Energy-Efficient Management & Provisioning of Generative AI Services for 6G Networks

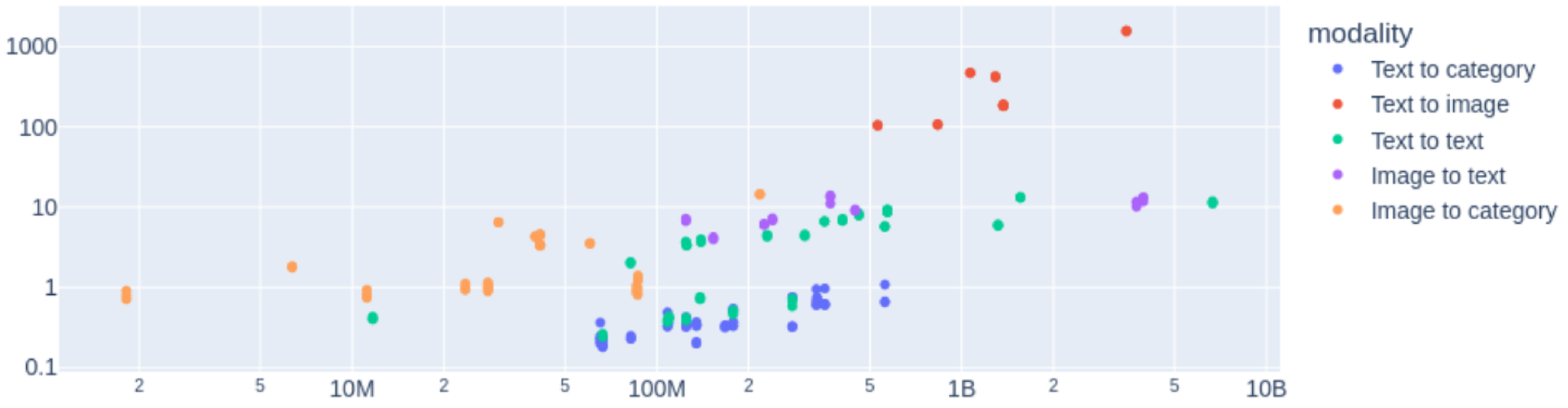
Adrien Sardi[†✳], Marie Line Alberi-Morel[†], Frédéric Giroire[✳♣], Sara Alouf[✳], Joanna Moulierac[✳♣]

[†] Bell Labs, Nokia Networks France,
[✳] INRIA Sophia-Antipolis,
[♣] Université Côte d'Azur, CNRS

1. Generative AI for 6G Networks

Inference		Training	
Google Search	ChatGPT Query	Netflix Operations (2021)	Training of GPT-4
0.3 Wh	2.9 Wh	94,000 MWh	62,300 MWh

Consumption of different applications [2]

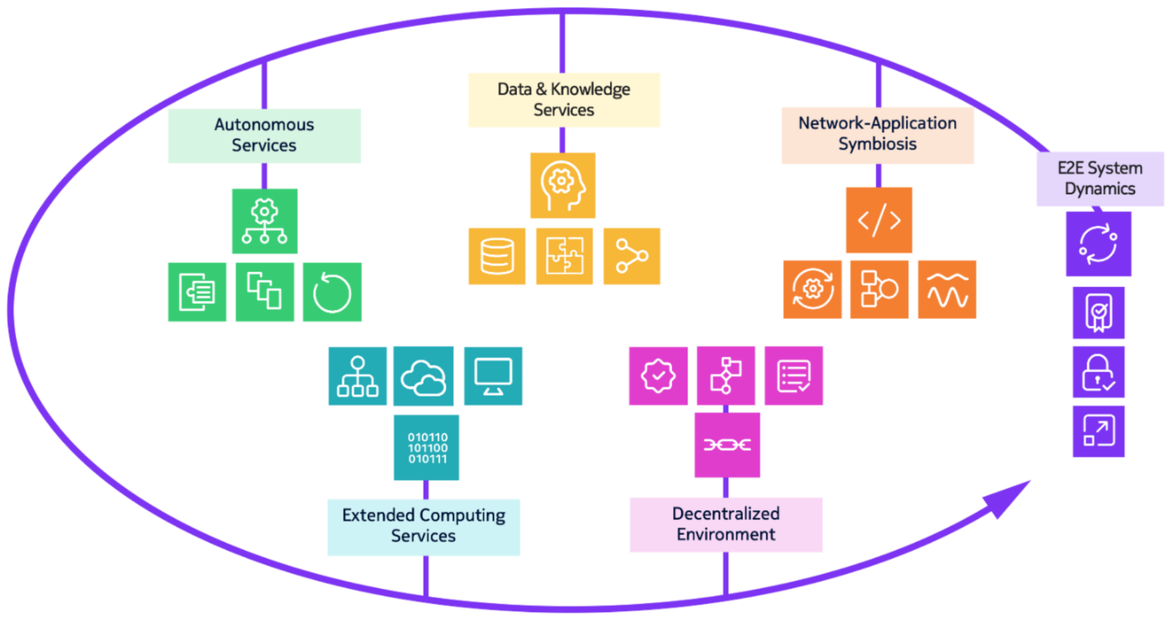


Carbon Emissions (CO_{2eq}) as a function of the model size (number of parameters) using `Code Carbon` [1].

2. Goals

1. Refine existing **Power Models**.
2. Deep dive into the application of **New AI Methods** to efficiently reduce the overall **Energy Consumption**.
3. Efficiently **Reconfigure Networks** when demand or energy availability has changed.

3. Integration in the UNEXT© environnement



The **Nokia Unified Networking Experience (UNEXT©) architecture** components [3]

References

[1] Sasha Luccioni, Yacine Jernite, and Emma Strubell. "Power Hungry Processing: Watts Driving the Cost of AI Deployment?" In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. New York, NY, USA: Association for Computing Machinery, June 5, 2024, pp. 85–99. ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658542. URL: https://dl.acm.org/doi/10.1145/3630106.3658542 (visited on 01/29/2025).

[2] *Netflix's Energy Consumption Nearly Doubled in 2019*. PCMAG. Feb. 10, 2020. URL: https://www.pcmag.com/news/netflixs-energy-consumption-nearly-doubled-in-2019 (visited on 03/11/2025).

[3] A. Sefidcon, C. Vulkán, and M. Gruber. *UNEXT: A Unified Networking Experience*. https://onestore.nokia.com/asset/213573. 2024.