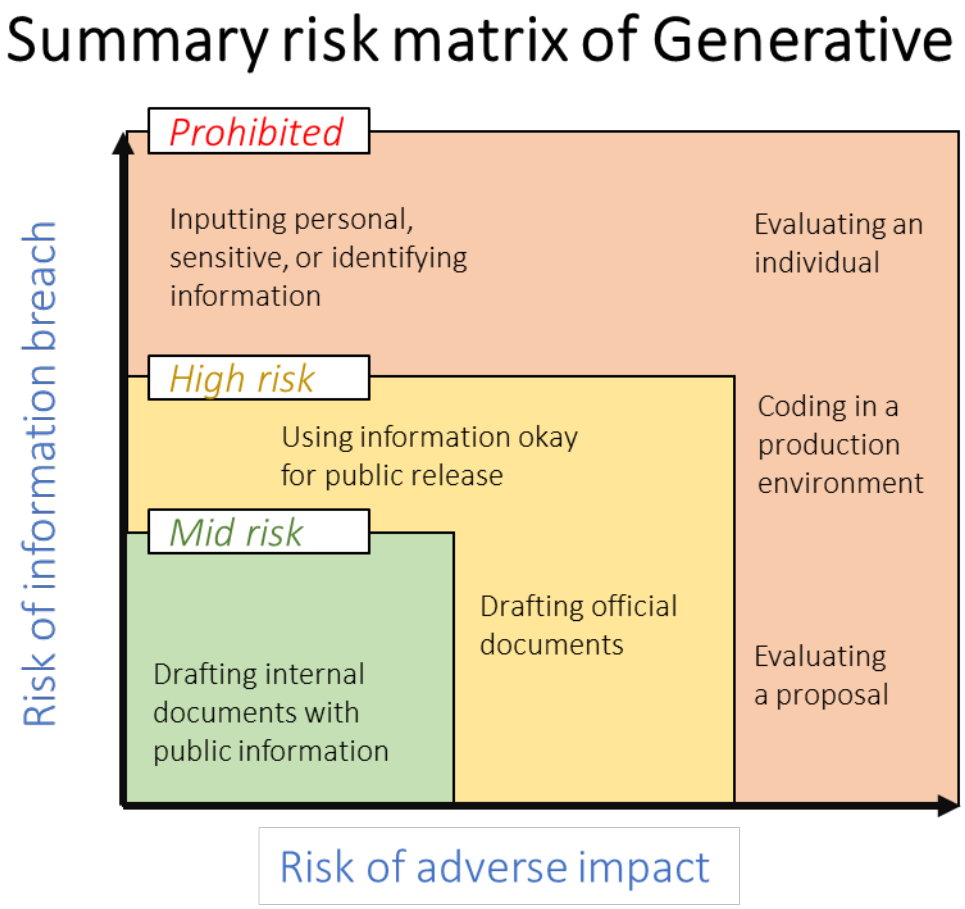
The risk presented by Generative AI tools varies by use case, with the risk spectrum ranging from mid-risk to high-risk to intolerable risk.

Generative AI risk is determined by two key factors:

1. Risk of information breach: the potential harm if the information exchanged with a Generative AI is released to an unintended audience. This can include entering personally identifiable information, sensitive records, or confidential business information into Generative AI. Additionally, any information entered into Generative AI may be subject to

the Public Records Act. If you wouldn't share the information in a public forum, don't share it with a Generative AI.

2. **Risk of adverse impact:** the potential harm of using the output for a decision, task, or service. This impact can be different for different populations and should be considered from an equity lens, such as adverse impacts to people of a certain race, age, gender identity, or disability status. Not only can AI be biased, but it can also provide false information. In general, if Generative AI is used in relation to City processes that can alter an individual or community's rights, freedoms, or access to services, it should be thoroughly reviewed by multiple users before any document is finalized or action is taken.



When Engaging in High-risk Use Cases

Keep in mind the tone and specific language in the AI output. Generative AI is trained on a global context and may not use the vocabulary or tone consistent with the City and its values. Simple examples include replacing "citizen" with "resident" in documents, and capitalizing "City" when

referring to the City of San José. These documents, like any others, require thorough review before moving from draft to final product.

Cite verifiable sources for all facts and figures (past memos, newspapers, research papers, etc.). ChatGPT or other Generative AI are not definitive sources. Facts should be accompanied by links or citations to sources that the general public could find, such as news articles or research papers. ChatGPT and other AI can fabricate sources if asked, so do not rely on them for finding citations either. Find sources directly and confirm they are legitimate before using.

Anything that would not be released or shared with the public should not be input into the AI. This includes information such as draft RFP requirements that should not be public yet, vendor transactions, procurement approvals, or internal City decisions.

Additional details on risk can be found in the Appendix.

Concluding Thoughts

Generative AI presents users an opportunity to work better, faster, and smarter. However, because the technology and the laws surrounding it are evolving and present unknown risks, its adoption comes with ethical considerations. Remember the fundamental rules when using any Generative AI:

1. Never submit personal or confidential information into a Generative AI.
2. Review, revise, test, and fact check any output from a Generative AI.
3. Be transparent when content was drafted using Generative AI.
4. Return to this document often, as guidance on usage will change rapidly.

By keeping the above guidance in mind when using generative AI tools, we can ensure the safe and responsible use of AI by the employees of the City. If you or your department has any questions, comments, or concerns around using Generative AI, please contact your team at digitalprivacy@sanjoseca.gov. The Privacy office can provide users trainings, set up AI evaluations, and help your team do the best with Generative AI.

Appendix

A Definition of Generative Artificial Intelligence

Generative Artificial Intelligence, commonly referred to as “Generative AI” or “GenAI”, is an “automated system” used to generate “content”.

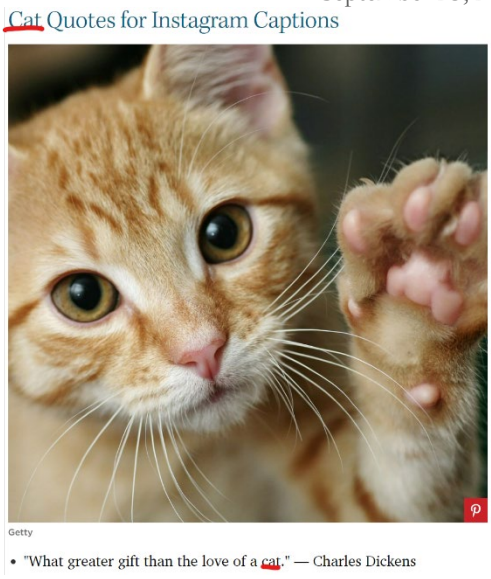
An "automated system" is any system, software, or process that uses computation as part of a system to generate outputs, outcomes, make or aid decisions, inform policy implementation, collect data or observations, or otherwise interact with individuals and/or communities.[6]

“Content” includes text, emails, presentations, images, video, audio, architectural documents, diagrams, and other forms of media.

Generative AI uses massive datasets to generate content that someone would want given a prompt (see definition of “prompt” below). For example, ChatGPT has collected data on millions of webpages to identify sentence patterns that commonly come next after someone types a phrase. Online information is paired with human training where algorithm developers manually judge and correct the output of the system. For example, it may have required a combination of millions of webpages and a human developer to train ChatGPT that “Jack fell down, and broke his crown” should be completed by “and Jill came tumbling after.”

[6] Definition derived from the United States White House Office of Science and Technology Policy AI Bill of Rights (October 2022) and the National Institute of Standards and Technology Artificial Intelligence Risk Framework (January 2023)

Billions of images are shared online every day, along with hundreds of thousands of hours of video[7] and countless text posts. Much of this information is connected to other information on the internet. For example, pictures of cats are often connected with captions that have the word "cat" in them. These connections allow a computer to, after millions of connections, "learn" what a cat looks like. Eventually, a computer can create an image of a cat based on all the previous images it has seen.



AI systems apply this same approach to music, books, poems, voices, videos, and anything else created on the internet.

Prompts and Generative AI

Generative AI relies on a user (e.g., a person) to "prompt" the AI to generate content. "Prompts" are any direction provided by a user. Examples of Generative AI include:

1. Creating text based on a prompt
2. Creating a picture or video based on a prompt
3. Making an audio file of a famous person saying something they did not say
4. Creating a movie scene based on a text prompt and pictures of the characters

Examples of prompts include:

1. **Text prompt to generate text content.** For example: "Tell me a story about three people becoming friends despite their differences"
2. **Text prompt to generate picture/video content.** For example: "Draw a cow with long hair and an ornate bell"
3. **Voice and text prompt to generate audio content.** For example: [Upload a recording of Tim Cook] "Say 'I'll just warn you now, I don't know how to use a computer' in the voice provided."

[7] (QUT), Q.U.of T. (2022) *3.2 billion images and 720,000 hours of video are shared online daily. can you sort real from fake?*, *QUT*. Queensland University of Technology

4. **Image and text prompt to generate picture/video:**  "Re-draw this bear with cleaner lines and give the bear a crown. Then show a clip of the bear running."

Details for Understanding Generative AI Risk

Understanding "Risk of Information Breach"

General rule: If the information exchanged with a Generative AI system would be harmful to a person or community if made public, it is a high or intolerable risk. Services like ChatGPT have been compromised in the past and leaked personal information.[8] Until private applications with higher security are deployed in the City, all information exchanged with Generative AI has a reasonable risk of being compromised.

Mid-risk information includes non-identifying and non-confidential information. For example, a simple email response or instructive documents often contain only general information that would not present any risk if made public.

High-risk information includes personally identifiable information (e.g., full name, birth date, email address) and confidential business information that may have larger implications to City processes. **Until a private application is deployed with security measures approved by the Cybersecurity Office, no high-risk information shall be provided to a Generative AI system.**

Prohibited risk information includes highly sensitive and identifying information. This includes data such as credit card numbers, bank account information, social security numbers, and other information that requires rigorous security measures and compliance standards before being processed.

Understanding "Risk of Adverse Impact"

General rule: If you are using Generative AI in relation to City processes that can alter an individual or community's rights, freedoms, or access to services, it is at least high risk and should be thoroughly reviewed before any document is finalized or action is taken. Additionally, any action that could reasonably lead to the City engaging in legal infringements on intellectual property are prohibited.

[8] "OpenAI says a bug leaked sensitive ChatGPT user data" https://www.engadget.com/openai-says-a-bug-leaked-sensitive-chatgpt-user-data-165439848.html

Mid-risk impact includes tasks associated with drafting internal messages, internal documentation, and idea generation. These tasks can be sped up with the support of Generative AI, but require many more steps before reaching a public impact.

High-risk impact includes tasks associated with official City documents or messaging. It also includes uses that require substantial editing and review before usage. These tasks require thorough review at the time of generation before using in any work context. Special care should be taken when a task may impact individuals differently across factors such as race, age, gender identity and disability (e.g., a memo about tree canopy inequity in neighborhoods).

Prohibited risk impact includes tasks that undermine trust in the City through false statements or news; deny people due process such as in resource allocation, job evaluations, and purchasing decisions; or expose the City to substantial security or legal risks. Generative AI does not have reasoning behind the content it produces and cannot justify a decision.

Examples of Generative AI Use Cases by Risk Level

Examples of Mid-risk Use Cases

1. Drafting messages to staff and trusted partners

Generative AI tools can help users draft emails or other messages to staff and trusted partners. ChatGPT is a tool commonly used for this purpose. You can prompt ChatGPT to provide formal sounding language from general framing of the message. You can also have it draft emails in different tones by asking for a different tone.

Additional Guidance:

1. You may be inclined to use ChatGPT to help with email replies. Do not copy your current email thread into ChatGPT. The email was sent to select people and may be confidential.
2. Be mindful about the purpose of the email, and if it is appropriate to use Generative AI for drafting it. For example, Vanderbilt University received heavy backlash for using ChatGPT to draft an email in response to a school shooting.[9]

Example:

[9] Korn, J. (2023) *Vanderbilt University apologizes for using CHATGPT to write mass-shooting email | CNN business, CNN*. Available at: https://www.cnn.com/2023/02/22/tech/vanderbilt-chatgpt-shooting-email/index.html

1. Prompt ChatGPT with the following: "I am the lead product manager for housing technology initiatives. We interview users to gather product requirements, prioritize features, and work with software developers on implementation. Draft me an email asking the software developers how long the housing database will take to implement, and what risks to implementation they see."
2. Carefully read through the email, perform final fact-check and other edits to the draft email. Manually add in personal information or internal confidential details before sending.
3. Cite at the end of the message "*some of this content was drafted using ChatGPT. All facts, figures, and statements were reviewed by the sender to be accurate."*
4. If someone replies to your email asking for what you would like to see in the database, you can return to ChatGPT and prompt it with the following: "I want the database to be easily understood by our field staff, draft this request to the developers". Read the draft, fact-check, and manually add information as needed. Cite ChatGPT at the end of the message as done previously.

2. Framing written content not intended for official release

Generative AI can be useful for creating an outline or structure for your written content. This can include an outline for a cover letter, long-form writing, project documentation, or speaker notes for a presentation. When the written content is not intended for official public release, it presents less risk than official City publications (like memos or policies). ChatGPT is the most common tool for this use case. You can write a few key points you would like to detail, any themes you want present, the kind of voice you would like, and how long you would like the content. Remember that information you input into a Generative AI system may be subject to a Public Records Act request.

Additional Guidance: Unless you have a Generative AI trained to your context—a feature likely not available for another year—the tool will provide generic language that does not apply to the City. For example, ChatGPT may use the word "citizens" rather than "residents" when referring to the people we serve because it is not used to San José's specific circumstances. As always, make sure to review, revise, and fact-check any output from Generative AI.

Example prompt steps:

1. Prompt ChatGPT with the following: "I am writing an instruction manual for how City staff should add content to the City's website. Draft an outline that can be posted on our

intranet. It should have a section about how to create lists, add hyperlinks, and show pictures using a content management system. Draft in a formal tone but make the text clear and approachable."

2. Review, revise, and fact-check. Manually enter any confidential or private information into the draft or final version.
3. Cite that you used ChatGPT in the drafting process. See how to cite Generative AI in the "Citing Generative AI" section under "Guidance while Using Generative AI".

3. Learning from a document

You may copy a large public document into a Generative AI tool and ask the AI questions about the document. ChatGPT is the most common tool for this use case.

Additional Guidance: The document or information you paste into the Generative tool AI should already be public information.

Example prompt steps:

1. Ask ChatGPT to "Summarize the following document. Let me know if there is any mention of California, cities, or San José."
2. Copy the text from a public document and paste it into ChatGPT. An example document would be the text from this news article: https://www.fiercewireless.com/wireless/san-jose-plans-smart-city-infrastructure-verizon-and-at-t. Copy text and paste into ChatGPT
3. You do not need to cite ChatGPT unless you quote specific text outputted from ChatGPT in future written content.

4. Brief list of other mid-risk use cases

1. Helping you come up with a name for your team. For example, "give me ten names for a team focused on AI and Privacy that uses the acronym 'SAFE'." Be sure the name is not already used by another team in the City.
2. Learning about a new topic in a way that you can understand. For example, "explain quantum mechanics to me like I'm five". Verify anything you learn from ChatGPT before applying the knowledge in a City context.
3. Helping you find the right word for a concept. For example, "what is the word for the second-to-last episode in series". Once the AI provides the word, search the word on Google (or elsewhere) to confirm it means what you think it means.

Examples of High-risk Use Cases

1. Drafting memos and or other public-facing City documents

Generative AI tools can help users draft memos and other public-facing documents more efficiently. Because the content is meant for public release, it is treated as high-risk and should **be reviewed and edited multiple times before release**. ChatGPT is the tool most used for this purpose.

Additional guidance: The City expects users to produce their own research that informs memos, such as information related to policy changes and program changes. Memos, press releases, and other publications also have their own City-specific formats and standards to follow. Consult your supervisor to make sure your memo follows City standards.

Example prompt steps:

1. Provide context around the memo but only provide public details: "We are building an encampment management work order system at the City to better coordinate services and have just completed the detailed design. We will be presenting to the City council on the latest update."
2. Then request ChatGPT to draft a memo: "Draft a memorandum with the following sections and key points: Introduction: (add bullet points), Human-Centered Design work (add bullet points), Requirements (add bullet points), Next-steps (add bullet points).
3. After initial memo draft, prompt ChatGPT to "Draft a conclusion summarizing all the prior sections."
4. Manually add in any non-public information to the draft memo produced by ChatGPT.
5. Carefully read through memo, perform fact-check and other edits to memo to maintain a tone consistent with City documents.
6. Cite verifiable sources (past memos, newspapers, research papers, etc.) for all facts and figures in the memo.
7. If required, cite that you used ChatGPT in the drafting of the memo. See how and when to cite Generative AI in the "Citing Generative AI" section under "Guidance while Using Generative AI".

2. Writing an RFP/RFB/Vendor relations

Generative AI tools can help users draft RFP/RFBs more efficiently. Because the content is meant for public release, it is a high-risk use case. ChatGPT is the tool most used for this purpose.

Additional guidance: RFPs and other publications have their own City-specific formats and standards to follow. Consult your supervisor and purchasing business partner to make sure your memo follows City standards. Guidance can be found on the City intranet, including the "Strategic Procurement Guidelines for RFP and Contract Requests To Finance-Purchasing". Take special care not to provide ChatGPT information that is not meant to be public yet. For example, if the specific requirements of the RFP are not meant to be public yet, do not input them into your prompts.

Example prompt steps:

1. Provide context around the procurement document without providing non-public details: "We are procuring an encampment management work order system at the City to better coordinate services."
2. After ChatGPT responds, ask to draft the procurement document: "Draft an RFP with the following sections and key points: Introduction: (add bullet points), Scope of Work (add bullet points), Requirements (add bullet points), Cost Breakdown (add bullet points)."
3. Manually add in any non-public information to draft document produced by ChatGPT.
4. Carefully read through the document, perform fact-check and other edits.
5. If required, cite that you used ChatGPT in the drafting process. See how and when to cite Generative AI in the "Citing Generative AI" section under "Guidance while Using Generative AI".

3. Writing advertisements, social media posts

Generative AI tools can help users draft promotional material. Because the content is meant for public release, it is a high-risk use case. Chat GPT is the tool most used for this purpose.

Additional guidance: None, refer to AI principles and guidance for high-risk use cases.

Example prompt steps:

1. Provide ChatGPT with details around the needed post and audience, for example: "Draft a cute tweet of less than 240 characters that reminds families that tomorrow is Walk and Roll to school day"
2. Review output, edit to make personal to San José and relevant department or office, and post.

4. Writing job postings or job descriptions

If you provide a Generative AI with a list of qualities you want and a role title, it can help you draft a formal-sounding job description. Because the content is meant for public release, and job requirements can have a substantial impact on who applies, it is a high-risk use case.

Additional guidance:
The City expects users to follow existing standards on the format and content of job postings based on classifications. Consult your Human Resources business partner or your Department's HR representative for information on job classifications and postings.

Additionally, be mindful of the language used in the requirements, responsibilities, and tone used in the job posting. Check if the job description seems to use language stereotypically associated with a specific race or gender. Use gender-neutral language: Avoid using gender-specific pronouns (he, she) and job titles (fireman, firewoman). Instead, opt for inclusive terms such as "they" and "fire officer." Remove gender-coded words: Avoid using adjectives that may be associated with a specific gender, such as "aggressive" or "nurturing." Use neutral descriptors, like "results-driven" or "collaborative."[10]

Example prompt steps:

1. Provide ChatGPT with the previous public posting for an Analyst I position and request that ChatGPT "Draft a similar job description, but with a focus on using information to inform park capital projects."
2. Manually add in any non-public information to draft document produced by ChatGPT.
3. Carefully read through the document and edit for a more neutral tone, perform fact-check and other edits.

[10] Krakovetskyi, O. (2023, March 22). Eliminating Bias in Job Descriptions with ChatGPT – The DevRain Tech Blog - Medium. *Medium*. https://medium.com/devrain/eliminating-bias-in-job-descriptions-with-chatgpt-72b92ebc7911

5. Creating images or video

Some Generative AI tools such as Stable Diffusion and Dall-E can create images or video clips based on text prompts. The City needs to maintain its legitimacy as a trustworthy source when using video and images, which requires substantial precautions whenever using AI-generated visual content.

Additional Guidance:

1. **Use only for illustrative purposes.** For historical events, use real images rather than generated. For example, if you want a picture of a giraffe wearing a suit and tie for your presentation, generate it. If you are proposing a new visual diagram or abstract concept, you can also generate it. If you want a picture of the Mayor at City Hall, find an actual picture.
2. **Require a citation embedded into the image or video at all times**. Images and videos can easily be taken out of their original context and misinterpreted as reality. To prevent a news article or other secondary source from treating an image as fact, all images and frames of a video must specify that they were generated using an AI system. The citation shall be included in the image itself, and cannot be removed without editing or cropping the image.

Example Use Case:

1. Provide a prompt: “drawing of falcon and its chicks on top of a skyscraper”
2. Choose your image
3. Embed the citation into the image: “Image generated by DALLE-2”
4. Add alt text into the image that clearly states the image was generated by an AI system

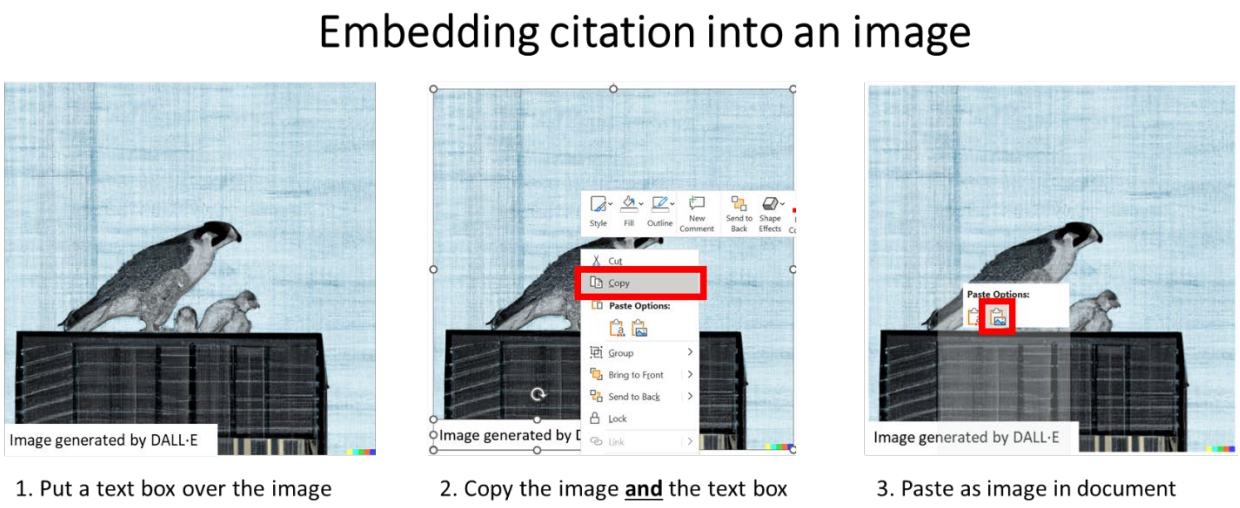


Figure 1: Three steps of embedding a text box into an image in order to cite the Generative AI system used for the image.

6. Creating presentation slides

If you provide Generative AI with some public information, it can create a presentation for you. Currently this feature is very new but may soon be integrated into existing applications like PowerPoint. Currently there is no clear leader in Generative AI for presentations, but a few examples are beautiful.ai and gamma.app. Presentations are automatically high risk because they go beyond text into images, which can present false information if the audience believes the image is real.

Suggested use: Provide a public document, or an outline with public information.

Additional guidance: Similar guidance to other public-facing documents, but with the additional requirement to cite all AI Generated images clearly on the image.

Example prompt steps:

1. Provide the AI with an article or outline from publicly available information. For example, an article about the evolution of cats.[11]
2. Generate the presentation
3. Fact check all content, review for tone and language
4. Cite Generative AI images

[11] Article used in example was Vsadmin. (2022, April 8). *The Evolution of Cats*. Killarney Cat Hospital. https://www.killarneycat.com/the-evolution-cats/

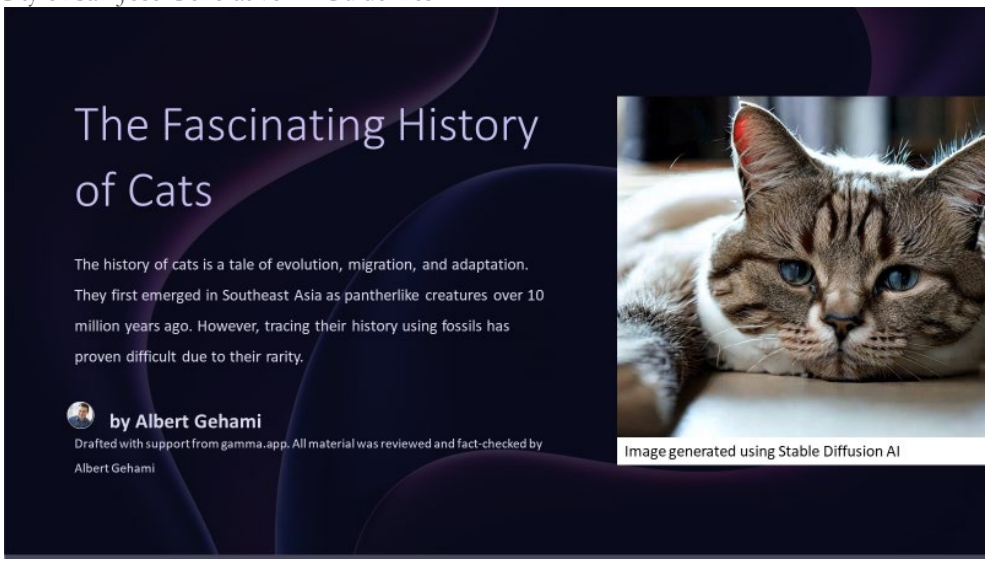


Figure 2: Image of a generated presentation with support from the AI tool gamma.app. The image on the right was generated using Stable Diffusion AI and was cited as such

7. Brief list of other high-risk use cases

Remember to follow the AI principles and general guidance for high-risk use cases.

1. Creating diagrams. For example, “create a flow chart of a tree turning into wood pulp and then into paper”. Replace pictures before publishing
2. Drafting papers. For example, “Here is my outline for my research paper, and my findings, draft a complete paper.”

Examples of Prohibited Use Cases

1. Programming or coding

Why it is prohibited: Code generated by an AI may be outdated, copyrighted, have identified vulnerabilities, or rely on other code that no longer works. The generated code is not cited to a date (like a stack overflow post would be), so it is unclear when the code would have been good.

What can you do with Generative AI: AI can help frame your coding problem, and help you draft pseudo-code to solve your problem conceptually. You can request code snippets for help defining syntax, and can be useful for testing projects in a low-risk, non-production environment.

2. Evaluations and Decisions

Why it is prohibited: Evaluating job applicants using AI has led to countless scandals of biased application reviews.[12] This evaluation issue also extends to other areas such as evaluating proposals or an existing employee.[13] AI-based evaluations expose the City to public protest across many key City functions such as hiring and purchasing.

Additionally, Generative AI shall not be used to determine highly sensitive decisions such as an individual's health plan, cost of bail, conviction of a crime, grades, or admissions to a program.

What can you do with Generative AI: AI can help flag key words and identify phrases within a document (see the mid-risk use case). However, the actual evaluation must be made by a person.

3. Language Translation

Why it is prohibited: Large Language Models like ChatGPT are not yet demonstrably better for translation than something like Google Translate.[14] Google Translate is also an AI system, but is built for specifically translating text, compared to modern Generative AI systems like ChatGPT, which attempt be a more general AI system for more problems.

Future Generative AI systems may be substantially better than existing translation AI systems, but they will require an evaluation of their performance before the City should use them over something like Google Translate. **Translations should be confirmed by a fluent speaker of both languages whenever possible.**

What can you do with Generative AI: Test out the system in a risk-free environment (e.g., you and a coworker testing the system), and report any translation system you would like to use to the Digital Privacy Officer for an algorithm evaluation. Contact digitalprivacy@sanjoseca.gov.

[12] Vallance, B. C. (2022, October 13). AI tools fail to reduce recruitment bias - study. *BBC News*. https://www.bbc.com/news/technology-63228466

[13] In general, people react worse to negative evaluations from AI than they do to negative evaluations from people. Lopez, Alberto, and Ricardo Garza. "Consumer bias against evaluations received by artificial intelligence: the mediation effect of lack of transparency anxiety." *Journal of Research in Interactive Marketing* (2023).

[14] Google Translate is also an AI system, but is built differently than the modern Generative AI systems like ChatGPT or Google Bard, which attempt to act as a more general AI system for more problems than just translating text

4. Creating voice or other audio

Why it is prohibited: Replicating a person's voice with AI in any City document or recording would undermine the trust of staff and the residents. Potential legal concerns also exist regarding replicating a person's voice. Do not generate audio through AI.

Additional Guidance around Generative AI

Be Aware of Targeted Cyber Attacks Using Generative AI

Although City staff are already familiar with handling cyber risks like phishing and malware, the advent of generative AI introduces heightened cybersecurity risks as the attacks can be more complex and personalized. Cyber threat actors may use generative AI in their attacks in the following ways:

- **Writing AI-powered, personalized phishing emails**: With the help of generative AI, phishing emails no longer have the tell-tale signs of a scam—such as poor spelling, bad grammar, and lack of context. Plus, with AI like ChatGPT, threat actors can launch phishing attacks at unprecedented speed and scale.
- **Generating deep fake data**: Since it can create convincing imitations of human activities—like writing, speech, and images—generative AI can be used in fraudulent activities such as identity theft, financial fraud, and disinformation.
- **Cracking CAPTCHAs and password guessing**: Used by sites and networks to comb out bots seeking unauthorized access, CAPTCHA can now be bypassed by hackers. By utilizing AI, they can also fulfill other repetitive tasks such as password guessing and brute-force attacks.

Detecting Generative AI

Software developers are building tools, like GPTZero, GPT Radar, and Originality.AI, designed to detect if a body of writing was created by a generative AI tool. These tools are in early stages of development and their detection accuracy rate may not always be accurate and should be used with caution. For example, there have been numerous incidents of instructors using ChatGPT detection tools falsely accusing students of plagiarism, endangering their grades and even diplomas.

Despite the limited accuracy of these tools, they allow residents to check if City documents were generated by AI regardless of whether users cite their usage or not. To build trust with residents,

users need to be proactive in communicating its usage of AI. Residents finding out on their own can cause reputation harm to the City.

Generative AI & Copyright

Numerous copyright lawsuits are springing up in which artists are suing AI companies like Stability AI and Midjourney for unauthorized use of their intellectual property to train the Generative AI systems. Large companies like Getty Images and Shutterstock are also joining suit against AI companies.

The US Copyright Office determined that art created solely by AI isn't eligible for copyright protection. Artists can attempt to register works made with assistance from AI, but they must show significant "human authorship." The office is also currently executing an initiative to "examine the copyright law and policy issues raised by artificial intelligence (AI) technology."