December 10, 2025

Reducing the Environmental Impact of GenAI Assistant Tools: Working Together to Support San Francisco's Climate Leadership

The City and County of San Francisco is proud to hold "A-List" status as a Global Climate Action Leader and remains deeply committed to adopting AI technologies that align with our ambitious climate goals.

As we implement Copilot Chat and other generative AI (GenAI) assistant tools to enhance service delivery, we are also engaging with our industry partners to promote greater transparency around the City's energy and water use associated with GenAI technologies. Research shows that GenAI models rely on significant electricity for computation and require large volumes of freshwater for both server cooling and electricity generation during data-center operations (Li et al., 2025)

While comprehensive reporting will take time, each of us can begin making a difference today by taking some simple voluntary steps to reduce the environmental footprint of our AI use. Because GenAI uses energy every time we run a query—and these repeated uses can surpass the one-time energy cost of training (Desislavov, 2023)—being thoughtful about how we use these tools matters. Lowering the energy impact of GenAI assistant tools begins with optimizing the prompt journey, from how we frame inputs to how we manage outputs (Verdadero et al., 2025)

Tips for Environmentally Conscious GenAI Assistant Tool Prompting

These tips apply across all GenAI assistant tools, including Copilot Chat, Copilot 365, ChatGPT, Gemini, and Claude.

1. Keep Outputs Short

The length of a GenAI assistant tool's response is one of the strongest drivers of its energy consumption (Verdadero et al., 2025). Longer outputs use more computing power and cooling, leading to a larger environmental impact.

Do this:

- **Set clear limits.** Add instructions like *"Under 100 words"* or *"5 bullets max."*

- **Be specific about what you need.** For example: *"Summarize key recommendations only"* or *"Provide three short takeaways."*
- **Stop early.** If the answer looks sufficient, click *Stop generating* instead of letting the model finish a long reply.

2. Write Fewer, Shorter, and Better Prompts

Each prompt you send to a GenAI assistant tool triggers inference computations that consume electricity and water through data center cooling and power generation. Recent studies show that shorter, well-structured prompts can reduce energy and water use during large language model inference (Jegham et al., 2025; Rubei et al., 2025). While the impact of a single query is small, longer, unclear, or more frequent prompts collectively increase computing demand and environmental footprint. Writing fewer, shorter, and clearer prompts improves response quality, saves time, and lowers resource use.

Do this:

- **Be concise.** Ask your main question directly.
- **Provide context early.** Include the audience, tone, and goal in your initial prompt rather than repeating them later.
- **Avoid repeat generations.** Use the same chat thread for related questions to maintain context and avoid repeating work. However, very long conversations increase processing and environmental costs, so start a new chat when the topic changes.
- **Be mindful with alternatives.** Many search engines now embed GenAI features (e.g., *Google AI Overviews*, *Bing Copilot Search*) that also increase compute load. Choose *"Web Results Only"* or add *"-ai"* to your query to disable generative summaries.

3. Limit High-Compute Tasks

Tasks involving images consume substantially more electricity, and likely more associated resources, than standard text-only queries (Luccioni et al., 2024). Similarly, reasoning models, which perform extended chain-of-thought processing to solve complex problems, require considerably more computational resources per query (Xu et al., 2025). Each regeneration compounds that impact.

Do this:

- **Minimize image generation.** Use GenAI assistant tools to generate images only when truly necessary. Opt for smaller or lower-resolution outputs when possible to reduce energy and resource use.

- **Use reasoning models selectively.** Whenever possible, only use reasoning-enabled or chain-of-thought models for tasks that genuinely require extended deliberation or complex problem-solving. For straightforward queries, a standard model is usually sufficient and far more efficient.

4. Use Cloud Storage Wisely

Cloud servers like OneDrive and Share Point draw continuous power for storage, operation, and cooling, even when files are not being accessed (Masanet et al., 2020). As AI-generated content grows, regularly deleting unnecessary AI-generated content helps limit long-term storage growth and the energy required to support it. *However, coordinate with your department's IT or records management team to ensure your cleanup practices align with City retention policies!*

Do this:

- **Keep only the final file.** Delete drafts, partial outputs, and duplicates of GenAI-generated content.
- **Clean up often.** Remove old or unused AI-generated files and enable auto-delete for temporary files.
- **Archive the rest.** Move older AI-generated files into your OneDrive or SharePoint Archive folder or a designated long-term storage library, and avoid keeping extra active copies.

5. Educate and Share

San Francisco's climate leadership depends on collective responsibility. Small, consistent actions by each City employee add up to measurable reductions in our overall digital footprint. User behavior plays a critical role in achieving system-level sustainability, how we use AI tools matters as much as the tools themselves (Verdadero et al., 2025).

Do this:

- **Share these tips with your team.** Discuss these tips in team meetings or digital literacy sessions.
- **Promote mindful AI use.** Encourage colleagues to think before prompting and to apply these practices in their own work.

- **Lead by example.** Demonstrate efficient prompting and resource-aware habits in your everyday use of GenAI assistant tools.

References

- Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2023). *Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning*. Sustainable Computing: Informatics and Systems, 38, 100857. https://doi.org/10.1016/j.suscom.2023.100857

- Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., & Hendawi, A. (2025). *How hungry is AI? Benchmarking energy, water, and carbon footprint of LLM inference*. arXiv:2505.09598 [cs.CY]. https://doi.org/10.48550/arXiv.2505.09598

- City of San José. (n.d.). *Generative AI guideline*. Retrieved October 2025, from https://www.sanjoseca.gov/your-government/departments-offices/information-technology/itd-generative-ai-guideline
- Li, P., Yang, J., Islam, M.A., & Ren, S. (2025). *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*. Retrieved November 25, 2025 from https://arxiv.org/pdf/2304.03271
- Luccioni, A. S., Jernite, Y., & Strubell, E. (2024). *Power Hungry processing: Watts driving the cost of AI deployment?* arXiv:2311.16863 [cs.LG]. https://doi.org/10.48550/arXiv.2311.16863
- Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). *Recalibrating global data center energy-use estimates*. Science, 367(6481), 984–986. https://doi.org/10.1126/science.aba3758
- Rubei, R., Moussaid, A., di Sipio, C., & di Ruscio, D. (2025). *Prompt engineering and its implications on the energy consumption of Large Language Models*. arXiv:2501.05899 [cs.SE]. https://doi.org/10.48550/arXiv.2501.05899
- Verdadero, L., Drobnjak, I., Bosilkovski, H., Wu, Z., Fischer, E., Perez-Ortiz, M. (2025). *Smarter, smaller, stronger: resource-efficient generative AI and the future of digital transformation*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000394521
- Xu, Y., Dong, H., Wang, L., Sahoo, D., Li, J., & Xiong, C. (2025). *Scalable chain-of-thoughts via elastic reasoning*. arXiv:2505.05315 [cs.LG]. https://doi.org/10.48550/arXiv.2505.05315

Acknowledgements

We extend our sincere appreciation to the following contributors for their expertise and guidance in the development of this document:

- **Dr. Riccardo Rubei**, Postdoctoral Researcher, Mälardalens University
- **Nidhal Jegham**, Graduate Student and Teaching Assistant in Statistics, University of Rhode Island
- **Professor Ivana Drobnjak**, Professor of Computational Healthcare and Director of the Undergraduate Program, Department of Computer Science, University College London
- **Dr. Sasha Luccioni**, AI & Climate Lead, Hugging Face

We are deeply grateful for their insights. Any remaining errors or omissions are our own.